



Article Title: Fake Profiles Identification in Online Social Networks Using Random Forest Algorithm

Fake Profiles Identification in Online Social Networks Using Random Forest Algorithm

Glancy Sherina M ¹, Dr. Maria Jesi², M. D. Amala Dhaya, M.E, (Ph.D)³

¹ Student, ²Associate Professor, ³Assistant Professor

Computer Science and Engineering, Loyola Institute of Technology and Science,

¹glancysherina908@gmail.com, ²mariajesi.CSE@lite.edu.in, ³amaladhayamd@gmail.com

ABSTRACT

Every day several people are creating their profiles on the social network platforms. The social network sites not only providing advantages to the users and also provide security issues to the users as well their information. To analyze, who are encouraging threats in social network need to classify the social networks profiles of the users. From the classification, can get the genuine profiles and fake profiles on the social networks. Different classification methods for detecting the fake profiles on the social networks. In this paper proposing Machine learning and Natural language Processing (NLP) techniques to improve the accuracy rate of the fake profiles detection using the Random forest algorithm.

Keywords: Natural language Processing, Classification, Random Forest algorithm, Networks profiles

1 Introduction

Online identification theft has been a primary problem considering it affected millions of people's worldwide. Victims of identification theft may suffer unique types of penalties; for illustration, they would lose time/cash, get dispatched to reformatory, get their public image ruined, or have their relationships with associates and loved ones damaged. Spatio-temporal mining on social network to determine circle of customers concerned in malicious events with the support of latent semantic analysis [1].The main objective is to detect the fake profiles from the social network sites like Facebook Sites. The presented process used Facebook profile to notice false profiles.

2 Related Works

LinkedIn is greatly preferred through the folks who're in the authentic occupations. With the speedy development of social networks, persons are likely to misuse them for unethical and illegal conducts. Creation of a false profile turns into such adversary outcomes which is intricate to identify without apt research. The current solutions which were virtually developed and theorized to resolve this contention, mainly viewed the traits and the social network ties of the person's social profile. However, in relation to LinkedIn such behavioral observations are tremendously restrictive in publicly to be had profile data for the customers



Article Title: Fake Profiles Identification in Online Social Networks Using Random Forest Algorithm

by the privateness insurance policies. The limited publicly available profile data of LinkedIn makes it ineligible in making use of the existing tactics in fake profile identification.

3 Research Methodology:

3.1 Collection of Datasets

Various attributes to include in the dataset are number of friends, followers, status count. Dataset is resulting to training and testing data. Classification algorithms are trained using training dataset and testing dataset is used to determine efficiency of algorithm.

From the dataset used, More than 80 percent of accounts are used to train the data, 20 percent of accounts to test the data.

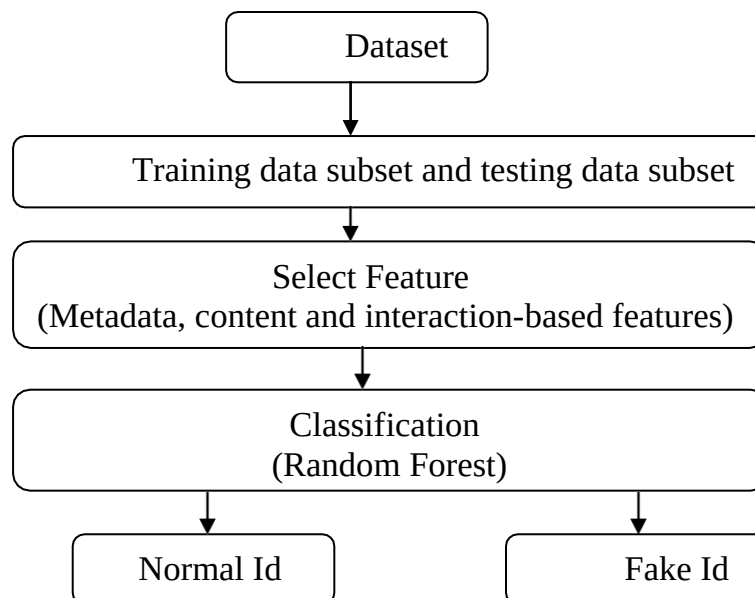


Figure 1: *flow chat*

3.2 Pre-processing

For the successful application pre-processing is required. The data which is acquired from different resources are sometime in raw form. It may contain some incomplete, redundant, inconsistent data. Therefore in this step such redundant data should be filtered.

Data should be normalized

3.3 Feature Extraction

This step is focus on identifying and using most relevant attribute from the dataset. Through this process irrelevant and redundant information is removed for the application of classifiers.

3.4 Applied Random Forest Algorithm

Random Forest is the machine learning algorithmic program that uses a textile approach to make a bunch of call trees with random set. A model is trained a lot random sample of the dataset to realize sane prediction performance many times. The output of all the choice trees within the tree, combined to create the ultimate



Article Title: Fake Profiles Identification in Online Social Networks Using Random Forest Algorithm

prediction. The performances of the classification algorithms Random Forest, Neural Network are used for the detection of fake accounts.

