iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** VI **Month of publication:** June 2022

DOI: https://doi.org/10.22214/ijraset.2022.44548

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue VI June 2022- Available at www.ijraset.com

Identification of Forged Profiles in Online Social Networks Using Machine Learning and NLP

Praneeth Narisetty[1], Sai Sravan Rakesh Saripalli[2]
Lovely Professional University

Abstract: ***Many people today use social networking sites as a part of their everyday lives. They create their own profiles on the social network platforms every day, and they interact with others regardless of their location and time. In addition to providing users with advantages, social networking sites also present security concerns to them and their information to them. We need to classify the social network profiles of the users to figure out who is encouraging threats on social networks. From the classification, we can figure out which profiles are genuine and which are fake. As far as detecting fake profiles on social networks is concerned, we currently have different classification methods. However, we must improve the accuracy of detecting fake profiles in social networks. We propose the use of a machine learning algorithm and Natural Language Processing (NLP) technique in this paper so as to increase the detection rate of fake profiles. This can be achieved using Support Vector Machines (SVM) and Naïve Bayes algorithms.***

Keywords: Social Networks, Machine Learning, NLP, Naïve Bayes, Support Vector Machine

I. INTRODUCTION

The Internet has developed into a well-known place for recreation, attracting hundreds of thousands of users who spend billions of minutes using it. Among the services offered by online social networks (OSNs) are social interaction-based platforms such as Facebook and MySpace, understanding dissemination-centric platforms such as Twitter and Google Buzz, and social interaction features similar to those seen on present-day social networking sites like Flickr. On the other hand, strengthening security concerns and protecting OSN privacy continue to be important bottlenecks and viewed as a mission. People who use Social Networks (SNs) share unique amounts of information about themselves. Our individual knowledge is exposed to the public in part or entirely, which makes us a prime target for various types of assaults, including identification theft. An identity theft occurs when someone uses a character's knowledge for their own gain. There was a primary problem with online identification theft during the earlier years, as it affected millions of people worldwide. In addition to losing time and money, identification theft victims may be sent to reformatory, have their public image destroyed, or have their relationships with family and associates damaged. There are currently no verification processes for ordinary users' debts on most social media platforms and their privacy and safety policies are extremely vulnerable. In fact, most social media applications default their settings to minimal privacy; hence, SN's became the best platform for fraud and abuse. It is possible for serious as well as naive attackers to commit identity theft and impersonation attacks through Social Networking services. In addition, social networking web sites require users to provide correct understanding to set up an account. If such bills were hacked, customers' online sharing would result in catastrophic losses. The information on online networks can also be static or dynamic. Dynamic knowledge is the detail that is relayed to the network with the assistance of the system; static knowledge is the detail that can be supplied at the time of profile creation, whereas static knowledge is the information that can be provided by the individual during the creation of the profile. The vast majority of current research relies on both static and dynamic data, whereas dynamic knowledge includes demographic information and interests. People's runtime habits and location in networks are considered dynamic knowledge, whereas static knowledge includes demographic information and interests. However, this doesn't apply to a great number of social networks, where only some static profiles are visible, and dynamic profiles are not usually apparent to individuals. The researcher presented several methods for detecting fake identities and malicious content material on social networks. These methods each had their own strengths and weaknesses. There are many problems associated with social networking, including privacy concerns, bullying, misuse, trolling, and many others. There are many instances of false profiles being used on social networking sites. False profiles are those that aren't specific, such as those containing false credentials, or those that aren't gender-based. Many false Facebook accounts engage in malicious and undesirable activities, causing problems for customers in the social community. People create fake profiles to defame a person or a group of people online, promote and campaign for characters or groups of people, or to do social engineering. In order to protect users from spam, phishing, and other threats, Facebook has its own security system, called the Facebook Immune System (FIS). To a large extent, the FIS has been unable to discover fake Facebook profiles produced by customers.

