Online Social Media Fake Profile Detection Using Data Mining Technique

Suneet Joshi[1], Dr. Deepak Singh Tomar[2]
PhD Research Scholar, CSE Dept., MANIT, Bhopal[1]
Associate Professor, CSE Dept, MANIT, Bhopal[2]
suneetjoshi_2000@yahoo.com[1], deepaktomarmanit@gmail.com[2]

Abstract: *The data mining techniques are helpful to deal with complex real-world problems. It is used for prediction and classification problems much frequently. In this proposed work, the data mining technique is employed for recognizing fake profiles in online social media. Due to a number of social media marketing agencies and spammers, the fake profiles in social media is continuously increasing. It is a serious issue in different social media platform (such as Twitter and Facebook). In this context, the proposed work is focused on designing a data model that is able to evaluate and classify the fake profiles according to their different attributes. This work implements the SVM based classification technique with RBF kernel for obtaining higher accuracy. The experimental outcomes demonstrate the proposed data model is accurate and achieves 94.5% accurate classification outcomes.*

Keywords: *Online social media, fake profile, classification, pattern recognition, comparison.*

1. INTRODUCTION

Social media is one of the online platforms where every age group members are spending their time. Most of them are genuinely usages their profile for communicating with their loved one and sharing their data among OSN (online social media) users [1]. On the other hand, some of the users are creating their fake profiles for different intentions such as advertisements, spreading hate, phishing innocent peoples [2]. Therefore fake profiles in social media are a crucial issue now in these days. In this context recently a number of studies are performed for identification or classification of un-trusted profiles [3] over different social media platforms i.e. Facebook [4] and Twitter [5].

The aim of the proposed work is to design an accurate data mining technique that works on social media fake profile dataset. Therefore a method is introduced in this paper, which first combines the fake profile and normal user profile data. In further, the data is preprocessed for handing the missing attributes and null values [6]. After improving the quality of learning dataset the feature selection approach is implemented. The aim of the feature selection technique is to reduce the dimensions of the initial dataset [7]. After computing the valuable dataset attributes the dataset is partitioned in two parts first is used for training of classification algorithm. That is 70% of entire dataset instances. Additionally, 30% of instances are selected for testing of the trained data model. After splitting the data SVM classifier is trained with the help of radial basis function and training data [8]. After training the trained model is used for classifying the test data. The performance evaluation of the designed classification technique is performed using N-cross validation technique [9].

This section provides a basic overview of the proposed data model for fake profile classification. In further sections, the design of the system and their performance evaluation is described. Finally, the conclusion is reported based on experimental analysis and some future extension of the work is also included.

2. PROPOSED WORK

The main aim of the proposed work is to accurately classify the fake profiles on a social media platform. In this context, a data model is designed for classifying the available attributes in fake profile dataset. In addition to that, some additional effort on the raw dataset is applied to improve the performance of the classifier. This section describes the proposed classification technique.

A. Dataset

The dataset is the key to any data mining and machine learning system. In this work, the supervised learning technique is used for the classification problem. The data set for the proposed system is taken from the Github [10]. This dataset contains two different CSV files where one of the files contains the attributes of fake profile and other CSV file contains the attributes of a legitimate user's profile. The dataset contains 34 attributes, among some of the attributes, are contains the URLs, Strings, and numerical values. In addition to that 1338 instances are available in fake user profile dataset and 1482 instances are available in normal user's dataset. Because the dataset is in raw format, therefore, a significant amount of attributes are containing the missing values. Thus the dataset is noisy and needs additional preprocessing and feature selection approaches. Additionally the dataset consist of two files for user definitions, therefore, the classification problem is concluded as binary.

