Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 4, April : 2023

FAKE PROFILE DETECTION USING MACHINE LEARNING

Ms. VUNDI JITYA[1],Ms.ADAPALA RUCHITHA[2] , Ms. BOYEDY NANDHINI [3],Ms.BODHU PUJITHA[4], Mrs. MAKKAPATI LAKSHMI PRASANNA

1.BTECH, VIJAYA INSTITUE OF TECHNOLOGY FOR WOMEN, ENIKEPADU, VIJAYAWADA, ANDHRAPRADESH, INDIA-521108 vundi.jitya.003@gmail.com

2. BTECH, VIJAYA INSTITUE OF TECHNOLOGY FOR WOMEN, ENIKEPADU, VIJAYAWADA, ANDHRAPRADESH, INDIA-521108

3. BTECH, VIJAYA INSTITUE OF TECHNOLOGY FOR WOMEN, ENIKEPADU, VIJAYAWADA, ANDHRAPRADESH, INDIA-521108

4. BTECH, VIJAYA INSTITUE OF TECHNOLOGY FOR WOMEN, ENIKEPADU, VIJAYAWADA, ANDHRAPRADESH, INDIA-521108

5. ASSISTANT PROFESSOR, COMPUTER SCIENCE AND ENGINEERING, VIJAYA INSTITUE OF TECHNOLOGY FOR WOMEN, ENIKEPADU, VIJAYAWADA, ANDHRAPRADESH, INDIA-521108 makkapatiprasanna@gmail.com

ABSTRACT

Online impersonation and fraudulent accounts are a problem on the social network, which is an essential part of our lives. According per the "Community Standards According to Facebook's "Enforcement Report" from March 2018, 3–4% of its active accounts at the time were fraudulent, and 583 million bogus accounts had only been removed in the first quarter of 2018. In this project, we offer a model that might be applied to identifying if an account is real or false. It is unnecessary to manually examine each account because our model, which uses Support Vector Machine as a classification technique, can process a big dataset of accounts at once. We are concerned with the community of Fake Accounts, and our issue is one of classification or clustering. We employ artificial neural networks and machine learning to assess the likelihood that a Facebook friend request is genuine or not. We also describe the relevant classes and libraries. We also examine the sigmoid function and the selection and application of the weights. Finally, we take into account the social network page's parameters, which are crucial to the solution that is offered. The existence of bots and phoney profiles is another risk factor for personal data being collected for illicit purposes. Bots are computer programmes that can compile data about users without their knowledge. Web scraping is the term for this activity. The fact that this behaviour is legal makes it worse. To access private information on a social networking site, bots can be disguised or appear as a false friend request.

1 INTRODUCTION

Facebook is the most widely used form of social media, with 2.46 billion users worldwide as of 2017 [1]. Social media platforms generate income from user-provided data. The typical user is unaware that when they use a social media network's service, their rights are forfeited. Social media businesses stand to gain significantly at the cost of the user. Facebook generates income from adverts and data each time a user publishes a new location, new images, expresses their likes and dislikes, and tags other users in anything they post. The average American user generates roughly $26.76 per quarter, to be more precise. With millions of users, that sum grows quite quickly. In the current digital era, the growing reliance on computer technology has made the average person more susceptible to crimes like data breaches and potential identity theft. These attacks are frequently carried out without warning and without informing the individuals whose data was compromised. Social networks currently have little reason to strengthen their data security. Social networking sites like Facebook and Twitter are frequently the targets of these hacks. Banks and other financial institutions are another possible target. In the current generation, everyone's social life is now entwined with online social networks. It has gotten simpler to add new friends and stay in touch with them and their updates. Online social networks have an impact on a variety of fields, including science, education, community activism, employment, and business. These online social networks have been the subject of research to see how they affect people. Instructors may quickly reach their students using this, creating a welcoming environment for them to learn. Teachers are becoming more familiar with these sites and using them to provide online classroom pages, assign assignments, hold conversations, and other activities that greatly enhance learning. Employers can utilise these social networking sites to find brilliant candidates who are enthusiastic about their jobs and whose backgrounds can be easily checked. Each user of a social networking site has a profile and can communicate with friends, share updates, and connect with people. Each user of a social networking site has a profile and can communicate with friends through the site.exchange updates and connect with like-minded individuals. These online social networks employ web 2.0 technology, enabling user interaction. Social networking services are expanding quickly and altering how individuals communicate with one another. People with similar interests are brought together by online groups, making it simpler for users to establish new connections. Over the past few years, online social networks (OSNs) like Facebook, Twitter, and LinkedIn have grown in popularity. OSNs are used by people to stay in touch with one another, share news, plan events, and even operate their own online businesses. With more than 2.2 billion monthly active users and 1.4 billion daily active users, the Facebook community is still expanding, up 11% from the previous year. Machine learning techniques were used to identify phoney accounts in order to detect them on social media sites after pre-processing the generated data. The classification results of the algorithms Support Vector, Random Forest, and Neural Network: Fake account detection is done by machines. The described algorithms' accuracy rates for identifying bogus accounts are compared, and the algorithm with the highest accuracy rate is noted. On a social media site like Twitter, our study focuses on finding phoney profiles and clever BOTS. Because they are automated and may be used without a human, phoney profile bots are employed increasingly frequently these days. As technology advances, A. I. is now used in every field of work and replacing humans, making it more difficult to detect bots than human-made fake profiles. Bots and fake profiles created for stilling personal data of users on social media platforms like Twitter as well as for spreading fake news and rumours can have

a significant impact on society. Hence, based on Twitter data metrics like followers, tweets, following, etc., we develop a model that recognises sophisticated bots and human-generated phoney identities. As the Twitter API allows us to retrieve a user's real-time Twitter data, we are using the Twitter dataset for our model.

2. RELEATED WORK

1) Myo MS, Nyein NM (2018) Using a blacklist to detect fake accounts on Twitter. Pages 562–566 in: IEEE 17th international conference on computer and information and information science

2) Identifying clusters of bogus accounts in online social networks. Cao X, David MF, and Theodore H. Pages 91–101 of the 8th ACM Workshop on Artificial Intelligence and Security

3) Buket E, Ozlem A, Deniz K, and Cyhun A. 2017. Detection of Twitter bogus accounts. Pages. 388–339 in IEEE's 2nd international conference on computer science and engineering

5) Yeh-Cheng C, Shyhtsun FW (2018) FakeBuster: a powerful tool for activity-based fake account detection. International Symposium on Parallel Architectures, Algorithms, and Programming, 9th IEEE. pp 108–110

6) Myo MS, Nyein NM (2018) Using a blacklist to detect fake accounts on Twitter. International Conference on Computer, Information, and Information Science, 17th IEEE. pp \s562–566

7) Qiang C, Michael S, Xiaowei Y, and Tiago P (2012) Detecting bogus accounts in large-scale social online services. 9th USENIX conference on designing and implementing networks. pp 1–14

8) Fakebook: identifying false profiles in online social networks, Mauro C, Radha P, Macro S. International IEEE conference on mining and analysing social networks. pp 1071–1078

