IOSR Journal of Engineering (IOSR JEN) **www.iosrjen.org**
ISSN (e): 2250-3021, ISSN (p): 2278-8719
PP 58-64

Indentify the Real and Fake Activities from Streaming Data: An Overview

Miss. Amruta Patil[1] Prof. Sandip Mane[2]
PG Student KES's RIT Department of Computer Engineering, Islampur Dist Sangli MH India1
Prof. KES's RIT Department of Computer Engineering, Islampur Dist Sangli MH India[2]

Abstract: *Identity deception on social media platforms has become a growing problem with tremendous increase in number of accounts on social media. These fake identities can be used by offenders for various malicious purposes. This research aims at studying various classification techniques to classify fake vs. real identities on online social media platforms. Social networks have become popular due to the ability to connect people around the world and share videos, photos, and communications. One of the security challenges in these networks, which have become a major concern for users, is creating fake accounts. In this paper, a new model which is based on similarity between the users' friends' networks was proposed in order to discover fake accounts in social networks. Similarity measures such as common friends, cosine, Jaccard, L1-measure, and weight similarity were calculated from the adjacency matrix of the corresponding graph of the social network. To evaluate the proposed model, all steps were implemented on the Twitter dataset.*
Keywords: *Identity deception, Social media, Cyber crimes, Machine learning, Classification.*

I. Introduction

Social media platforms (such as Twitter, Facebook, LinkedIn, Instagram) are one of the crucial means for communication and information dissemination over the internet. Much can be learned about peoples habitat by analyzing their behavior over the social media. This helps offenders to commit various cyber crimes such as cyber bullying, skewing perceptions, misdirecting users to malicious websites, fraud, identity impersonation, dissemination of pornography, terrorist propaganda, to spread malware etc. Since identity deception provides means for offenders to commit such crimes it has become necessary to identify the fake identities over social media platforms. These fake identities can be created by bots or humans. The fake identities by bots generally target large group of peoples at a time. Whereas, fake accounts by humans generally target specific individual or limited number of peoples. This system represents an approach detect the fake identities created by humans on social media platforms. In order to detect identity deception we have applied Random Forest algorithm for machine learning. Also various preprocessing steps such as stop word removal, Porter's algorithm for stemming lexical analysis are applied on the data extracted through Twitter API. Accounts for bots are removed during data cleaning phase of preprocessing based on certain parameters such as presence of profile image, name etc. , accounts of known celebrities are also removed from the corpus. The fake accounts are created using two random human data generator APIs and validated based on tests such as Mann Whitney U Test and Chi Square Test.

II. Literature Survey

The classification in machine learning is based on training /learning from a training dataset. This learning can be categorized into three types: supervised, semi supervised and unsupervised learning. In supervised learning class labeled data is present at the beginning. In semi supervised learning some of the class labels are known .Whereas, in unsupervised learning class labels are not available. Once the training phase is finished features are extracted from the data based on term frequency and then classification technique is applied.

Estee et. al. [1] trained the classifier by applying previously used features for bot detection in order to identify fake accounts created by human on Twitter. The training is based on supervised learning. They have tested for 3 different classifiers i.e. Support Vector Machine (SVM) with linear kernel, Random Forest (RF) and Adaboost. For SVM, the svmLinear library in R software is used. Here the boundary based on feature vectors is created for classification. For RF model, the RF library in R software is used. RF model creates variations of trees and mode of class outcome is used to predict identity deception. For boosting model, the Adaboost function in R is used. Adaboost is used along with decision trees where each feature is assigned different weight to predict outcome. These weights are iteratively adjusted and output is evaluated for effectiveness of identity deception prediction at each iteration. This process is repeated until best result is obtained. Among these 3 classifiers RF reached the best result.