B. Proposed methodology

The proposed fake profile classification strategy is highlighted in figure 1. The components of the given system are explained as follows:

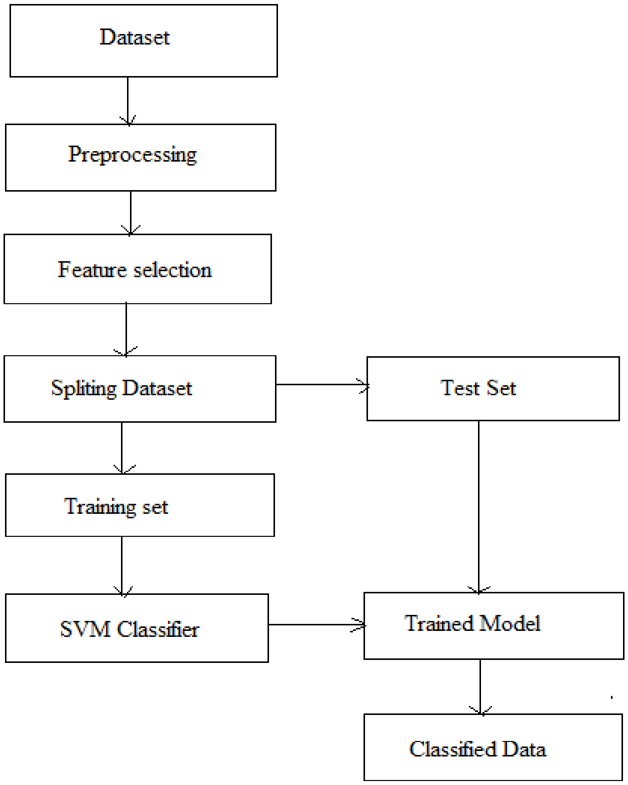

Figure 1: proposed methodology

Dataset: the description of the raw dataset is provided in the previous section. In this phase, the system accepts both the raw dataset files i.e. fake user and normal user. Thus first of all both the set of files read with the help of system and combined to prepare a combined dataset.

Preprocessing: the dataset is contains missing values, which is required to handle. The missing values can impact on the performance of the classifier. On the other hand in fake profiles data contains many missing values as compared to the normal profile. Therefore in place of removal of the missing values, we encode the attributes. In order to encode the dataset the following process is used:

Table 1: preprocessing

```
Input: combined dataset D
Output: encoded dataset $E_d$
Process:
  1. $[Row, col] = ReadDataset(D)$
  2. $for(i = 1; i \leq col; i++)$
       a. $if(D_i == null \,||\, D_i == missing)$
       b. $for(j = 1; j \leq row; j++)$
            i. $if(D_{ij} != null)$
                 1. $D_{ij} = 1$
            ii. Else
                 1. $D_{ij} = 0$
            iii. End if
       c. End for
  3. $E_d = E_d.Add(D_i)$
  4. End for
```

According to the above given algorithm, the process first scans the attributes one by one. If the attribute's values are missing or null in any column then the system replaces all the values by using 0 and 1 encoding. Thus, if the value is available then the values are replaced by 1 and if the value is not available then it is replaced by using 0. By using this scheme all the dataset is encoded or altered. The final attributes and its values are preserved in a new dataset namely E_d.

Feature selection: the size of the dataset is significantly higher, and it takes time for learning and classification. Therefore the dimensionality reduction technique is used here for reducing the dataset size. In this context, the correlation coefficient technique is implemented with the entire encoded dataset. Based on this process the un-useful attributes from the dataset is reduced. The process of feature selection is described in below table 2.

Table 2: feature selection

```
Input: encoded dataset $E_d$, Number of features F
Output: reduced dataset $R_d$
Process:
  1. $for(i = 1; i \leq col; i++)$
       a. $X = E_d^i$
       b. $Y = C$
       c. $CC_i = ComputeCorelation(X, Y)$
       d. $Rank.Add(CC_i)$
  2. $end\ for$
  3. $Rank.sort(\quad)$
  4. $R_d = R_d.Add(Rank, F, E_d)$
  5. $return\ R_d$
```

According to the above-given process, each attribute is evaluated with the respect of class labels. During this, the correlation coefficient is calculated for each column. The calculated value is given here as the rank of the attributes. After computing ranks of all the columns, the rank values are sorted. After sorting the attributes rank the numbers of required features are separated from the dataset which is used for the further classification process.

Splitting dataset: the aim of this process is to prepare the two different sets of data for cross-validation process. Among one part of data is used for training and the second part of data is used for testing of the trained data model.

Training set: the training set is prepared with the 70% of attributes which is used directly with the classifier for training purpose.

Test set: the 30% of randomly selected data instances are used for testing of the trained classifier. As we told we are using the cross-validation process therefore both the set of data is different for training and testing using the random selection of instances.

SVM classifier: according to the nature of data the given problem is a binary classification problem. Therefore the SVM (support vector machine) is an efficient and accurate binary classifier. That is a supervised learning technique and here we use the radial basis function as the kernel function during the SVM training and testing.

Trained model: the training dataset is used for SVM training and after training the model is prepared for classification.

Classified data: the test dataset is classified using the trained SVM classifier. The trained data model provides the outcomes for each fold and the mean value of the validation is used for producing and preparing the confusion matrix.

C. Proposed algorithm

This section provides the combined steps of a process for classifying the fake profile dataset. Table 3 contains the steps of processes involved.

Table 3: proposed algorithm

```
Input: fake profile data F, normal profile Date N, number of Features M
Output: classified data C
Process:
  1. $F^{'} = readData(F)$
  2. $N^{'} = readData(N)$
  3. $D = F^{'} + N^{'}$
  4. $P = preProcessData(D)$
  5. $SF = ComputeFeatures(P, M)$
  6. $[Train, Test] = Split(SF, 70,30, Random)$
  7. $Train_{model} = SVM.Train(Train, RBF)$
  8. $C = Train_{model}.Classify(Test)$
  9. Return C
```

According to the process given in table 3, the proposed system first accept both the parts of data as input. Additionally, both the part of data is combined to form combined dataset D. after preparing the dataset it is preprocessed. During the preprocessing of data, it is encoded to rectify the missing values and null values in the dataset. After processing the dataset the feature selection is performed on modified data. Additionally, the selected features are sub-divided in two parts training set and testing set. The training set is used to train the SVM classifier and after training the test dataset is used. After the classification of test data, the performance of the classification is computed.

3. RESULT ANALYSIS

The experiments are performed with the proposed system using the fake social media profile dataset. In order to measure the performance of the system, N-cross validation technique is used. Additionally based on different numbers of experiments the mean values are referred.

Fake 0.998	0.002
0.097	Genuine 0.903

Figure 2: normalized confusion matrix

The normalized confusion matrix of the obtained classification outcome is demonstrated in figure 2. According to the obtained confusion matrix, the true positive rate of the accurate classification is 0.998 and the true negative rate is 0.903. On the other hand, the false negative rate of classification is 0.002 and a false positive rate of classification is 0.097. Based on the above matrix the additional parameters are also computed.

$$precesion = \frac{TP}{TP + FP}$$

Thus from the above equation, the precision of the classification algorithm is for fake 99 and for the normal user are 90. The mean precision rates for both the class labels have become 94.5. Similarly, the recall is computed using the following formula:

$$recall = \frac{TP}{TP + FN}$$

Thus for the recall of the fake profile are 90 and for the normal user are 99. Thus the mean of classification recall is 94. Finally, the F-score is computed using the following formula:

$$F - Score = 2 * \frac{precision * Recall}{precision + recall}$$

The obtained f-score of the classification system is 94.25. Figure 3 demonstrates the performance of the proposed system using the Bar graph.

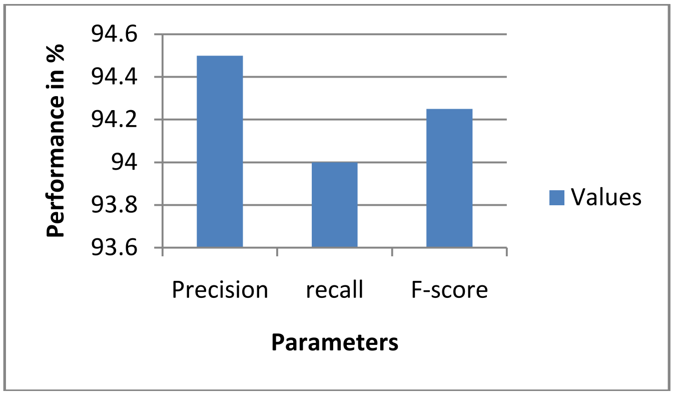

Figure 3: classification performance

The given values here are converted into percentage before reporting.

