FAKE PROFILE IDENTIFICATION IN SOCIAL NETWORK USING MACHINE LEARNING AND NLP

Mr. Ramachandramurthy[1], n chandrakanth[2], rahul[3],d rishi kumar[4],p sathwik[5],ch likhitha[6],amreen[7]

[1]Assistant Professor,department of information technology,malla reddy institute of engineering and technology(autonomous),dhulapally,secundrabad

[2,3,4,5,6,7]UG Students,department of information technology,malla reddy institute of engineering and technology(autonomous),dhulapally,secundrabad

ABSTRACT

At present social network sites are part of the life for most of the people. Every day several people are creating their profiles on the social network platforms and they are interacting with others independent of the user's location and time. 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 we need to classify the social networks profiles of the users. From the classification, we can get the genuine profiles and fake profiles on the social networks. Traditionally, we have different classification methods for detecting the fake profiles on the social networks. But, we need to improve the accuracy rate of the fakeprofile detection in the social networks. In this paper we are proposing Machine learning and Natural language Processing (NLP) techniques to improve the accuracy rate of the fake profiles detection. We can use the Support Vector Machine (SVM) and Naïve Bayes algorithm.

Keywords: Machine learning, Natural Language Processing, fake reviews, online product, E-commerce, Product review monitoring.

I. INTRODUCTION :

Social networking has end up a well-known recreation within the web at present, attracting hundreds of thousands of users, spending billions of minutes on such services. Online Social network (OSN) services variety from social interactions-based platforms similar to Face book or MySpace, to understanding dissemination-centric platforms reminiscent of twitter or Google Buzz, to Social interaction characteristic brought to present systems such as Flicker. The opposite hand, enhancing security concerns and

protecting the OSN privateness still signify a most important bottleneck and viewed mission.

When making use of Social network's (SN's), one of a kind men and women share one-of-a-kind quantities of their private understanding. Having our individual know-how entirely or in part uncovered to the general public, makes us excellent targets for unique types of assaults, the worst of which could be identification theft. Identity theft happens when any individual uses character's expertise for a private attain or purpose. During the earlier years, 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. At present, the vast majority of SN's does no longer verifies ordinary users" debts and has very susceptible privateness and safety policies. In fact, most SN's applications default their settings to minimal privateness; and consequently, SN's became a best platform for fraud and abuse. Social Networking offerings have facilitated identity theft and

Impersonation attacks for serious as good as naïve attackers. To make things worse, users are required to furnish correct understanding to set up an account in Social Networking web sites. Easy monitoring of what customers share on-line would lead to catastrophic losses, let alone, if such bills had been hacked.

Profile information in online networks will also be static or dynamic. The details which can be supplied with the aid of the person on the time of profile creation is known as static knowledge, the place as the small print that are recounted with the aid of the system within the network is called dynamic knowledge. Static knowledge includes demographic elements of a person and his/her interests and dynamic knowledge includes person runtime habits and locality in the network. The vast majority of current research depends on static and

dynamic data. However this isn't relevant to lots of the social networks, where handiest some of static profiles are seen and dynamic profiles usually are not obvious to the person network. More than a few procedures have been proposed by one of a kind researcher to realize the fake identities and malicious content material in online social

networks. Each process had its own deserves and demerits.

The problems involving social networking like privacy, online bullying, misuse, and trolling and many others. Are many of the instances utilized by false profiles on social networking sites. False profiles are the profiles which are not specific i.e. They're the profiles of men and women with false credentials. The false Face book profiles more commonly are indulged in malicious and undesirable activities, causing problems to the social community customers. Individuals create fake profiles for social engineering, online impersonation to defame a man or woman, promoting and campaigning for a character or a crowd of individuals. Face book has its own security system to guard person credentials from spamming, phishing, and so on. And the equal is often called Facebook Immune system (FIS).

The FIS has now not been ready to observe fake profiles created on Facebook via customers to a bigger extent.

II. LITERATURE REVIEW

Fake Profile Identification in Social Network using Machine Learning and NLP, In present times, social media plays a key role in every individual life. Everyday majority of the people are spending their time on social media platforms. The number of accounts in these social networking sites has dramatically increasingday-by-day and many of the users are interacting with others irrespective of their timeand location. These social media sites have both pros and cons and providesecurity problems to us also for our information. To scrutinize, who are giving threats in these networking sites we need to organize these social networking accounts into genuine accounts and fake accounts. Traditionally, we are having different classification methods to point out the fake accounts on social media. But we must increase the accuracy rate in identifying fake accounts on these sites. In our paper we are going with Machine Learning technologies and Natural Language processing (NLP) to increase the accuracy rate of detecting the fake accounts. We opted for Random Forest tree classifier algorithm.
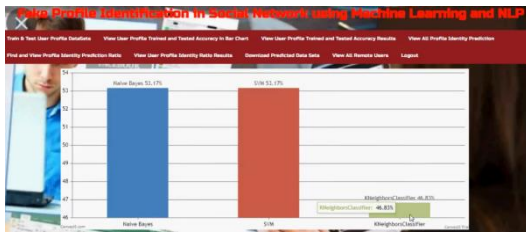
III. MODULES

Service Provider

In this module, the Service Provider has to login by using valid user name and password.