Senet. al. [2] performed supervised learning based on features obtained from FakeLike_data and RandLike_data. They have experimented with different classification algorithms such as Logistic Regression (LR), Random Forest (RF), Support Vector Machine (SVM) with RBF kernel, AdaBoost with Random Forest as base initiator,XGBoost and simple feed forward neural network i.e. Multi-Layer Perceptron (MLP) in order to detect the fake likes on instagram. For MLP they have used 2 hidden layers with 200 neurons each. Both layers use sigmoid activation function and output layer has a dropout of 0.2 in order to prevent over fitting. Here, MLP outperforms other methods.

Sedhaiet. al. [3] trained three different classifiers i.e. Naïve Bayes (NB), Logistic Regression (LR) and Random Forest (RF) using semi supervised Learning. These three classifiers use different classification techniques i.e. generative, discriminative and decision tree based classification models. The dataset used was from Twitter. Twitter Id is detected as spam if at least two classifiers of these three detect it as spam otherwise it is detected as ham. They have called this framework as S^3D (Semi Supervised Spam Detection).It gives best result as compared to any individual classifier.

Xiao et. al. [4] performed supervised learning in order to extract best feature set from the LinkedIn data. They have trained three classifiers i.e Logistic Regression (LR) with L1 regularization, Support Vector Machine (SVM) with radial basis kernel function and a Random Forest (RF) a nonlinear tree based ensemble learning method. Except regularization LR tries to find parameters using maximum likelihood criterion. Whereas with regularization there is tradeoff between fitting and having fewer variables to be chosen in the model. In this paper, they use L1 penalization to regularize the LR model. This technique maximizes the probability distribution of the class label y given a feature vector x and also reduces number of irrelevant features by using penalty term to bound the coefficients in L1 norm. The SVM looks for an optimal hyperplane as a decision function in high dimensional plane. While RF combines many weak classifiers (decision trees) to form strong classifier .For each decision tree training data is sampled and replaced to get training data of same size. Then at each node m features are selected at random to split decision tree. The common output class is considered as result of RF. Here RF gives the best result for classification of fake identities.

Ikramet. al. [5] used supervised two class SVM classifier implemented using scikit learn (an open source machine learning library for python) in order to automatically distinguish between like farm users from normal (baseline) users. They have compared this classifier with otherwell known supervised classifiers such as Decision tree, AdaBoost, K- Nearest Neighbor (KNN), Random Forest (RF) and confirmed that two class SVM is best in detecting like farm users on Facebook.

Dickerson et. al. [6] used Indian Election Dataset (IEDS) extracted from twitter for training. They tried for six high level classifiers such as SVM, Gaussian naïve Bayes, AdaBoost, Gradient Boosting, RF and Extremely Randomized Trees.The classifiers were built and trained on top of scikit-learn, a machine learning toolkit supported by INRIA and Google. Here, AdaBoost performed best on the reduced feature set and gradient boosting performed best on full feature set where reduced feature set involved only those features that did not involve sentiment analysis.

Ikramet. al.[7] System used supervised two class SVM classifier implemented using scikit learn (an open source machine learning library for python) in order to automatically distinguish between like farm users from normal (baseline) users. They have compared this classifier with otherwell known supervised classifiers such as Decision tree, AdaBoost, K- Nearest Neighbor (KNN), Random Forest (RF) and confirmed that two class SVM is best in detecting like farm users on Facebook.

Peddintiet. al. [8] developed a classifier that converts the four class classification problem into two binary classification problems: one that classifies each account as anonymous or non anonymous and other classifies each account as identifiable or non identifiable. The results of two classifiers are combined to classify each account as 'anonymous', 'identifiable' or 'unknown' for Twitter data. Both the binary classifiers use Random Forest (RF) with 100 trees as a base classifier. The choice of the classifier and number of trees is based on cross validation performance and out of bag error. These classifiers are also cost sensitive meta classifiers, where higher cost is imposed for misclassifying instances as anonymous or identifiable. The dataset used here was from Twitter.

Oentaryoet. al. [9] used supervised and unsupervised learning methods and tested for four prominent classifiers: naïve Bayes (NB), Random Forest (RF) and two instances of generalized linear model i.e. Support Vector Machine (SVM) and Logistic Regression (LR). This study involves Twitter dataset generated by users in Singapore and collected from 1 January to 30 April 2014 via the Twitter REST and streaming API. Here LR outperforms the other techniques and gives best result for classification of accounts as Broadcast bots, Consumption Bot, Spam Bot and Human.