4. CONCLUSION

The social media is one of the growing platforms; add a significant amount of new users are also appearing in these platforms. Among some of them are genuine users and some of them are fake. These fake users are creating various social and financial issues directly or indirectly. The growing fake users are increasing spam and also can be involved in phishing kinds of criminal acts. Therefore identification of these fake profiles is much essential for clean social media networks. In this presented work the main aim is to design and develop an accurate data mining technique for fake profile identification.

Therefore the proposed work involves a simulation of fake profile classification method. This method usages a third party social media fake profile data set [1] for experimentation and system design. The obtained dataset is noisy therefore some additional effort on data preprocessing and feature extraction is performed. After that radial function based SVM classifier is trained and tested. The obtained performance of the proposed work demonstrates the proposed technique is accurately classifying the data. The obtained mean accuracy of the classification system is 94.5%. Thus the proposed technique is suitable for real-world application of fake profile classification.

In the near future, the following work is proposed to extend the given system.

1. Improvement of classification accuracy.
2. Implementing the technique with ensemble classifier.
3. Performing the comparative performance study with a similar classification technique.

REFERENCES

[1] Steffen Staab, “Trends & Controversies: Social Networks Applied”, IEEE INTELLIGENT SYSTEMS, 1541-1672/05/$20.00 © 2005 IEEE

[2] Tayfun Tuna, Esra Akbas, Ahmet Aksoy, M. Abdullah Canbaz, Umit Karabiyik, Bilal Gonen, Ramazan Aygun, “User Characterization for Online Social Networks”, Copyright 2016, Springer-Verlag Wien, Social Network Analysis and Mining.

[3] Oana Goga, Giridhari Venkatadri, Krishna P. Gummadi, “The Doppelgänger Bot Attack: Exploring Identity Impersonation in Online Social Networks”, IMC’15, October 28–30, 2015, Tokyo, Japan. c 2015 ACM. ISBN 978-1-4503-3848-6/15/10

[4] Mohammadreza Mohammadrezaei, Mohammad Ebrahim Shiri, and Amir Masoud Rahmani, “Identifying Fake Accounts on Social Networks Based on Graph Analysis and Classification Algorithms”, Hindawi, Security and Communication Networks, Volume 2018, Article ID 5923156, 8 pages

[5] Shamstabriz M. Asadullah, Dr. S. V. Viraktamath, “Classification of Twitter Spam Based on Profile and Message Model Using SVM”, International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056, Volume: 04 Issue: 05 | May -2017

[6] Devi Priya R, Sivaraj R, “A REVIEW OF MISSING DATA HANDLING METHODS”, International Journal On Engineering Technology and Sciences – IJETS™ ISSN (P): 2349-3968, ISSN (O): 2349-3976 Volume 2 Issue 2, February 2015

[7] Zena M. Hira and Duncan F. Gillies, “A Review of Feature Selection and Feature Extraction Methods Applied on Microarray Data”, Hindawi Publishing Corporation Advances in Bioinformatics Volume 2015, Article ID 198363, 13 pages

[8] SANA ULLAH JAN, YOUNG-DOO LEE, JUNGPIL SHIN, AND INSOO KOO, “Sensor Fault Classification Based on Support Vector Machine and Statistical Time-Domain Features”, VOLUME 5, 2017, 2169-3536, 2017 IEEE

[9] Jacopo Acquarelli, Twan van Laarhoven, Jan Gerretzen, Thanh N. Tran, Lutgarde M.C. Buydens, Elena Marchiori, “Convolutional neural networks for vibrational spectroscopic data analysis”, Analytica Chimica Acta 954 (2017) 22-31, © 2016 Elsevier B.V.

[10] https://github.com/harshitkgupta/Fake-Profile-Detection-using-ML.