4 Existing system

In Existing system I have used two algorithms like Support Vector Machine (SVM) and Naïve Bayes algorithm. It only predicted the values. Does not suitable for large amount of data's. In existing system system used Machine learning and Natural language Processing (NLP) techniques. For classification to use the Support Vector Machine (SVM) and Naïve Bayes algorithm. Not suitable for large data sets. Does not perform very well when the data set has more noise i.e. target classes are overlapping.

5 Proposed system

In proposed system using Random Forest (RF) algorithm. In this algorithm provide improved the detection accuracy rate compare then previous technique. Will be develop web application by using dijango.

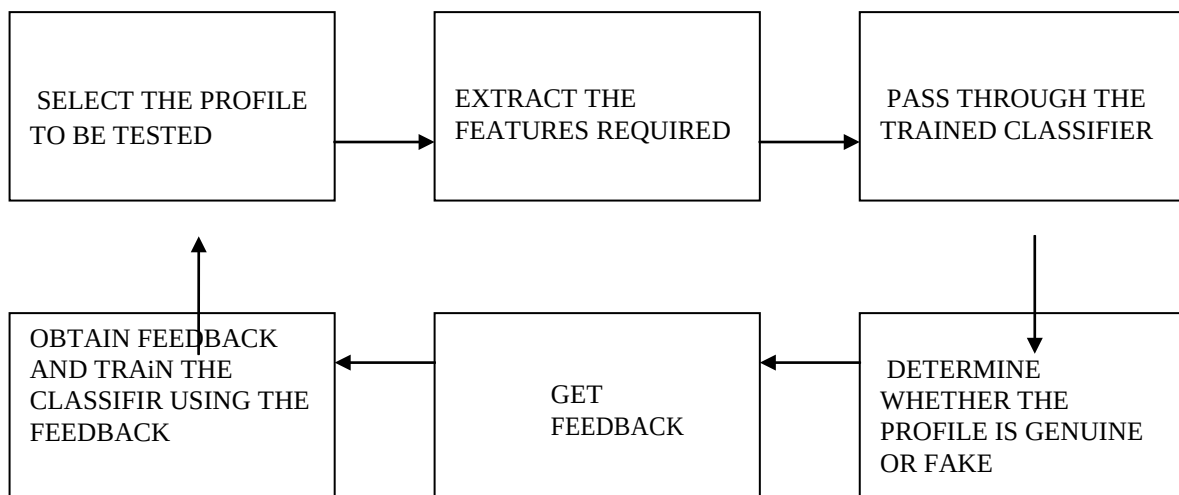


Figure 2: Architecture diagram

The detection process starts with the selection of the profile that needs to be tested. After selection of the profile, the suitable attributes (i.e. features) are selected on which the classification algorithm is implemented. The attributes extracted is passed to the trained classifier. The classifier gets trained regularly as new training data is feed into the classifier. The classifier determines the whether the profile is fake or genuine.

The classifier may not be 100% accurate in classifying the profile so; the feedback of the result is given back to the classifier. This process repeats and as the time proceeds, the no. of training data increases and the classifier becomes more and more accurate in predicting the fake profiles.



Article Title: Fake Profiles Identification in Online Social Networks Using Random Forest Algorithm

5.1 Random Forest algorithm

Random forest is a Supervised Machine Learning Algorithm that is used widely in Classification and Regression problems. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression. Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model. Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset.

6 Results and Discussion

This figure shows the profile type prediction status.

Figure 3: Profile type prediction status

7 Conclusion

By using proposed techniques, easily detect the fake profiles from the social network sites. In this project, the Facebook dataset was used to identify the fake profiles. Improved the detection accuracy rate.

8 Future Scope

Main problem is that a person can have multiple Facebook accounts which makes them an advantage of creating fake profiles and accounts in online social networks. The idea is of attaching Aadhar card number when signing up an account so that we can restrict to create a single account and there is no chance of fake profiles at any moment.

References

1. Z. Halim; M. Gul; N. ul Hassan; R. Baig; S. Rehman; and F. Naz, "Malicious users' circle detection in social network based on spatiotemporal co-occurrence," in *Computer Networks and Information Technology (ICCNIT)*, 2011 International Conference on, July, pp. 35–390



Article Title: Fake Profiles Identification in Online Social Networks Using Random Forest Algorithm

2. Dr. S. Kannan; Vairaprakash Gurusamy, "Preprocessing Techniques for Text Mining", 05 March 2015.
3. Liu Y; Gummadi K; Krishnamurthy B; Mislove A," Analyzing Facebook privacy settings: User expectations vs. reality", in: Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference, ACM, pp. 61–70.
4. Mahmood S; Desmedt Y," Poster: preliminary analysis of google?'s privacy. In: Proceedings of the 18th ACM conference on computer and communications security", ACM 2011, pp. 809–812.
- Stein T; Chen E; Mangla K," Facebook immune system. In: Proceedings of the 4th workshop on social network systems", ACM 2011 pp
5. Saeed Abu-Nimeh; T. M. Chen, and O. Alzubi, "Malicious and Spam Posts in Online Social Networks," Computer, Vol: 44, no: 9, IEEE2011, pp. 23–28.