Viswanathet. al. [10] uses unsupervised machine learning approach for training. The dataset used is from Facebook. They use K-Nearest Neighbors technique for this classification. In KNN data is classified based on majority vote of its neighbors, with test data being assigned to the class most common among its k nearest

neighbors where k is a positive integer typically small in value. The classification is done into the four classes i.e. Black market, Compromised, Colluding, Unclassified.

MosaYahyazadeh and Mahdi Abadi[11] fixed-width clustering algorithm, similarity algorithm. System addressthese issues and propose an online unsupervised method, called BotOnus, for botnet detection that does not require a priori knowledge of botnets. Itextracts a set of flow feature vectors from the network traffic at the end of eachtime period, and then groups them to some flow clusters by a novel onlinefixed-width clustering algorithm. Flow clusters that have at least two members,and their intra-cluster similarity is above a similarity threshold, are identifiedas suspicious botnet clusters, and all hosts in such clusters are identified as botinfected.

Muhammad Hassan Arif· Jianxin Li et.al[12] System investigates the performance of learning classifier systems(LCS), which are rule-based machine learning techniques, insentiment analysis of twitter messages and movie reviews,and spam detection from SMS and email data sets. In thisstudy, an existing LCS technique is extended by introducinga novel encoding scheme to represent classifier rules in orderto handle the sparseness in feature vectors, which are generatedusing the term frequency inverse document frequency ofword n-grams and sentiment lexicons.

Estee Van Der Walt and Jan Eloff[13]System trained the classifier by applying previously used features for bot detection in order to identify fake accounts created by human on Twitter. The training is based on supervised learning. They have tested for 3 different classifiers i.e. Support Vector Machine (SVM) with linear kernel, Random Forest (RF) and Adaboost. For SVM, the svmLinear library in R software is used. Here the boundary based on feature vectors is created for classification. For RF model, the RF library in R software is used. RF model creates variations of trees and mode of class outcome is used to predict identity deception. For boosting model, the Adaboost function in R is used. Adaboost is used along with decision trees where each feature is assigned different weight to predict outcome. These weights are iteratively adjusted and output is evaluated for effectiveness of identity deception prediction at each iteration. This process is repeated until best result is obtained. Among these 3 classifiers RF reached the best result.

Indira Senet. al.[14] System performed supervised learning based on features obtained from Fake Like_data and RandLike_data. They have experimented with different classification algorithms such as Logistic Regression (LR), Random Forest (RF), Support Vector Machine (SVM) with RBF kernel, AdaBoost with Random Forest as base initiator,XGBoost and simple feed forward neural network i.e. Multi-Layer Perceptron (MLP) in order to detect the fake likes on instagram. For MLP they have used 2 hidden layers with 200 neurons each. Both layers use sigmoid activation function and output layer has a dropout of 0.2 in order to prevent over fitting. Here, MLP outperforms other methods.

SurendraSedhai and AixinSun[15] System trained three different classifiers i.e. Naïve Bayes (NB), Logistic Regression (LR) and Random Forest (RF) using semi supervised Learning. These three classifiers use different classification techniques i.e. generative, discriminative and decision tree based classification models. The dataset used was from Twitter. Twitter Id is detected as spam if at least two classifiers of these three detect it as spam otherwise it is detected as ham. They have called this framework as S3D (Semi Supervised Spam Detection).It gives best result as compared to any individual classifier.