9) Using machine learning approaches, Duraipandian M. (2019) evaluated the performance of the routing algorithm for Manet. J Trends Compute Sci Smart Technol (TCSST), 1(1), 25–38

10) Yazan, B., D., Georgos, S., Jorge, L., Matei, R., Konstatin, and H. (2016) Integro: Leveraging victim prediction for robust fake

3.PROPOSED WORK AND ALGORITHM

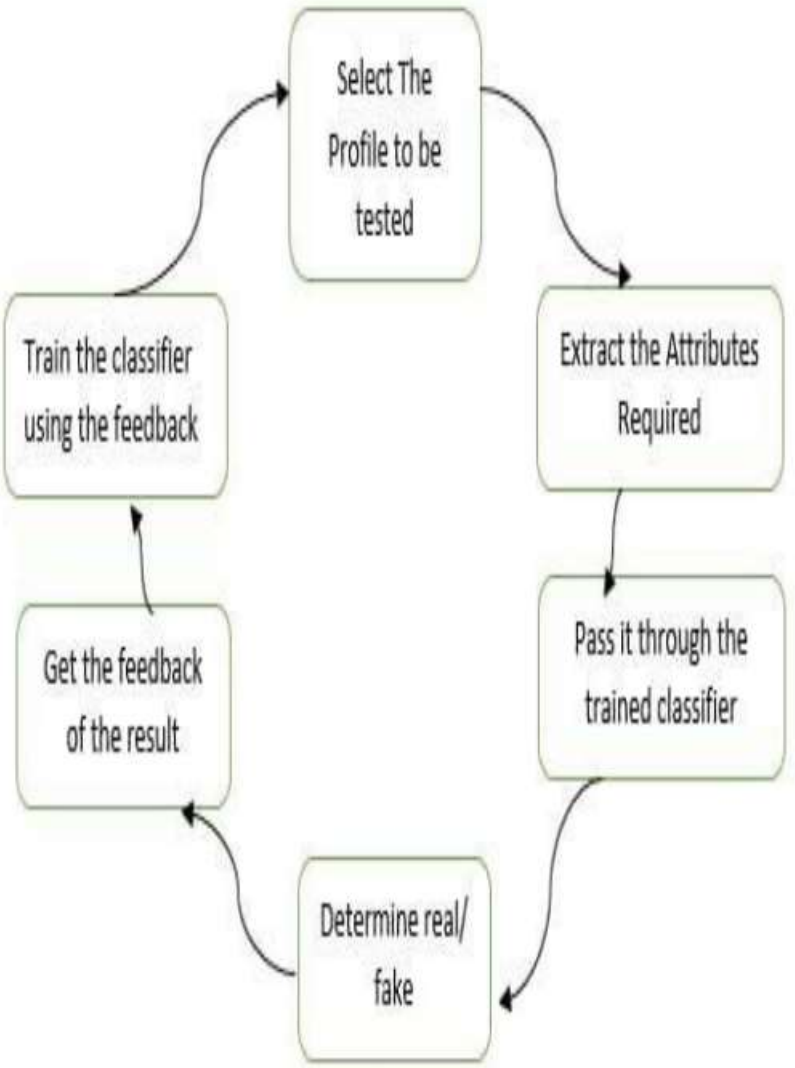
In our approach, we use artificial neural networks and machine learning to assess the likelihood that a friend request is genuine or not. To keep both old and new phoney data profiles, we use Microsoft Excel. The data is subsequently kept by the algorithm in a data frame. A training set and a testing set will be created from this data collection. To train our model, we would want a set of data from the social media platforms. The attributes we use for the training set when determining whether a profile is false are the following: Account age, Gender, User age, Link in the description, Number of messages exchanged, Number of friend requests sent, Entered location, Location by IP, and Fake or Not. Each of these variables is evaluated before being given a value. For the gender parameter, for instance, a value of (1) is assigned to the training set for Gender if it can be identified whether the profile is a female or a man. Other parameters are subjected to the same procedure. We also take into account a person's country of origin.

ADVANTAGES

• In this research, we came up with a framework through which we can detect a fake profile using machine learning algorithms so that people's social lives are secured. These metrics include:
(I) social reputation; (ii) global engagement; (iii) subject engagement; (iv) likeability; and (v) credibility.

ARCHITECTURE

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 4, April : 2023



4 METHODOLOGIES

Supervised Machine Learning algorithm can be broadly classified into Regression and Classification Algorithms.

Regression Algorithms: Regression finds correlations between dependent and independent variables. Regression algorithms therefore aid in the prediction of continuous variables such as real estate prices, economic trends, climatic patterns, oil and gas prices (a crucial job in today's world!), etc.

Linear Regression

- Decision Tree Regressor
- Random Forest Regressor

Linear Regression:

One of the simplest and most widely used Machine Learning methods is linear regression. It is a mathematical technique for performing predictive analysis. For continuous/real/numeric variables like sales, salary, age, and product price, among others, linear regression generates predictions. The linear regression algorithm, also known as linear regression, demonstrates a linear connection between a dependent (y) and one or more independent (y) variables. Given that linear regression demonstrates a linear connection, it can be used to determine how the dependent variable's value changes as a function of the independent variable's value. The connection between the variables is represented by a sloping straight line in the linear regression model.

Decision Tree Regression:

Decision tree regression can be used to perform non-linear regression in machine learning. The decision tree regression algorithm's primary job is to divide the information into more manageable chunks. The values of all data points that relate to the issue statement are plotted using the subsets of the dataset. This algorithm divides the data collection into decision and leaf nodes, producing a decision tree. When the data collection has not undergone enough change, ML experts favour this model. One should be aware that even a small shift in the data can have a significant impact on the decision tree's structure. Additionally, one should avoid over-pruning the decision tree regressors. since there won't be enough remaining end nodes to make the forecast. One should not overly prune the decision tree regressors in order to have numerous end nodes (regression output values). This develops a model with a tree-like structure that can forecast data in the future and generate useful ongoing output.

Random Forest Regression Algorithm:

Another popular method for non-linear regression in machine learning is random forest. A random forest employs multiple decision trees to predict the outcome as opposed to decision tree regression (single tree). With the help of this algorithm, a decision tree is constructed using k randomly chosen data points from the provided dataset. The worth of any new data point is then predicted using a number of decision trees. A random forest algorithm will forecast multiple output values because there are numerous decision trees. To determine the final result for a new data point, you must discover the average of all the predicted values. This occurs as a result of the numerous decision trees that must be mapped using this method. more processing capacity. Trees run parallel; it's a bagging method, not a boosting technique. i.e., there is no interaction between these trees as you construct trees.

Classification Algorithms:

An algorithm called classification discovers functions to categorise the dataset into groups based on different criteria. On the basis of what it learns from the training dataset, a computer programme divides the data into different groups.

- Logistic Regression
- Decision Tree Classifier
- Random Forest Classifier
- KNN - K Nearest Neighbour

Logistic Regression:

Under the category of supervised learning is logistic regression. Using a predetermined collection of independent variables, it is used to predict the categorical dependent variable. In a categorical dependent variable, the outcome is predicted by logistic regression. As a result, the result must be a discrete or categorical number. Rather than providing the precise values of 0 and 1, it provides the probabilistic values that fall between 0 and 1In logistic regression, we fit a "S" shaped logistic function, which forecasts two maximum values, rather than a regression line. (0 or 1). By using various kinds of data to categorise the observations, logistic regression can be used to quickly determine the most efficient factors that were used for classification.

Classification Algorithm (Decision Tree):

A supervised learning method called a decision tree can be used to solve classification and regression problems, but it is typically favoured for doing so. It is a tree-structured classifier, where internal nodes stand in for a dataset's features, branches for the decision-making process, and each child node for the classification result. The Decision Node and Leaf Node are the two elements in a decision tree. While Leaf nodes are the results of decisions and do not have any additional branches, Decision nodes are used to make decisions and have numerous branches. It is a graphical representation for obtaining all feasible answers to a decision or issue based on predetermined conditions. Since it functions like a tree, it is known as a judgement tree. begins with the base node and grows on additional branches to create a structure resembling a tree. The CART algorithm, which means for Classification and Regression Tree algorithm, is used to construct a tree. It

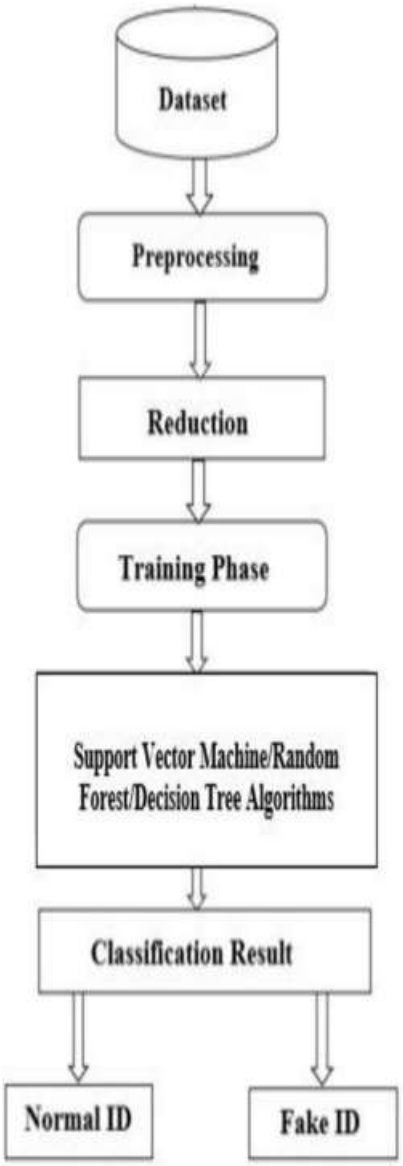
can be applied to classification and regression.

Random Forest Algorithm:

Popular machine learning algorithm Random Forest is a part of the guided learning methodology. In order to increase the dataset's predictive accuracy, a classifier called Random Forest uses multiple decision trees on different segments of the input data. It can be applied to ML issues involving both classification and regression. It is founded on the idea of ensemble learning, which is a method of combining various classifiers to address complex issues and enhance model performance. Even for the large dataset, it operates effectively and predicts the outcome with a high degree of accuracy. When a significant amount of the data is absent, accuracy can still be maintained.**K-Nearest Neighbour (KNN)**

Algorithm:

.

