International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

Fake Instagram Profile Identification And Classification Using Machine Learning

1. Mr K. Praveen Kumar, 2. L. Sampadha,3. Kasa Mamatha,4. G. Sai Prakash Reddy

1(Associate Professor, Sphoorthy Engineering College,Hyderabad

Email: Praveen.koram@gmail.com)

2(CSE Department, Sphoorthy Engineering College, Nadergul

Email: sampadhareddy2001@gmail.com

Kasamamatha2304@gmail.com

Prakashreddy7544@gmai.com)

1. ABSTRACT-

With the increase in Internet usage, Instagram is now considered a very important platform for advertising marketing and social interaction. It is used by millions of users but, some users tend to misuse the platform by creating false identities. In recent years though Internet is a boon, online social networks are susceptible to threats by cyber criminals and spammers. Moreover, the popularity of social media users is determined by followers and hence users resort to different wrong means to promote increased profile followers. Researchers have offered a lot of feasible solutions for social media applications. Fake engagement is one of the significant problems in Online Social Networks (OSNs) which is used to increase the popularity of an account in an inorganic manner. The detection of fake engagement is crucial because it leads to loss of money for businesses, wrong audience targeting in advertising, wrong product predictions systems, and unhealthy social network environment. This study is related with the detection of fake and automated accounts which leads to fake engagement on Instagram. Prior to this work, there were no publicly available dataset for fake and automated accounts. For this purpose, two datasets have been published for the detection of fake and automated accounts. For the detection of these accounts, machine learning algorithms like Naive Bayes, Logistic Regression, Support Vector Machines and Neural Networks are applied. Additionally, for the detection of automated accounts, cost sensitive genetic algorithm is proposed to handle the unnatural bias in the dataset. To deal with the unevenness problem in the fake dataset, Smotenc algorithm is implemented. For the automated and fake account detection datasets, 86% and 96% classification accuracies are obtained, respectively.

KEYWORDS: ***Logistic Regression, Random Forest Algorithm, median imputation, Maximum likelihood estimation, k cross validation, overfitting, out of bag data, recall, identity theft, Angler phishing***

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

2.INTRODUCTION-

Online Social Networking has grown extremely throughout the last few years. Online Social networks such as Facebook, Instagram, RenRen, Linked In, Google + have become increasingly popular over the last few years. People use online social network to keep in touch with each other's, share news, organize events and even advertisement of their own e–businesses. For the period between 2013 and 2019 around 3.14 million U.S dollars have been spent on sponsoring political ads on Facebook by nonprofit organizations. The continues growing Facebook community more than 2.3 billion monthly active users with an increase of 12% on a year –over-year basis.

In the second social networking sites Instagram has reported about one billion of subscribers with 337 million monthly active users.online social network has also attracted interest of researchers for mining and analyzing their massive amount of data, exploring and studying users behaviors as well as detecting abnormal activities In scientist have made a study to predict analyze and explain customer loyalty of social media online community. The online social network operator can increase the credibility of user metrics and enable third parties to consider user account. In the present generation, information security and privacy are among the primary requirements of social network users, maintaining security and privacy are more important. As recently banks and financial institutions have started to analyze the loan of Facebook and Instagram accounts. The open nature of online social networks the amount of personal data for a subscriber to vulnerable attack. In this paper, a support vector machine classification algorithm has been used. The support vector machine algorithm uses fewer features hence it can correctly classify about 98% of the accounts of our provided training data set. We also validated the detection performance of our classifier classifies that the account is fake or real. In the present generation, information security and privacy are among the primary requirements of social network users, maintaining security and privacy are more important

In the present generation, information security and privacy are among the primary requirements of social network users, maintaining security and privacy are more important. As recently banks and financial institutions have started to analyze the loan of Facebook and Instagram accounts. The open nature of online social networks the amount of personal data for a subscriber to vulnerable attack. In this paper, a support vector machine classification algorithm has been used. The support vector machine algorithm uses fewer features hence it can correctly classify about 98% of the accounts of our provided training data set. We also validated the detection performance of our classifier classifies that the account is fake or real.