Cao Xiao, David Freeman and Theodore Hwa[16] System performed supervised learning in order to extract best feature set from the LinkedIn data. They have trained three classifiers i.e Logistic Regression (LR) with L1 regularization, Support Vector Machine (SVM) with radial basis kernel function and a Random Forest (RF) a nonlinear tree based ensemble learning method. Except regularization LR tries to find parameters using maximum likelihood criterion. Whereas with regularization there is tradeoff between fitting and having fewer variables to be chosen in the model. In this paper, they use L1 penalization to regularize the LR model. This technique maximizes the probability distribution of the class label y given a feature vector x and also reduces number of irrelevant features by using penalty term to bound the coefficients in L1 norm. The SVM looks for an optimal hyperplane as a decision function in high dimensional plane. While RF combines many weak classifiers (decision trees) to form strong classifier .For each decision tree training data is sampled and replaced to get training data of same size. Then at each node m features are selected at random to split decision tree. The common output class is considered as result of RF.

ID	Title of paper	Base classification technique	Other techniques tested	Dataset	Performance parameter
1.	Using Machine Learning to Detect Fake Identities : Bots vs. Humans [1]	Random Forest	SVM Linear,Adaboost	Twitterdataset	F1 score, PR-AUC, Accuracy
2.	Worth its Weight in Likes: Towards Detecting Fake Likes on Instagram [2]	Multi-Layer Perceptron	Logistic Regression, Random Forest,AdaBoost, XGBoost	Instagramdataset	Precision, Recall, AUC
3.	Semi-Supervised Spam Detection in Twitter Stream [3]	S3D (Naïve Bayes, Logistic Regression and Random Forest.)	Naïve Bayes , Logistic Regression, Random Forest	Twitter dataset	F1 score, Precision, Recall
4.	Detecting Clusters of Fake Accounts in Online Social Networks [4]	Random Forest	Logistic Regression,SVM	LinkedIn dataset	Recall, AUC
5.	Combating Fraud in Online Social Networks: Detecting Stealthy Facebook Like Farms [5]	SVM	Decision Tree, AdaBoost, KNN, Random Forest, Naïve Bayes	Facebook dataset	Precision , Recall, F1 Score
6.	Using Sentiment to Detect Bots on Twitter: Are Humans more Opinionated than Bots? [6]	AdaBoost, Gradient boosting	Gaussian naive Bayes, SVM, Random Forest, Extremely Randomized Trees	Twitter(Indian Election Dataset)	AUROC
7.	Combating Fraud in Online Social Networks: Detecting Stealthy Facebook Like Farms[7]	K- Nearest Neighbor (KNN),	Decision Tree, Random Forest (RF)	System taken data from UCI machine learning repository	Accuracy, semi structured dataset.
8	Mining Anonymity: Identifying Sensitive Accounts on Twitter [8]	Random Forests & Binary classifiers	_	Twitter dataset	Precision, Recall
9.	On Profiling Bots in Social Media [9]	Logistic Regression	SVM, Naïve Bayes, Random Forests	Twitter dataset (Singapore)	F1 Score, Precision, Recall,
10.	Towards Detecting Anomalous User Behaviour in Online Social Networks [10]	KNN	_	Facebook dataset	AUROC, TP rate, FP rate

III. Proposed System Overview

. The following figure shows the flow of the system. and results are evaluated based on performance metrics such as accuracy, precision, recall etc.

1. Data Acquisition: First of all the information for different Twitter accounts basedon certain parameters is extracted from Twitter using Twitter API.

2. Preprocessing: Then we will apply various preprocessing steps such as lexical analysis, stopword removal, stemming (Porters algorithm),index term selection and datacleaning in order to make our dataset proper.

2.1 Lexical analysis: Lexical analysis separates the input alphabet into,

1)Word characters (e.g. the letters a-z) and

2)Word separators (e.g space, newline, tab).

2.2 Stopword removal: Stopword removal refers to the removal of words that occurmost frequently in documents. The stopwords includes,

1) Articles (a, an, the,....)
2) Prepositions (in, on, of,...)
3) Conjunctions (and, or, but, if,....)
4) Pronouns (I, you, them, it....)
5) Possibly some verbs, nouns, adverbs, adjectives (make ,thing, similar....)