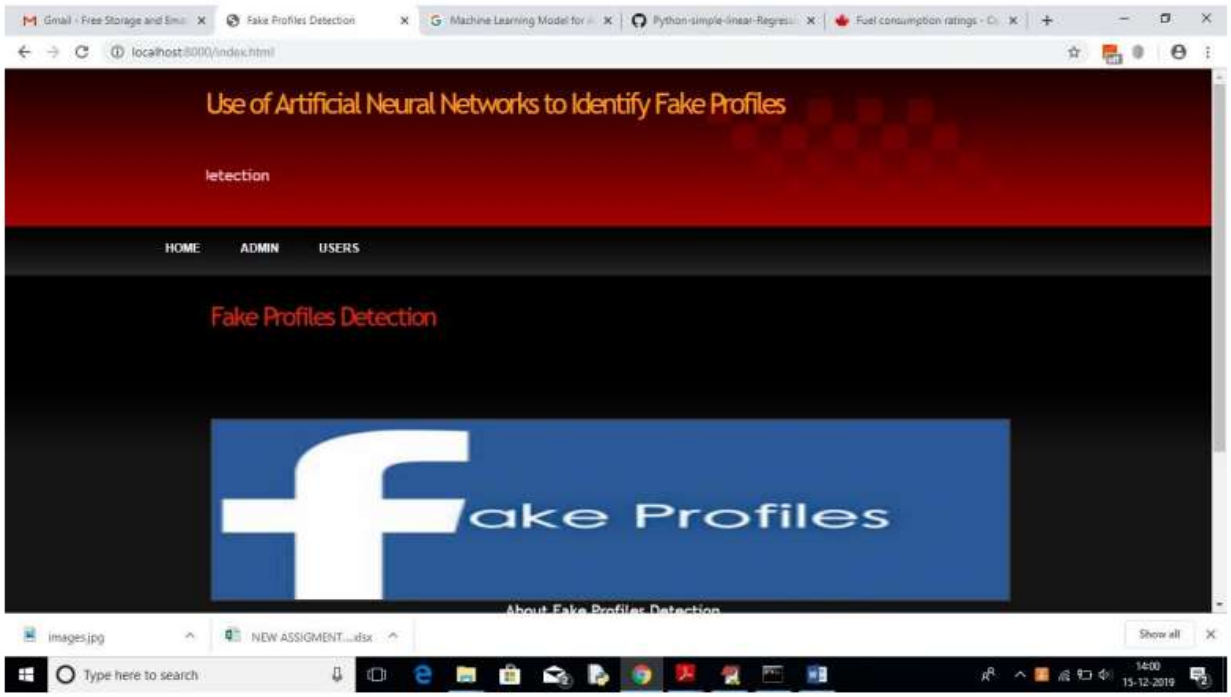


5 **RESULTS AND DISCUSSION**

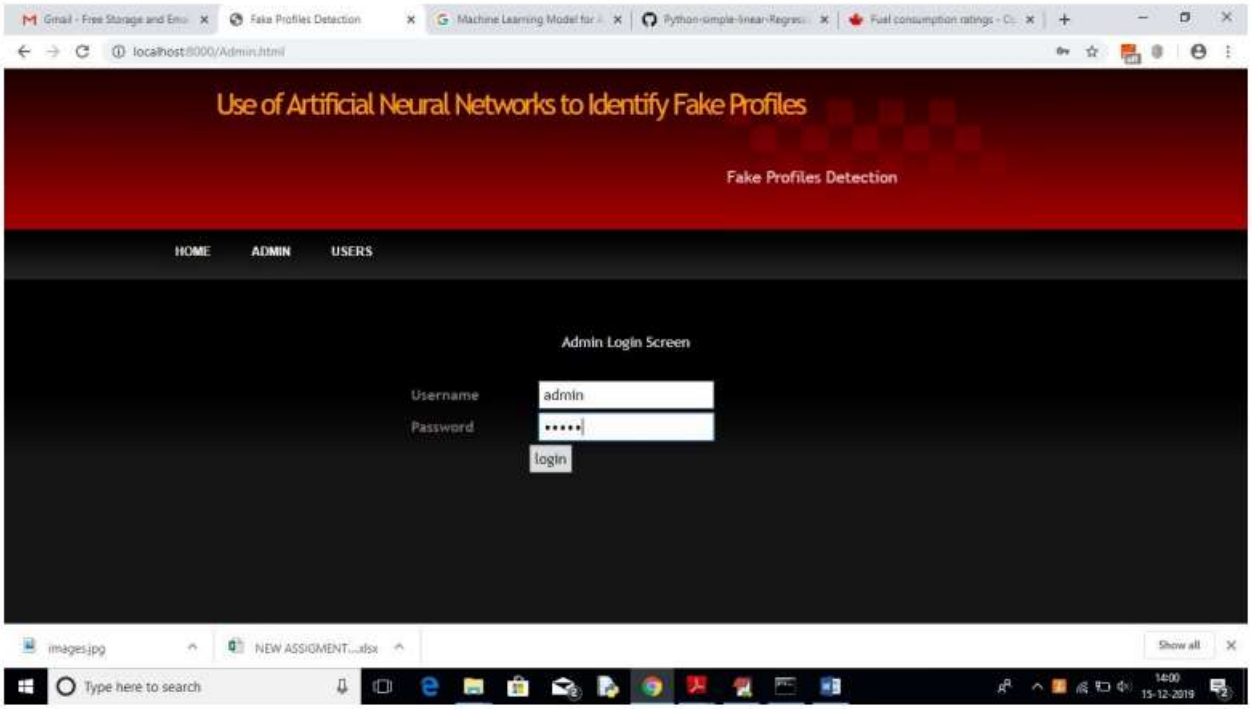
SCREENSHOTS

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 4, April : 2023

Deploy this application on DJANGO server and then run in browser enter URL as 'http://localhost:8000/index.html' to get below screen

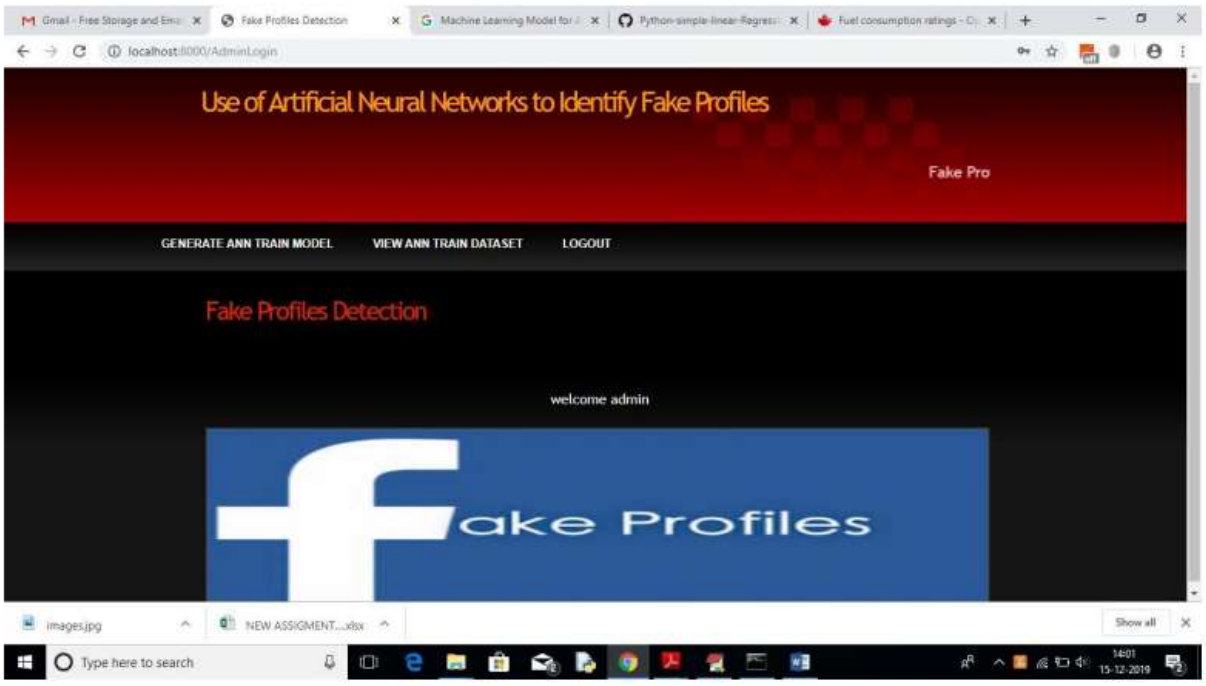


In above screen click on 'ADMIN' link to get below login screen

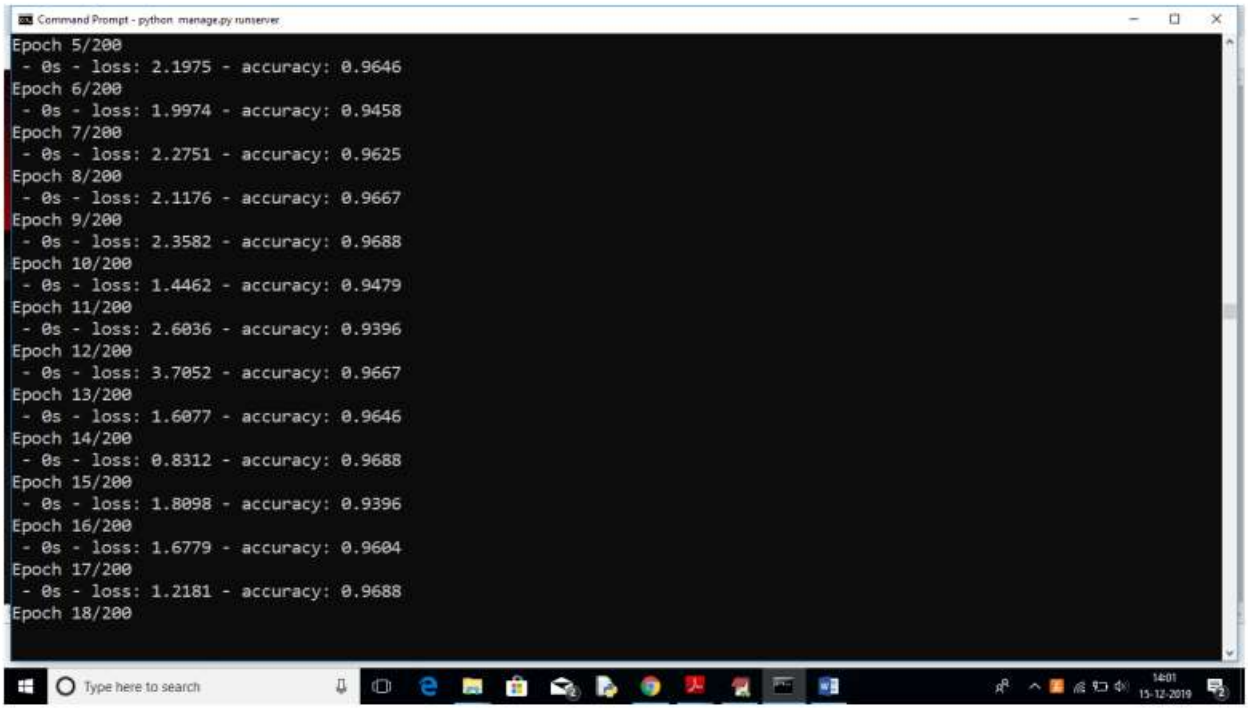


In above screen enter admin and admin as username and password to login as admin. After login will get below screen

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 4, April : 2023

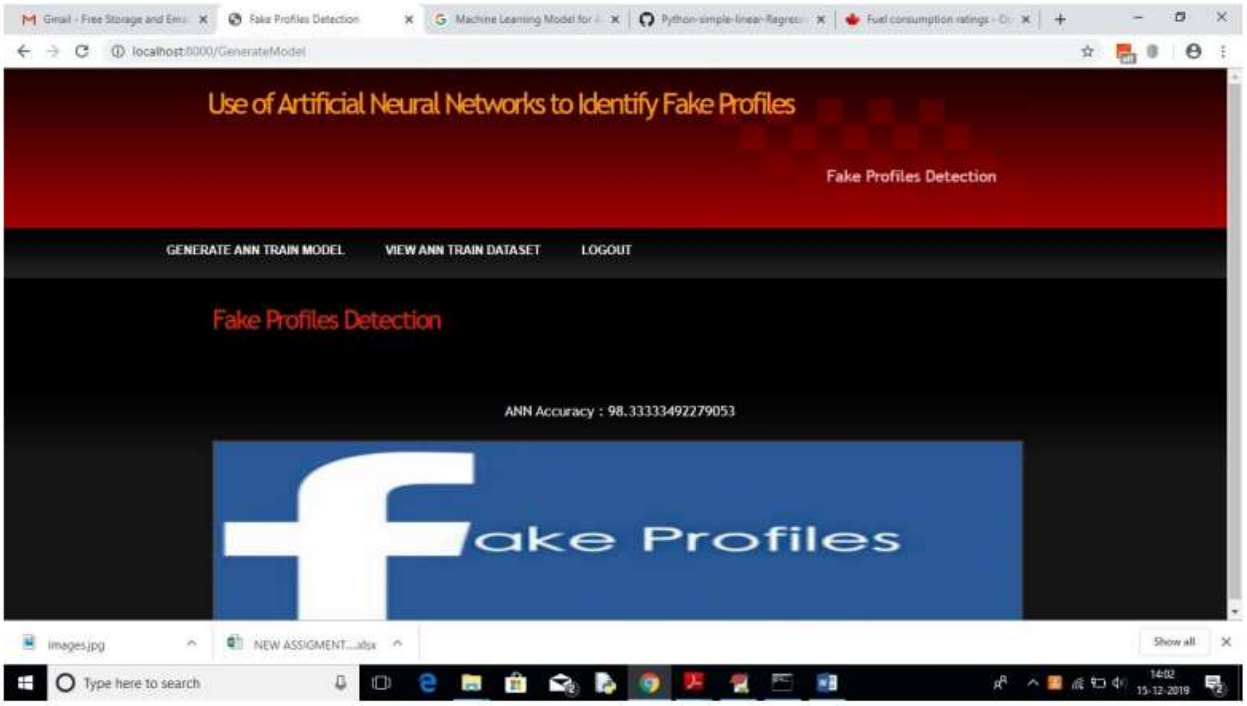


In above screen click on 'Generate ANN Train Model' to generate training model on dataset. After clicking on that link you can see server console to check ANN processing details with accuracy

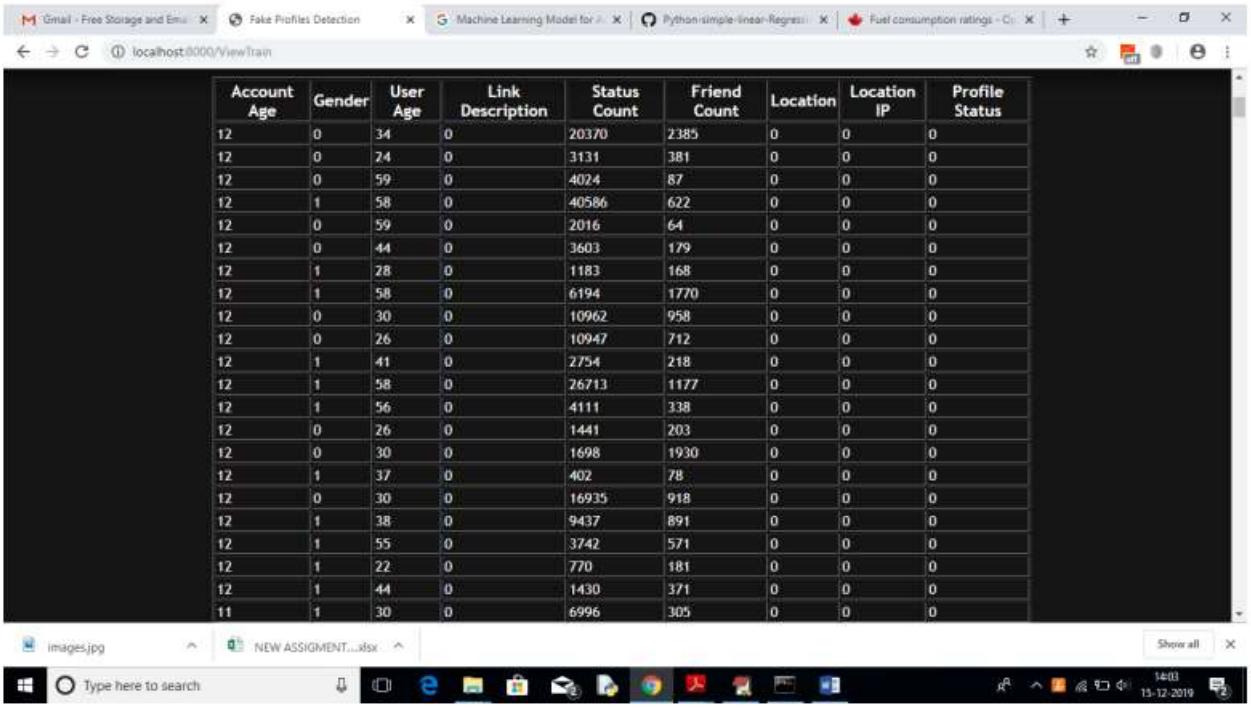


In above black console we can see all ANN details.