3. LITERATURE SURVEY-

Previously a lot of work has been done on other platforms like Facebook and Twitter, but not much work has been done for Instagram. Each of these Social Medias are different in terms of features that have to be considered, strategies used, etc. Few of the past work include [1] Worth its Weight in Likes: Towards Detecting Fake Likes on Instagram: This particular paper concentrates on analysing likes and identifying the genuine ones to reduce the effect of fake likes on Instagram influencer market. They used a simple feed-forward neural network Multi-Layer Perceptron (MLP) which obtained a precision about 83%. [2] Identifying Fake Profiles in LinkedIn: A number of features were considered to train the dataset using neural networks, SVMs, and principal component analysis. The precision rate achieved was 84%. [3] Detection of Fake Profiles in Social Media: This is a literature review to detecting fake social media accounts classified into the approaches aimed on analysing individual accounts. So our 2 proposed method is a novel approach in terms of the platform chosen and the algorithms used like Random Forest for classification.

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

4. IMPLEMENTATION-

Data pre-processing: The dataset features are presented in two types.

Categorical Features: Categorical Features has various categories of data for e. g languages, different tweets, profile colors.

Numerical Features: It has a numerical type of data.

Feature Extraction: In feature reduction phase has extracted the different features and reduce the dimensionality of features. Feature reduction has used four different techniques to reduce the dimensionality of features.

1. Principal Component Analysis (PCA)

2.Spearman's Rank-Order Correlation

3.Relevance and Redundancy analysis technique

4.Markov Blanket Technique

1.Principal Component Analysis (PCA): PCA is a dimension reduction technique this is used to reduce feature vector dimensions. It finds the top number of features that best describe the data and covers as much variance of it, unnecessary features by assigning a lower weight so they did not impact on the data mining process.

2. Spearman's Rank-Order Correlation: Spearman's Rank-Order Correlation is a type of feature selection filtering method. It measures the strength and direction of the monotonic relationship between two variables P and Q.

3.Relevance and Redundancy analysis technique: Relevance and Redundancy analysis technique is used for feature selection. In Relevance and Redundancy analysis technique Spearman's Rank-Order Correlation Was used to eliminate all pairs of features and user input as selected features.

4. Markov Blanket Technique: The Markov Blanket Technique for a node B in a Bayesian network is the set of nodes composed of B's Parents of its children. In Markov Blanket Technique the Markov Blanket of a node is its neighboring node.

5.Wrapper Feature Selection using RANDOM FOREST: One of the well-known feature selection methods is wrapper feature selection methods, where different feature subsets are selected and qualified by a learning model. The features subset with the highest predictive performance would be selected. All subsets of a set can be found using bit manipulation, there will be 2n subsets for a given set, where n is the number of features F, in a set. For example, there will be 23 subsets for the set {1, 2, 3}. This method provides the best performing feature set for that particular learning model.

5. RESULT AND ANALYSIS-

5.1 Dataset Description

Once the proposed Random Forest classifier had been trained, their effectiveness was evaluated. Using confusion matrix we can describe the performance of the classification model.

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

profile pic	nums/leng	fullname w	nums/leng	name==use	description	external UF	private	#posts	#followers	#follows	fake
1	0.33	1	0.33	1	30	0	1	35	488	604	0
1	0	5	0	0	64	0	1	3	35	6	0
1	0	2	0	0	82	0	1	319	328	668	0
1	0	1	0	0	143	0	1	273	14890	7369	0
1	0.5	1	0	0	76	0	1	6	225	356	0
1	0	1	0	0	0	0	1	6	362	424	0
1	0	1	0	0	132	0	1	9	213	254	0
1	0	2	0	0	0	0	1	19	552	521	0
1	0	2	0	0	96	0	1	17	122	143	0
1	0	1	0	0	78	0	1	9	834	358	0
1	0	1	0	0	0	0	1	53	229	492	0
1	0.14	1	0	0	78	1	1	97	1913	436	0
1	0.14	2	0	0	61	0	1	17	200	437	0
1	0.33	2	0	0	45	0	1	8	130	622	0
1	0.1	2	0	0	43	0	0	60	192	141	0
1	0	2	0	0	56	0	1	51	498	337	0
1	0.33	2	0	0	86	0	1	25	96	499	0
1	0	1	0	0	97	0	1	188	202	605	0
1	0	3	0	0	46	0	1	590	175	199	0
1	0	2	0	0	39	0	1	251	223	694	0
1	0.5	1	0	0	0	0	1	0	189	276	0

Figure 1: Instagram dataset

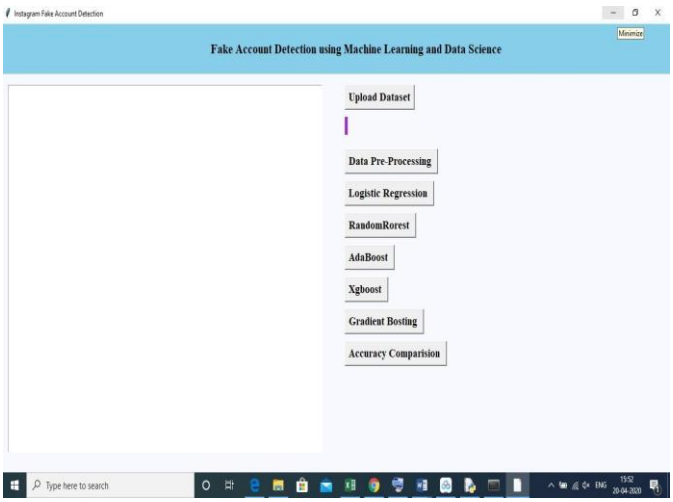

Figure 2: Fake account Detection using Machine learning techniques

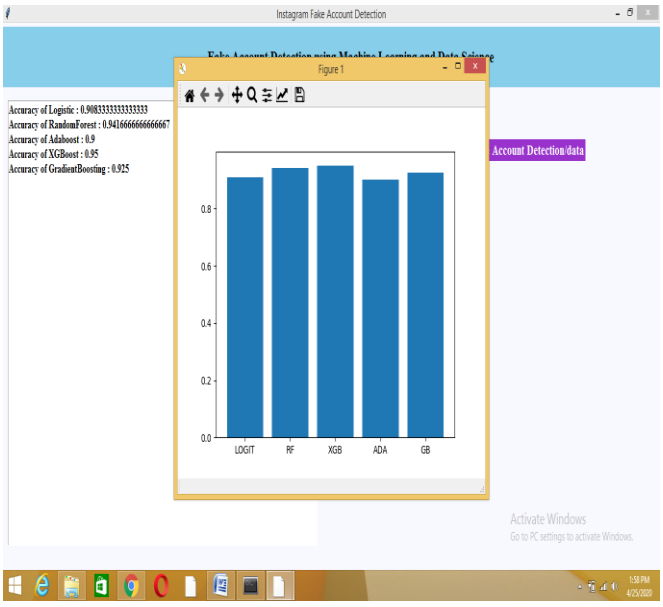

Figure 3: Comparison of machine learning techniques against Fake account detection

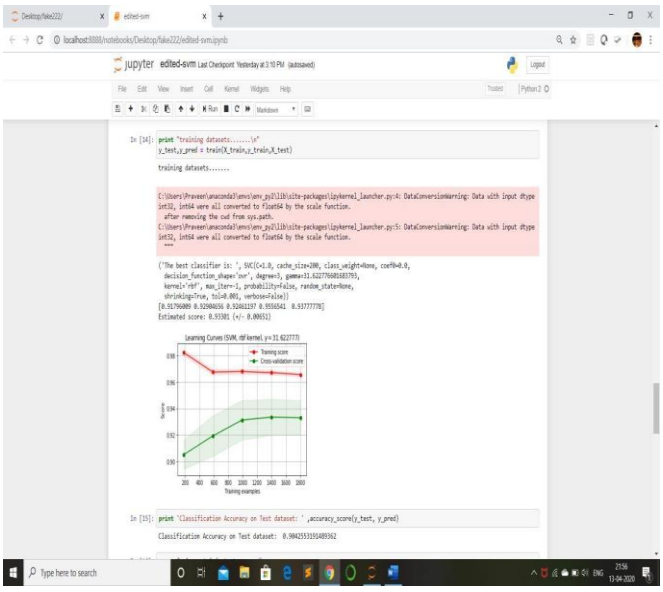

Figure 4: Among all algorithmic performance, XGboost performed well .

