Volume-10,Issue-2, (Aug-2018)
ISSN (O) :- 2349-3585

A Survey on Classification Techniques to Determine Fake vs. Real Identities on Social Media Platforms

Priyanka Shahane[1], Deipali Gore[2]
[1]M.E. Scholar, [2]Assistant Professor, Computer Department, P.E.S. Modern College of Engineering, Pune

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.

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 people's 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. This paper presents the survey of various supervised, semi-supervised and unsupervised machine learning algorithms to classify fake vs. real identities on social media platforms.

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.

Sen et. 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.

Sedhai et. 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.

Ikram et. 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 other well 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.

Fuller et. al. [7] used "person of interest statements" officially known as a Form 1168 as a dataset ,was provided from law enforcement personal at participating military bases. Person of interest statements are official reports written by a subject or witness in an official investigation. While considering the creation of an automated decision support tool for deception detection, three common classification methods were utilized : Artificial Neural Networks, Decision Tree, and Logistic Regression. Among these methods Artificial Neural Network (ANN) reached the best performance. An ANN is a collection of nodes arranged in layers. It has three main layers input layer, hidden layer and output layer. The nodes in hidden layer combines inputs from previous layers into a single output value. This output is then passed on to the next layer. A weight is associated with each unit in the network, it is determined by training a network on portion of data. The network performance is then tested on a holdout sample.

Peddinti et. 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

ISSN:-2349-3585 |www.ijrdt.org

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.

Oentaryo et. 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.

Viswanath et. 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.

III. PERFORMANCE PARAMETERS

The performance depends upon the method of classification and dataset used for classification. The various performance parameters used in above literature survey are as follows:

F1 score is a combination of precision and recall it is calculated as ,

F1 Score: 2*(Precision * Recall / (Precision + Recall)

Precision can be calculated as,

Precision: TP / TP + FP

Where,

TP (True Positive) test result is one that detects the condition when the condition is present.

FP (False Positive) test result is one that detects the condition when the condition is absent.

Recall (Sensitivity/ TP Rate) can be calculated as,

Recall: TP/ TP + FN

Where,

FN (False Negative) test result is one that does not detect the condition when the condition is present.

Specificity (TN Rate) can be calculated as,

Specificity: TN/ TN + FP

Where,

TN (True Negative) test result is one that does not detect the condition when the condition is absent.

FP Rate can be calculated as,

FP Rate: FP/ FP + TN

Accuracy can be calculated as,

Accuracy: sum (abs (Expected Output – Actual Output)) / 2

PR-AUC: Precision-Recall Area Under Curve (PR-AUC) is a statistical value of the area under the precision-recall curve.

AUROC (AUC): It is a Area Under ROC (Receiver Operating Characteristic) Curve. ROC is created by plotting TP Rate against FP Rate.

IV. COMPARATIVE ANALYSIS

The following table gives the comparative analysis of various classification techniques in machine learning used to detect identity deception on social media platforms. In the above literature survey given in section I, each paper tests for various classification techniques for classification of fake vs. real identities on social media platforms. The best classification technique is then selected based on performance parameters given in section III.

ISSN:-2349-3585 |www.ijrdt.org

Table 1. Comparative analysis of classification techniques to detect identity deception on social media

Sr. No.	Paper	Best 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	Twitter dataset	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	Instagram dataset	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.	Decision Support for Determining Veracity via Linguistic-based Cues [7]	Neural Network	Decision Tree, Logistic Regression	Form 1168	False +, Sensitivity, Specificity, Accuracy
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

ISSN:-2349-3585 |www.ijrdt.org

V. CONCLUSION

From this survey we conclude that the problem of detecting identity deception on social media can be solved by using various machine learning techniques such as SVM, RF, LR, NB, MLP,ANN and so on. Among these techniques Random Forest (RF) reach the best performance with accuracy of 87.11 %. Also, we notice that the performance of the system varies with classification technique and dataset used. Furthermore, the performance of system can be increased by using other techniques such as Deep Learning with different activation functions in future.

REFERENCES

[1] Estee Van Der Walt and Jan Eloff, "Using Machine Learning to Detect Fake Identities: Bots vs Humans," IEEE, 2018.

[2] Indira Sen et. al. "Worth its Weight in Likes: Towards Detecting Fake Likes on Instagram," ACM, 2018.

[3] Surendra Sedhai 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] Ikram et. 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] C. Fuller, D. Biros and R. Wilson "Decision Support for Determining Veracity via Linguistic-based Cues" ELSEVIER , 2009.

[8] S. Peddinti, K. Ross and J. Cappos "Mining Anonymity: Identifying Sensitive Accounts on Twitter," ARXIV ,2016.

[9] R. Oentaryo et. al. "On Profiling Bots in Social Media," ARXIV, 2016.

[10] B. Viswanath et. al. "Towards Detecting Anomalous User Behaviour in Online Social Networks," USENIX, 2014.