2.3 Stemming: Stemming replaces all the variants of a word with a single stem word. Variants include plurals, gerund forms (ing forms), third person suffixes, past tense suffixes, etc.).Example: connect: connects, connected, connecting, connection and so on. Here we will use Porter's algorithm for stemming.
2.4 Index term selection: Index term selection refers to the selection of features from large amount of available data which will be useful to classify the given data.
2.5 Data cleaning: During data cleaning step Bots are removed from the dataset based on certain parameters such as presence of name, profile image, number of followers, number of tweets etc. Also accounts of the known celebrities are removed from the given corpus.
3. Create fictitious accounts: Then fictitious accounts are created with the help of various random human data generator APIs and manually by us. The basis for creation of fictitious accounts is that the people generally lie on their age, gender ,image, location and the name most.

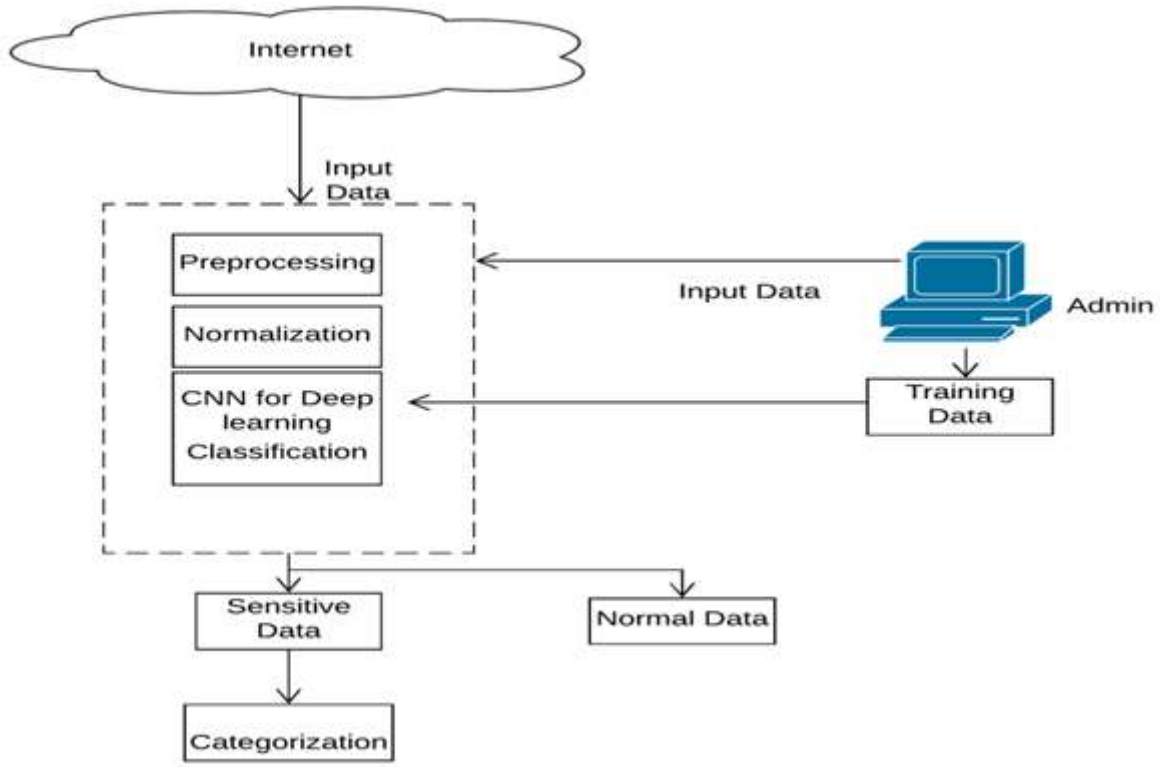

Figure 1: Proposed system architecture

4. Validate data: In an absence of deceptive human accounts and for the sake of the validity of the research, it was decided to ensure that the fabricated deceptive accounts are as far as possible aligned with the data contained in the original corpus. This was done to make the research results as realistic as possible. Most importantly, the following two statistical tests were employed to validate that the injected deceptive accounts are still representative of original mined corpus.
4.1 Mann-Whitney-U test: Mann-Whitney-U test proofs that the means of the two sets are similar per attribute.
4.2 Chi Square Test: The Chi Square test for independence proofs that the datasets are not correlated and therefore independent.
This means that both the deceptive and original corpus must have similar data and show the same distributions.
5. Then the deceptive accounts which pass the validation tests for fake accounts are injected into the system.
6. Create new features: The new features are created for classification purpose by manipulation of previously extracted features (for eg. taking ratios of features).
7. Supervised machine learning: Then supervised machine learning is applied in order to train the classifier. Here class labeled data is present at the beginning.
7.1 Random Forest: Random Forest algorithm for machine learning is applied to determine identity deception on social networks where multiple decision trees are created using randomly selected features from the feature set and the majority output class of all the decision trees is taken as output of the Random Forest.
8. The results are evaluated based on various performance metrics such as accuracy,F1 Score, PR-AUC..

IV. Algorithms Design

Algorithm 1 : Fuzzy Random Forest

Input : TrainFeature set {} which having values of numeric or string of train DB, TrainFeature set {} which having values of numeric or string of train DB, Threshold T, List L.

Output: classified all instances with weight.

Step 1 : Read all features from Test set using below

$$\text{TestFeature} = \sum_{j=1}^{n} (\text{T[j]})$$

Step 2: Read all features from Trainset using below

$$\text{TrainFeature} = \sum_{k=1}^{m} (\text{T[k]})$$

Step 3: Read all features from Trainset using below

Step 4 : Generate weight of both feature set

$$W = (\text{TrainFeature}, \text{TestFeature})$$

Step 5 : Verify Threshold

Selected_Instance= result = W > T ? 1 : 0;

Add each selected instance into L, when n = null

Step 6 : Return L.

Algorithm 2 : Weight Calculation Algorithm

Input: Attribute list from twitter data and policy base train DB.

Output: Each list with weight.

Here system have to find similarity of two lists:

$$\vec{a} = (a_1, a_2, a_3, \ldots) \text{ and } \vec{b} = (b_1, b_2, b_3, \ldots) \text{ ---------- (1)}$$

where a_n and b_n are the components of the vector (features of the document, or values for each word of the comment) and the n is the dimension of the vectors:

Step 1: Read each row R from Data List L

Step 2: for each (Column c from R)

Step 3: Apply formula (1) on c and Q

Step 4: Score=Calc(c,Q)

Step 5: calculate relevancy score for attribute list.

Step 6: assign each Row to current weight

Step 7: Categorize all instances

Step 8: end for end procedure

V. Results and Discussions

The below Figure 2 shows the result presented Classifier (Random Forest) is trained using supervised machine learning technique where we have class labeled data (injected fictitious accounts) is available for training the classifier. Here we test for three different cross validation techniques i.e 3 fold, 5 fold and 10 fold where 70 percent data is given for training and remaining 30 percent data goes for testing.

Results are evaluated based on various performance metrics such as accuracy, F1 Score, Precision and recall.

Accuracy = TP+TN / TP+TN+FP+FN

Precision = TP/ TP+FP

Recall = TP/ TP+FN

F1Score = 2 ∗ (Precision ∗ Recall) / (Precision + Recall)

Where,

TP = True Positive

TN = True Negative

FP = False Positive

FN = False Negative

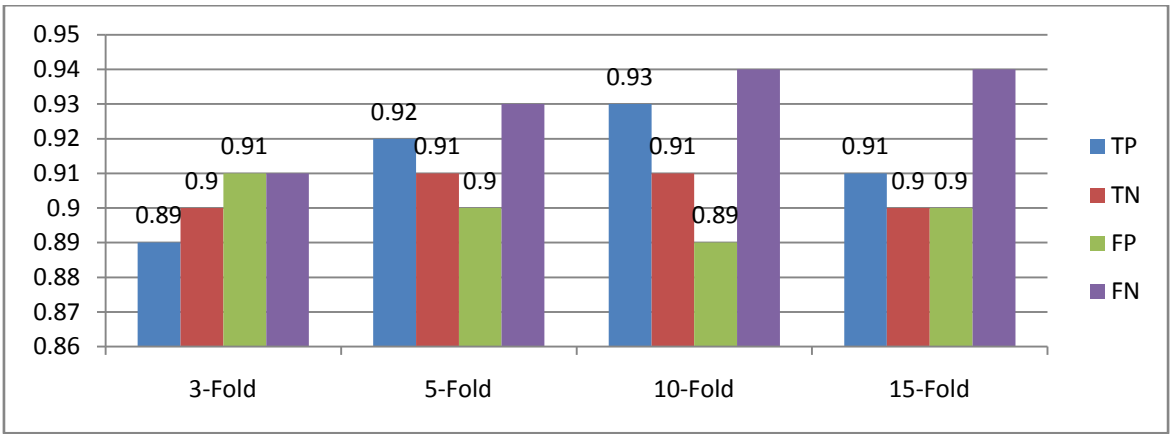

Figure 2: Time required in seconds with different support denominator with different dataset.

The above figure show the experimental analysis of proposed system confusion matrix of proposed system with various fold validation.

VI. Conclusion

Random Forest can be used to solve the problem of determining fake vs. real identities on social networks with maximum accuracy of 94.67 percent. The performance of given system varies with the dataset used for it. Also we found that 10 fold cross validation gives better results than that of 5 fold and 3 fold cross validation. Furthermore, accuracy can be increased in future by enrichment of features set and testing for other classification techniques such as deep learning (For e.g. Deep Convolutional Neural Networks) and evaluating it for different activation functions..

References

[1]. Estee Van Der Walt and Jan Eloff, "Using Machine Learning to Detect Fake Identities: Bots vs Humans," IEEE, 2018.
[2]. Indira Senet. al. "Worth its Weight in Likes: Towards Detecting Fake Likes on Instagram," ACM, 2018.
[3]. SurendraSedhai and Aixin Sun, "Semi-Supervised Spam Detection in Twitter Stream," IEEE , 2018.
[4]. Cao Xiao, David Freeman and Theodore Hwa , "Detecting Clusters of Fake Accounts in Online Social Networks," ACM , 2015.
[5]. Ikramet. al., "Combating Fraud in Online Social Networks: Detecting Stealthy Facebook Like Farms," ARXIV, 2016.
[6]. J. Dickerson, V. Kagan and V. Subhramanian "Using Sentiment to Detect Bots on Twitter: Are Humans more Opinionated than Bots?" IEEE, 2014.
[7]. Ikramet. al., "Combating Fraud in Online Social Networks: Detecting Stealthy Facebook Like Farms," ARXIV, 2016.
[8]. S. Peddinti, K. Ross and J. Cappos "Mining Anonymity: Identifying Sensitive Accounts on Twitter," ARXIV ,2016.
[9]. R. Oentaryoet. al. "On Profiling Bots in Social Media," ARXIV, 2016.
[10]. B. Viswanathet. al. "Towards Detecting Anomalous User Behaviour in Online Social Networks," USENIX, 2014.
[11]. M. Yahyazadeh and M. Abadi, "BotOnus: An online unsupervised method for botnet detection," The ISC International Journal of Information Security, vol. 4, pp. 51-62, 2012.
[12]. M. H. Arif, J. Li, M. Iqbal, and K. Liu, "Sentiment analysis and spam detection in short informal text using learning classifier systems," Soft Computing, pp. 1-11, 2017.
[13]. Estee Van Der Walt and Jan Eloff, "Using Machine Learning to Detect Fake Identities: Bots vs Humans," IEEE, 2018.
[14]. Indira Senet. al. "Worth its Weight in Likes: Towards Detecting Fake Likes on Instagram," ACM, 2018.
[15]. Surendra Sedhai and Aixin Sun, "Semi-Supervised Spam Detection in Twitter Stream," IEEE , 2018.
[16]. Cao Xiao, David Freeman and Theodore Hwa , "Detecting Clusters of Fake Accounts in Online Social Networks," ACM , 2015.