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

6. CONCLUSION-

In this paper, a new classification algorithm was proposed to improve detecting fake accounts on social networks, where the RANDOM FOREST trained model decision values were used to train a Neural Network model. To reach our goal we used dataset run it into the preprocessing phase where different feature reduction techniques were used to reduce the feature vector. In the classification phase, RANDOM FOREST learning algorithms were used. The results of the analyses showed that "RANDOM FOREST" has archived better accuracy results with all feature sets comparing with other classifiers, with classification accuracy around 98%. It was noticed that the Neural Network algorithm has the lowest classification accuracy compared with RANDOM FOREST. This occurred because the RANDOM FOREST algorithm reaches the global minimum of the optimized function, while the Neural Network using the gradient descent technique, and may reach the local minimum, not global minimum like RANDOM FOREST. It was also noticed that using the feature set provided by PCA, encountered a very low classification accuracy, while the wrapper RANDOM FOREST feature set achieves high classification accuracy. This happened because PCA performs dimension reduction and generates new features base on a linear combination of original features. But the wrapper RANDOM FOREST approach, and other feature selection techniques select the best set of original features, not a linear combination of all features. On the other words, feature selection selects the most effective original features, but PCA performs a linear combination of the original features event they are not effective. The Wrapper RANDOM FOREST feature set records a remarkable accuracy among the other feature selection technique sets because the Wrapper RANDOM FOREST technique not only selects the best features but also removes the redundancy.

7. ACKNOWLEDGEMENT-

This work has been carried out as part of our academic project to be submitted to the university. In this project work, we got the guidance and all inputs from our internal guide Mr. K. Praveen Kumar and we are thankful to our guide for his constant support and encouragement without which, the paper could not be completed.

8. REFERENCES-

[1] (2018) Detecting Fake Accounts on Social Media. [Online].Available:https://www.researchgate.net/publicatio n/330629456_Detecting_Fake_Accounts_on_Social_Media
[2] Political advertising spending on Facebook between 2014 and 2018.Internet draft. [Online].Available:https://www.statista.com/statistics/8913 27/ political-advertisingspending-facebook-by-sponsorcategory/
[3]Quarterly earning reports. Internet draft.[Online].Available:https://investor.fb.com/home/defau lt.a spx(2018)
[4] Statista.Instagram: number of monthly active users 2010- 2018. Internet draft. [Online]. Available: https://www.statista.com/statistics/282087/number-ofmonthly active-Instagram-users/
[5] R. Kaur and S. Singh, "A survey of data mining and social network analysis based anomaly detection techniques," Egyptian informatics journal, vol. 17, no. 2, pp. 199–216, 2016.
[6] L. M. Potgieter and R. Naidoo, "Factors explaining user loyalty in a social media-based brand community," South African Journal of Information Management, vol. 19, no. 1, pp. 1–9, 2017.
[7] Y. Boshmaf, D. Logothetis, G. Siganos, J. Ler´ıa, J. Lorenzo, M. Ripeanu, K. Beznosov, and H. Halawa, "´ Integro: Leveraging victim prediction for robust fake account detection in large scale osns," Computers & Security, vol. 61, pp. 142–168, 2016.
[8] (2013) Banque populaire dis-moicombiendamistu as sur Facebook, je tediraisi ta banquevataccorder un prłt. Internet draft. [Online].Available:http://bigbrowser.blog.lemonde.fr/2013/ 09/1/ popularitedis-moi-combien-damis-tu-as-sur-facebookje-te-diraisi-ta-banqueva-taccorder-un-pret/
[9] J. R. Douceur, "The sybil attack," in International workshop on peerto-peer systems. Springer, 2002, pp. 251– 260.
[10] (2012) Cbc.facebook shares drop on news of fake accounts. Internetdraft.[Online].Available:http: //www.cbc.ca/news/technolo gy/facebook-shares-droponnews-of-fake-accounts-1.1177067
[11] Y. Boshmaf, M. Ripeanu, K. Beznosov, and E. Santos-Neto, "Thwarting fake osn accounts by predicting their victims," in Proceedings of the 8th ACM Workshop on Artificial Intelligence and Security. ACM, 2015, pp. 81–89.
[12] (2018) Facebook publishes enforcement numbers for the first time.Internet draft [Online]. Available:https://newsroom.fb.com/news/2018/05/enforce mentnumbers/
[13] (2018) How concerned are you that there are fake accounts and bots on social media platforms that are used to try to sell you things or influence you? Internet draft. [Online]. Available: https://www.statista.com/statistics/881017/fake socialmedia-accounts-bots-influencing-selling-purchases-usa/
[14] (2012) Buying their way to Instagram fame. Internet draft. [Online].Available:www.nytimes.com/2012/08/23/fashion/t witterf ollowers-for-sale.html?smid=pl-share
[15] (2017) Welcome to the era of the bot as political boogeyman. Internet draft.[Online]. Available:https: //www.washingtonpost.com /news/politics/ wp/201 7/06/12/welcometo-the-era-of-the-bot-as-politicalboogeyman