II. RELATED WORK

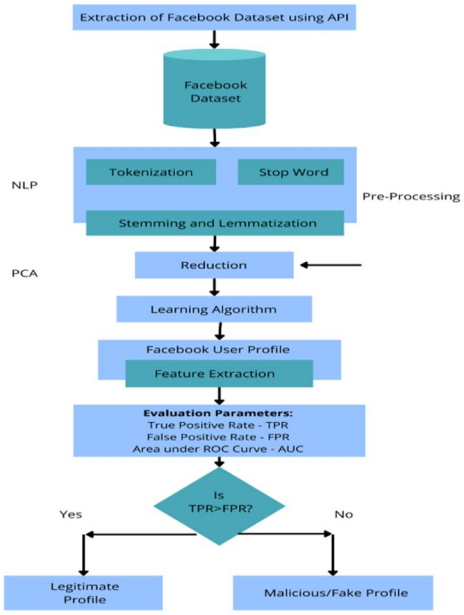
Although the prototype approach employs the best normal systems for normal language processing and human-computer interaction, Chai et al awarded this paper as a proof-of inspiration gain knowledge, the results realized by the user are significant. The researchers have discovered that the common language dialog-based approach is preferred by users, mainly beginner users, when compared to a fully deployed menu procedure. Also, they learned that sophistication in dialog management is more important than the ability to handle complex typical language sentences in an ecommerce environment. It is important to combine natural language dialog-based navigation with menu-driven navigation intelligently on ecommerce web sites to meet the unique needs of each user. Their approach has recently been revised with significant improvements in language processing, dialog management, and information management. Recently, they completed the development of a new iteration of the approach. According to them, informal interfaces that are based on average language present powerful alternatives to conventional menu-driven and search-based web sites. Many individuals in the authentic occupations prefer LinkedIn. Social networks have developed rapidly, and people are likely to misuse them for unethical and illegal conduct. Creating a false profile can lead to adversary outcomes difficult to identify without adequate research. Several solutions have been developed and theorized to resolve this contention based largely on the traits and social network ties of the individual's social profile. The LinkedIn privateness insurance policies, however, severely restrict the public availability of customer profile data regarding behavioral observations. LinkedIn is unable to identify fake profiles using existing tactics due to the limited public profile data available. It is therefore necessary to conduct a distinguishing study on how LinkedIn's fake profile identification system should be evaluated. Shalinda Adikari and Kaushik Dutta identified the crucial profile data and the knowledge mining procedure that are necessary for identifying false profiles on LinkedIn.

A spatial-temporal analysis of social networks was proposed to identify circles of customers who are involved in malicious events based on latent semantic analysis by Z. Halim et al. As the organization generated by spatial-temporal co-incidence is very similar to the organization generated by the actual organization, it may be very encouraging to compare the results resulting from spatial-temporal co-incidence with the original organization/ties within the social network. In order to get higher photos, we develop the number of nodes i.e. Actors once they have set the threshold level. If the feature set is correctly selected, Latent Semantic Indexing performs very well for identifying malicious content. There is a problem with this technique, in that users need to decide on how rich their feature set is, because if it is too small, most malicious content will not be detected. On the other hand, the larger the feature set, the better the performance.

III. BACKGROUND

A. Overview of the Proposed System

The purpose of this paper is to present a machine learning & natural language processing system to analyze false profiles in online social networks. We also introduce SVM and naive Bayes algorithms to improve the accuracy rate of detecting the fake profiles.



Working Procedure for Proposed System

A Facebook profile is used to detect false profiles through the proposed procedure. The method includes three phases;

1) NLP Pre-processing
2) Principal Component Analysis(PCA)
3) Learning Algorithms

A. NLP Pre-Processing

The importance of text pre-processing is essential to any NLP approach, and the following benefits can be gained from NLP pre-processing:

- Text content records should be indexed (or knowledge records) as little as possible
  - Words that stop appearing in textual content make up 20-30% of the total phrase count
  - Stemmed indexing could reduce it by 40-50%
- For the IR method to be more efficient and effective
  - The search engine will just be confused when stops words are used in textual content mining or shopping
  - The stuttering method is used to find similar words within a document.