After login successful he can do some operations such as Login, Train & Test User Profile Data Sets, View User Profile Trained and Tested Accuracy in Bar Chart,
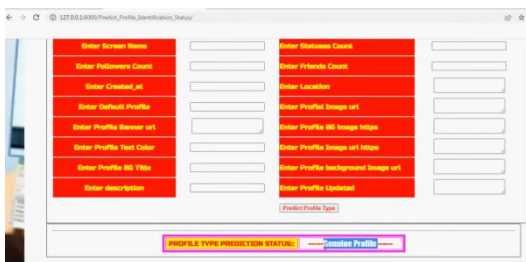

View User Profile Trained and Tested Accuracy Results,

View All Profile Identity Prediction,

Find and View Profile Identity Prediction Ratio, View User Profile

Identity Ratio Results, Download Predicted Data Sets, View All Remote Users

View and Authorize Users

In this module, the admin can view the list of users who all registered. In this, the admin can view the user's details such as, user name, email, address and admin authorizes the users.

Remote User

In this module, there are n numbers of users are present. User should register before doing any operations.

Once user registers, their details will be stored to the database. After registration successful, he has to login by using authorized user name and password. Once login is successful user will do some operations like register and login,predict profile identification status,view your profile.

IV. EXISTING SYSTEM

Chai et al awarded on this paper is a proof-of inspiration gain knowledge of. Even though the prototype approach has employed most effective normal systems in normal language processing and

human-pc interplay, the results realized from the user trying out are significant. By using comparing this simple prototype approach with a wholly deployed menu procedure, they've discovered that users, principally beginner users, strongly pick the common language dialog-based approach. They have additionally learned that in an ecommerce environment sophistication in dialog administration is most important than the potential to manage complex typical language sentences.

In addition, to provide effortless access to knowledge on ecommerce web sites, natural language dialog-based navigation and menu-pushed navigation should be intelligently combined to meet person's one-of-a-kind wants. Not too long ago, they have got accomplished development of a new iteration of the approach that includes enormous enhancements in language processing, dialog administration and information management. They believed that average language informal interfaces present powerful personalized alternatives to conventional menupushed or search-based interfaces to web sites.

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 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. For that reason, there is to conduct distinctive study on deciding on systems for fake profile identification in LinkedIn. Shalinda Adikari and Kaushik Dutta researched and identified the minimal set of profile data that are crucial for picking out false profiles in LinkedIn and labeled the appropriate knowledge mining procedure for such project.

Z. Halim et al. Proposed spatio-temporal mining on social network to determine circle of customers concerned in malicious events with the support of latent semantic analysis. Then compare

the results comprised of spatio temporal co incidence with that of original organization/ties with in social network, which could be very encouraging as the organization generated by spatio-temporal co-prevalence and actual one are very nearly each other. Once they set the worth of threshold to right level, we develop the number of nodes i.e. Actor so that they are able to get higher photo. Total, scan indicate that Latent Semantic Indexing participate in very good for picking out malicious contents, if the feature set is competently chosen. One obvious quandary of this technique is how users pick their function set and the way rich it's. If the characteristic set is very small then most of the malicious content material will not be traced. However, the bigger person function set, better the performance won.

Disadvantages

➢ The system is not implemented Learning Algorithms like svm, Naive Bayes.

➢ The system is not implemented any the problems involving social networking like privacy, online bullying, misuse, and trolling and many others.

V.PROPOSED SYSTEM

On this paper we presented a machine learning & natural language processing system to observe the false profiles in online social networks. Moreover, we are adding the SVM classifier and naïve bayes algorithm to increase the detection accuracy rate of the fake profiles.

An SVM classifies information by means of finding the exceptional hyperplane that separates all information facets of 1 type from those of the other classification. The best hyperplane for an SVM method that the one with the biggest line between the two classes. An SVM classifies data through discovering the exceptional hyperplane that separates all knowledge facets of one category from those of the other class. The help vectors are the info aspects which are closest to the keeping apart hyperplane.

Naive Bayes algorithm is the algorithm that learns the chance of an object with designated features belonging to a unique crew/category. In brief, it's a probabilistic classifier. The Naive Bayes algorithm is called "naive" on account that it makes the belief that the occurrence of a distinct feature is independent of the prevalence of other aspects. For illustration, if we're looking to determine false profiles based on its time, date of publication or posts, language and geoposition. Even if these points depend upon each and every different or on the presence of the other

facets, all of these properties in my view contribute to the probability that the false profile.

Advantages

➢ In the proposed system, Profile information in online networks will also be static or dynamic. The details which can be supplied with the aid of the person on the time of profile creation is known as static knowledge, the place as the small print that are recounted with the aid of the system within the network is called dynamic knowledge.

➢ In the proposed system, Social Networking offerings have facilitated identity theft and Impersonation attacks for serious as good as naïve attackers.

Overview of Proposed System :

This paper introduces a machine learning and natural language processing system designed to identify fraudulent profiles within online social networks. The system incorporates the SVM classifier and Naïve Bayes algorithm to enhance the accuracy of detecting fake profiles. By integrating these algorithms, we aim to significantly improve the system's capability to discern and flag deceptive or false identities within online social networks.

The process outlined in this study utilizes Facebook profiles to detect fraudulent accounts. The proposed methodology operates through three primary phases, each contributing to the overall detection process.

1. SVM
2. Naïve Bayes

SVM (Support Vector Machine)

In classification tasks a discriminant machine learning technique aims at finding, based on an *independent and identically distributed* (*iid*) training dataset, a discriminant function that can correctly predict labels for newly acquired instances. Unlike generative machine learning approaches, which require computations of conditional probability distributions, a discriminant classification function takes a data point x and assigns it to one of the different classes that are a part of the classification task. Less powerful than generative approaches, which are mostly used when prediction involves outlier detection, discriminant approaches require fewer computational resources and less training data, especially for a multidimensional feature space and when only posterior probabilities are needed. From a geometric perspective, learning a classifier is equivalent to finding the equation for a

multidimensional surface that best separates the different classes in the feature space.

SVM is a discriminant technique, and, because it solves the convex optimization problem analytically, it always returns the same optimal hyperplane parameter—in contrast to *genetic algorithms* (*GAs*) or *perceptrons*, both of which are widely used for classification in machine learning. For perceptrons, solutions are highly dependent on the initialization and termination criteria. For a specific kernel that transforms the data from the input space to the feature space, training returns uniquely defined SVM model parameters for a given training set, whereas the perceptron and GA classifier models are different each time training is initialized. The aim of GAs and perceptrons is only to minimize error during training, which will translate into several hyperplanes' meeting this requirement.

Naïve Bayes :

The naive bayes approach is a supervised learning method which is based on a simplistic hypothesis: it assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of

any other feature .Yet, despite this, it appears robust and efficient. Its performance is comparable to other supervised learning techniques. Various reasons have been advanced in the literature. In this tutorial, we highlight an explanation based on the representation bias. The naive bayes classifier is a linear classifier, as well as linear discriminant analysis, logistic regression or linear SVM (support vector machine). The difference lies on the method of estimating the parameters of the classifier (the learning bias).

While the Naive Bayes classifier is widely used in the research world, it is not widespread among practitioners which want to obtain usable results. On the one hand, the researchers found especially it is very easy to program and implement it, its parameters are easy to estimate, learning is very fast even on very large databases, its accuracy is reasonably good in comparison to the other approaches. On the other hand, the final users do not obtain a model easy to interpret and deploy, they does not understand the interest of such a technique.

Thus, we introduce in a new presentation of the results of the learning process. The classifier is easier to

understand, and its deployment is also made easier. In the first part of this tutorial, we present some theoretical aspects of the naive bayes classifier. Then, we implement the approach on a dataset with Tanagra. We compare the obtained results (the parameters of the model) to those obtained with other linear approaches such as the logistic regression, the linear discriminant analysis and the linear SVM. We note that the results are highly consistent. This largely explains the good performance of the method in comparison to others. In the second part, we use various tools on the same dataset (**Weka 3.6.0**, **R 2.9.2**, **Knime 2.1.1**, **Orange 2.0b** and **RapidMiner 4.6.0)**. We try above all to understand the obtained results.

VI.CONCLUSION

In this paper, we proposed machine learning algorithms along with natural language processing techniques. By using these techniques, we can easily detect the fake profiles from the social network sites. In this paper we took the Face book

Data set to identify the fake profiles. The NLP pre-processing techniques are used to analyze the dataset and machine learning algorithm such as SVM and Naïve Bayes are used to classify the profiles. These learning algorithms are improved the detection accuracy rate in this paper.

VII.REFERENCES

1. Michael Fire et al. (2012). "Strangers intrusion detection-detecting spammers and fake profiles in social networks based on topology anomalies." Human Journal 1(1): 26-39.Günther, F. and S. Fritsch (2010). "neuralnet: Training of neural networks." The R Journal 2(1): 30-38

2. Dr. S. Kannan, Vairaprakash Gurusamy, "Preprocessing Techniques for Text Mining", 05 March 2015.

3. Shalinda Adikari and Kaushik Dutta, Identifying Fake Profiles in LinkedIn, PACIS 2014 Proceedings, AISeL

4. 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.

5. 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.

6. 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.

7. Stein T, Chen E, Mangla K," Facebook immune system. In: Proceedings of the 4th workshop on social network systems", ACM 2011, pp

8. 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.

9. J. Jiang, C. Wilson, X. Wang, P. Huang, W. Sha, Y. Dai, B. Zhao, Understanding latent interactions in online social networks, in: Proceedings of the 10th ACM SIGCOMM Conference on Internet Measurement, ACM, 2010, pp. 369–382

10. Kazienko, P. and K. Musiał (2006). Social capital in online social networks. Knowledge-Based Intelligent Information and Engineering Systems, Springer