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 4, April : 2023



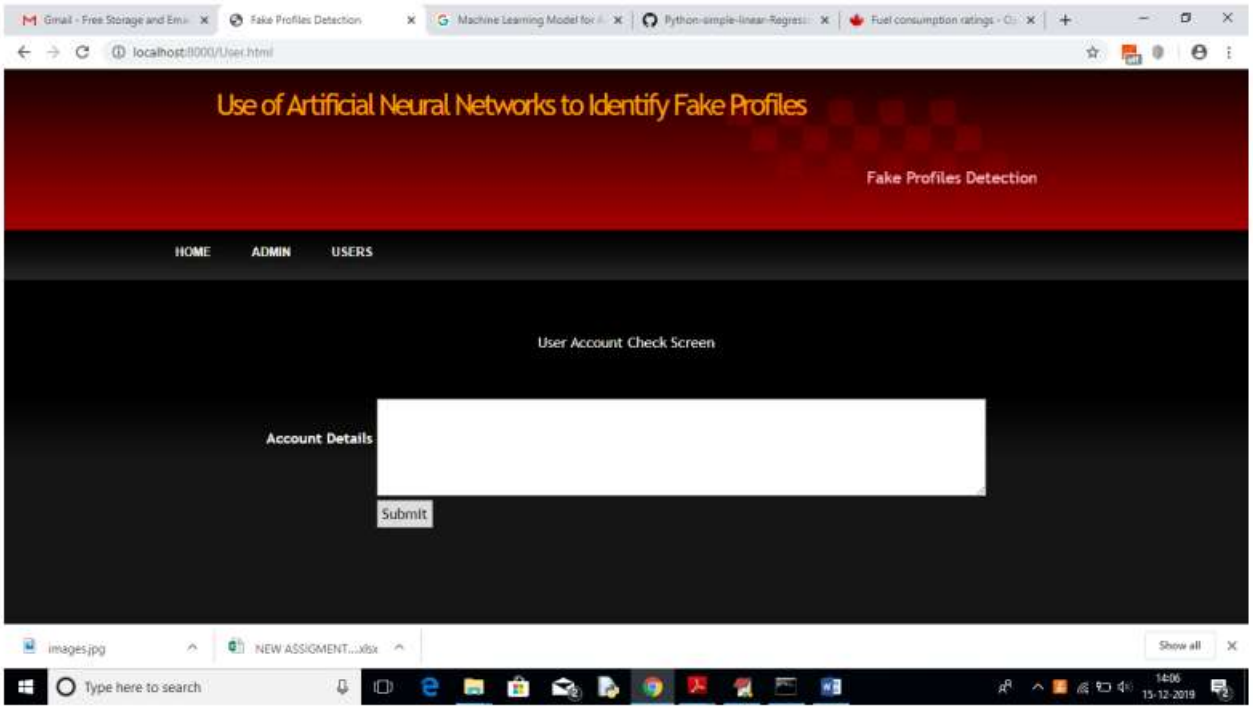
In above screen we can see ANN got 98% accuracy to train all Facebook profile. Now click on 'View Ann Train Dataset' link to view all dataset details



Account Age	Gender	User Age	Link Description	Status Count	Friend Count	Location	Location IP	Profile Status
12	0	34	0	20370	2385	0	0	0
12	0	24	0	3131	381	0	0	0
12	0	59	0	4024	87	0	0	0
12	1	58	0	40586	622	0	0	0
12	0	59	0	2016	64	0	0	0
12	0	44	0	3603	179	0	0	0
12	1	28	0	1183	168	0	0	0
12	1	58	0	6194	1770	0	0	0
12	0	30	0	10962	958	0	0	0
12	0	26	0	10947	712	0	0	0
12	1	41	0	2754	218	0	0	0
12	1	58	0	26713	1177	0	0	0
12	1	56	0	4111	338	0	0	0
12	0	26	0	1441	203	0	0	0
12	0	30	0	1698	1930	0	0	0
12	1	37	0	402	78	0	0	0
12	0	30	0	16935	918	0	0	0
12	1	38	0	9437	891	0	0	0
12	1	55	0	3742	571	0	0	0
12	1	22	0	770	181	0	0	0
12	1	44	0	1430	371	0	0	0
11	1	30	0	6996	305	0	0	0

In above screen we can see all train data and scroll down to view all records. Now ANN train model is ready and you can logout and click on 'User' link to get below screen.

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 4, April : 2023



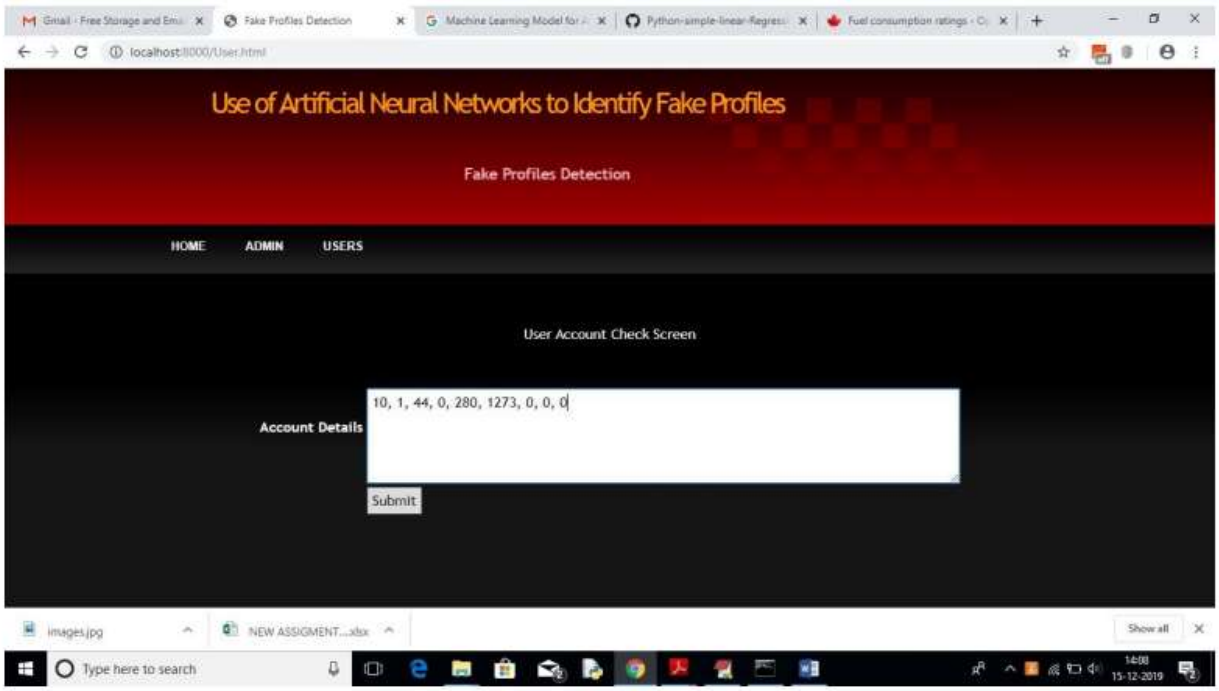
In above screen enter some test account details to get prediction/identification from ANN. You can use below records to check

10, 1, 44, 0, 280, 1273, 0, 0

10, 0, 54, 0, 5237, 241, 0, 0

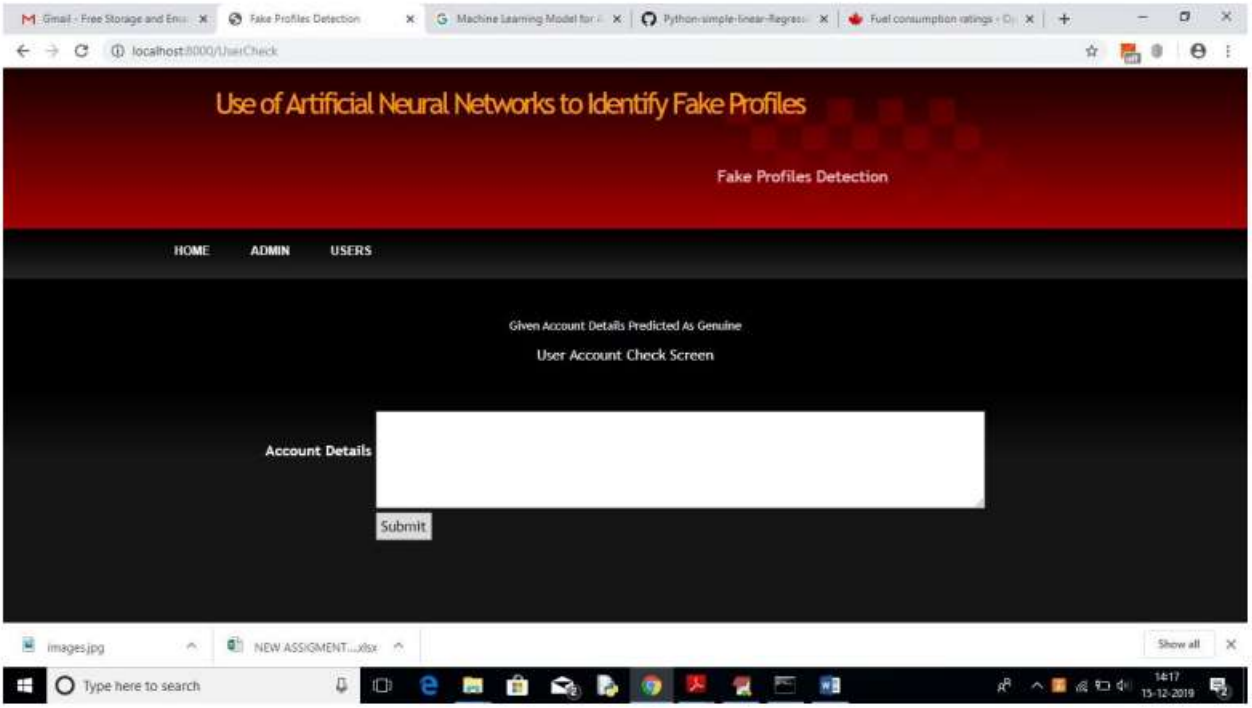
7, 0, 42, 1, 57, 631, 1, 1

7, 1, 56, 1, 66, 623, 1, 1

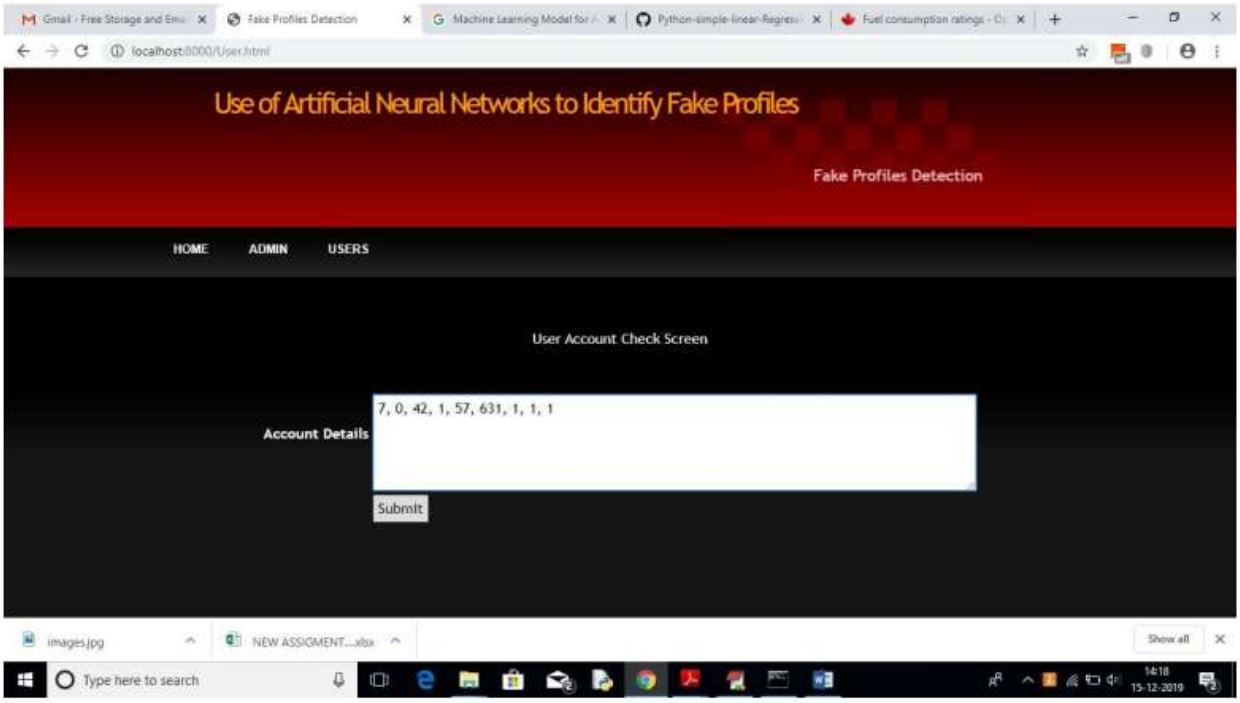


Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 4, April : 2023

For above input will get below result

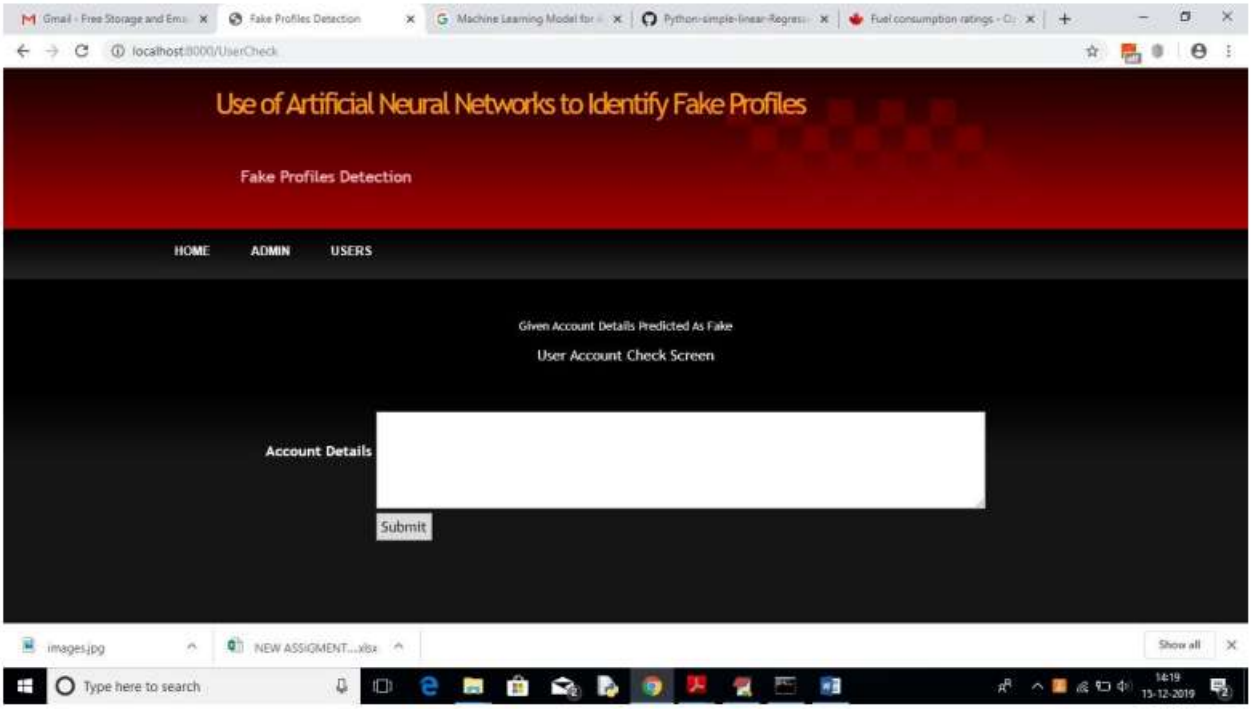


In above screen we can see the result predicted as genuine account



For above account details we got below result

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 4, April : 2023



In above screen we got result as fake for given account .

6. CONCLUSION AND FUTURE WORK

To determine whether or not a friend request is genuine, we employ machine learning, specifically an artificial neural network. Each equation is run through a Sigmoid function at each neuron (node). We make use of Facebook's or other social networks' training data sets.

By reducing the final cost function, modifying each neuron's weight, and back propagating patterns of bot activity, the described deep learning algorithm would be able to learn new patterns of behaviour.

We have provided a system that, when combined with Random Forest Classifier, allows us to identify phoney profiles in any online social network with an efficiency of up to 95%. By processing the posts and the profiles using NLP methods and neural networks, fake profile identification can be enhanced. We hope to categorise profiles in the future by including profile images as a feature. bias.

This project's main drawbacks are that it only uses visible data and has no real-time applications. Further jobs can be completed by running a CNN on the numerical, category, and profile photo data. Better outcomes might also occur from the addition of new parameters, the fusion of different models, and the creation of a real-time model. Depending on their size or specific significance in the recognition process, the areas in the model and data may be given varying degrees of prominence. It would be simpler to discover areas where exceedingly complicated problems, such as those that occasionally appear and the latter, must be located using this technique, for example. Although complicated, these hybrid models should produce better results. Combining these methods occasionally, though, might not make a big difference in the outcome. When that happens, the model will be ready for more social networking platforms like LinkedIn, Snap Chat, We Chat, QQ, etc.

7. REFRENCES

[1] S. Vidros, C. Kolias, G. Kambourakis, and L. Akoglu, “Automatic Detection of Online Recruitment Frauds: Characteristics, Methods, and a Public Dataset”, Future Internet 2017, 9, 6; doi:10.3390/fi9010006.

[2] B. Alghamdi, F. Alharby, “An Intelligent Model for Online Recruitment Fraud Detection”, Journal of Information Security, 2019, Vol 10, pp. 155176, https://doi.org/10.4236/iis.2019.103009.

[3] Tin Van Huynh1, Kiet Van Nguyen, Ngan Luu-Thuy Nguyen1, and Anh Gia-Tuan Nguyen, “Job Prediction: From Deep Neural Network Models to Applications”, RIVF International Conference on Computing and Communication Technologies (RIVF), 2020.

[4] Jiawei Zhang, Bowen Dong, Philip S. Yu, “FAKEDETECTOR: Effective Fake News Detection with Deep Diffusive Neural Network”, IEEE 36th International Conference on Data Engineering (ICDE), 2020.

[5] Scanlon, J.R. and Gerber, M.S., “Automatic Detection of Cyber Recruitment by Violent Extremists”, Security Informatics, 3, 5, 2014, https://doi.org/10.1186/s13388-014-0005-5

[6] Y. Kim, “Convolutional neural networks for sentence classification,” arXiv Prepr. arXiv1408.5882, 2014.

[7] T. Van Huynh, V. D. Nguyen, K. Van Nguyen, N. L.-T. Nguyen, and A.G.- T. Nguyen, “Hate Speech Detection on Vietnamese social media Text using the Bi-GRU-LSTM-CNN Model,” arXivPrepr. arXiv1911.03644, 2019.

[8] P. Wang, B. Xu, J. Xu, G. Tian, C.-L. Liu, and H. Hao, “Semantic expansion using word embedding clustering and convolutional neural network for improving short text classification,” Neurocomputing, vol. 174, pp. 806814, 2016.

[9] C. Li, G. Zhan, and Z. Li, "News Text Classification Based on Improved BiLSTM-CNN," in 2018 9th International Conference on Information Technology in Medicine and Education (ITME), 2018, pp. 890-893.
[10] K. R. Remya and J. S. Ramya, "Using weighted majority voting classifier combination for relation classification in biomedical texts," International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT), 2014, pp. 1205-1209.
[11] Yasin, A. and Abuhasan, A. (2016) An Intelligent Classification Model for Phishing Email Detection. International Journal of Network Security& Its Applications, 8, 55-72. https://doi.org/10.5121/imsa.2016.8405
[12] Vong Anh Ho, Duong Huynh-Cong Nguyen, Danh Hoang Nguyen, Linh Thi-Van Pham, Duc-Vu Nguyen, Kiet Van Nguyen, and Ngan Luu Thuy Nguyen."Emotion Recognition for Vietnamese Social Media Text", arXivPrepr. arXiv:1911.09339, 2019.
[13] Thin Van Dang, Vu Duc Nguyen, Kiet Van Nguyen and Ngan Luu Thuy Nguyen, "Deep learning for aspect detection on vietnamese reviews" in In Proceeding of the 2018 5th NAFOSTED Conference on Information and Computer Science (NICS), 2018, pp. 104-109.
[14] Li, H.; Chen, Z.; Liu, B.; Wei, X.; Shao, J. Spotting fake reviews via collective positive-unlabeled learning. In Proceedings of the 2014 IEEE International Conference on Data Mining (ICDM), Shenzhen, China, 14-17 December 2014; pp. 899-904.
[15] Ott, M.; Cardie, C.; Hancock, J. Estimating the prevalence of deception in online review communities. In Proceedings of the 21st international conference on World Wide Web, Lyon, France, 16-20 April 2012; ACM: New York, NY, USA, 2012; pp. 201-210.
[16] Nizam ani, S., Memon, N., Glasdam, M. and Nguyen, D.D. (2014) Detection of Fraudulent Emails by Employing Advanced Feature Abundance. Egyptian Informatics Journal, Vol.15, pp.169-174