1) Tokenization

The tokenization process consists of dividing up a text into phrases, phrases, symbols, or other significant factors known as tokens. Tokenization is aimed at exploring the phrases within this text. Tokenization is a vital component of both linguistics (as a way to segment textual content) and laptop science (as a method of lexical analysis). The list of tokens becomes input for further processing such as parsing or textual content mining. Textual knowledge is essentially a block of characters. For this reason, parsers must tokenize records as part of the requirement of knowledge retrieval strategies. Even though the text is already stored in codecs readable by computers, some challenges remain. For example, punctuation marks must be removed. Other characters, like brackets, hyphens, and so on, need to be handled as well.

2) Stop word Removal

Often, stop phrases are outdated phrases such as 'and', 'are', 'this' etc. They do not appear to be helpful for classification of records. Therefore, they must be removed. However, the development of such stop phrases records is problematic and inconsistent between textual sources. Moreover, each textual content report contains phrases that are not essential for text mining applications, resulting in a reduction in text knowledge and improved approach performance.

3) Stemming and Lemmatization

By stemming and lemmatizing a phrase, inflectional types & mainly derivationally associated types are consolidated into one fashioned base type. In most cases, stemming involves trimming the ends of words to achieve this, goal, and it is quite common to lose derivational affixes as part of this process. As a result of lemmatization, in most cases the goal is to reduce inflectional endings that come after the base or dictionary type of a word, to return to the base or dictionary type of the original word.

B. Principal Component Analysis (PCA)

By performing Principal Component Analysis, the basic understanding made possible from the table can be synthesized as a suite of orthogonal variables known as major accessories, and the similarity between observations, variables, and elements of maps revealed.

C. Learning Algorithms

The proposed system is based on the Support Vector Machine (SVM) and naïve Bayes algorithms, both of which are machine learning algorithms.

1) Support Vector Machine (SVM)

As part of the SVM classifying information, all facets of one type are separated by an exceptional hyperplane. A hyperplane with the largest line between two classes is the best hyperplane for an SVM method. A SVM classifies data by identifying the exceptional hyperplane that separates the knowledge facets of one category from the other. Keeping apart vectors are the information aspects that are closest to the support vectors.

2) Naïve Bayes

An object is placed into a certain crew/category based on whether it is most likely to belong to a certain class using a probabilistic classifier known as Naive Bayes.

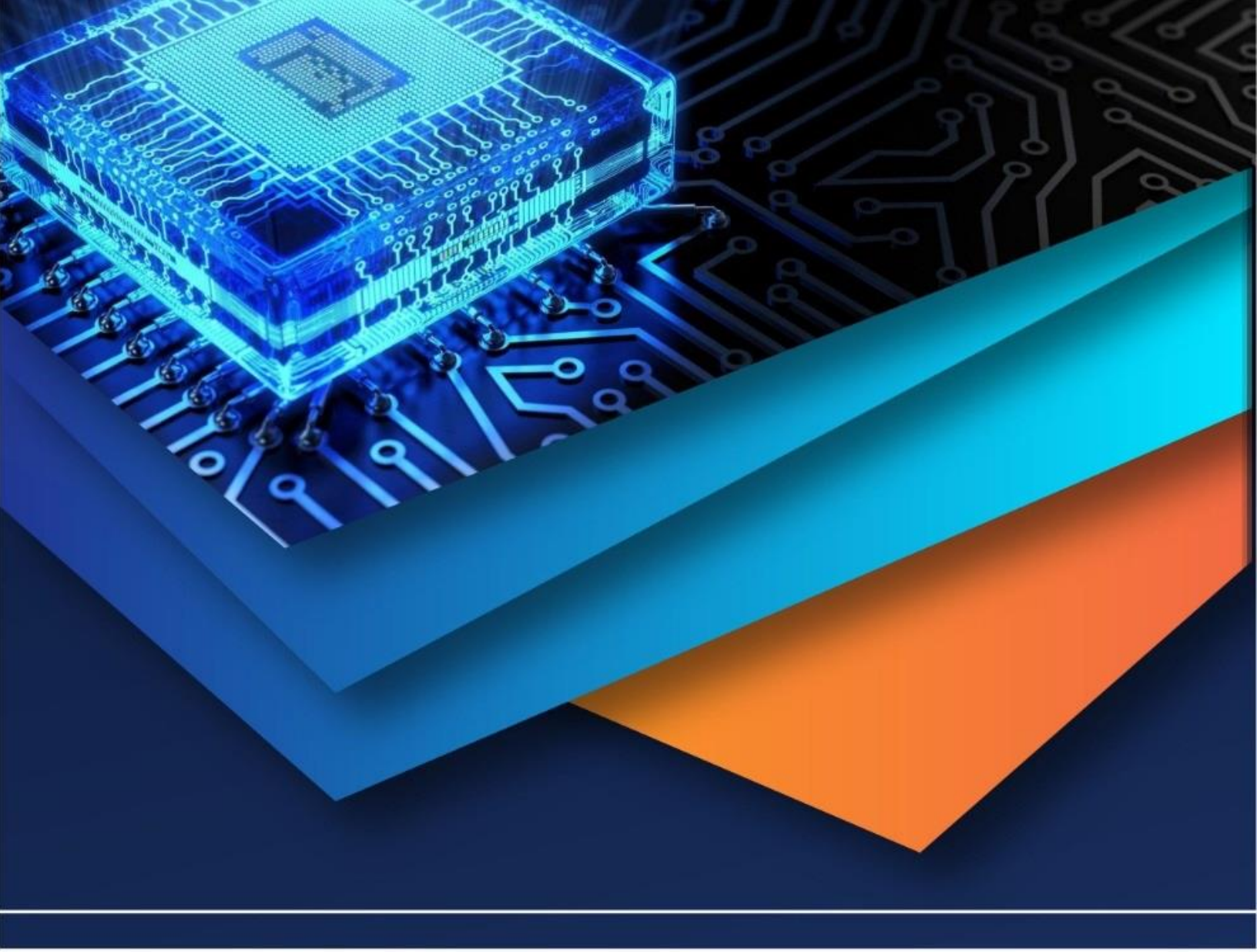
In its name, the Naive Bayes algorithm indicates that distinct features are independent of other features. We can identify false profiles by analyzing the time, date, language, and location of posts, for example. Even if they depend upon each other or on the presence of the other facets, all of these properties, in my opinion, contribute to the probability that the false profile exists.

IV. CONCLUSION

As part of this paper, we propose machine learning algorithms and natural language processing techniques for detecting fake social network profiles. In this paper, we take the Facebook dataset as a basis for identifying the fake profiles. To analyze the dataset and classify profiles, NLP pre-processing techniques are used, followed by machine learning algorithms like SVMs and Naïve Bayes. This paper exhibits an improved rate of detection accuracy after applying the learning algorithms.

REFERANCES

[1] Dr. S. Kannan, Vairaprakash Gurusamy,"Preprocessing Techniques for Text Mining", 05March 2015.
[2] Arushi Gupta, RishabhKaushal, "Improving Spam Detection in Online Social Networks",978-1-4799 7171-8/15/$31.00 ©2015 IEEE.
[3] Cody Buntain, Jennifer Golbeck, "Automatically Identifying Fake News in PopularTwitter Threads", 2017 IEEE International Conference on Smart Cloud
[4] ShlokaGilda,"Evaluating Machine Learning Algorithms for Fake News Detection",2017 IEEE 15th Student Conference on Research and Development (SCORED).
[5] S. Maheshwari, "How fake news goes viral: A case study", Nov.2016.
[6] Shalinda Adikari and Kaushik Dutta, Identifying Fake Profiles in LinkedIn, PACIS 2014 Proceedings, AISEL
[7] 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.
[8] 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.
[9] 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.
[10] Marco L. Della Vedova, Eugenio Tacchini, Stefano Moret, Gabriele Ballarin, Massimo DiPierro, Luca de Alfaro, "Automatic Online Fake News Detection Combining Content and Social Signals", ISSN 2305-7254,2017.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY