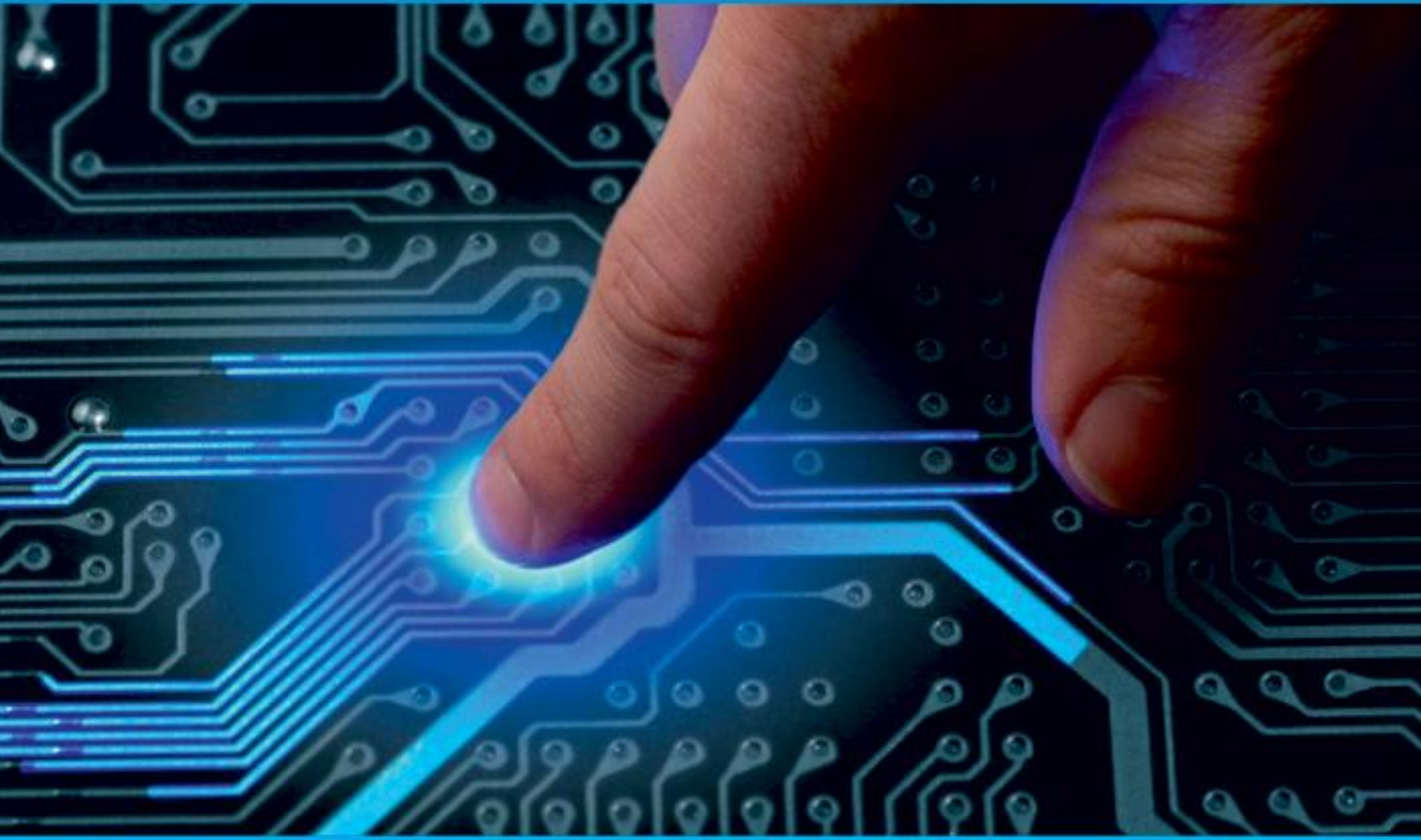
IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 12, Issue 2, April 2024

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.379

A Blockchain-Based Solution to identify and Block Fake Profiles

Sathya Priya R, Kowshikraj K, Mohana N, Naveen B, Preethi M

Assistant Professor, Department of Computer Science and Engineering, Knowledge Institute of Technology, Salem, India

Department of Computer Science and Engineering, Knowledge Institute of Technology, Salem, India

ABSTRACT: The networking sites have become an integral part of most people's lives. Every day, numerous individuals create their profiles on various product networking platforms and interact with others from anywhere, at any time. However, these platforms not only offer benefits to users but also pose security risks to their personal information. To identify potential threats on product networks, it is crucial to classify users' profiles as genuine or fake. Although traditional methods exist for detecting fraudulent profiles, it is necessary to improve the accuracy rate of fake profile detection on product networks. This paper proposes the use of Machine Learning and Natural Language Processing (NLP) techniques, such as the Support Vector Machine (SVM) and Naïve Bayes algorithm, to enhance the accuracy rate of fake profile detection.

KEYWORDS: Network profiles, Fake profile, Natural language processing (NLP), Machine learning, Support Vector Machine (SVM).

I. INTRODUCTION

When using product networks, different people share different amounts of their personal information. Having our information entirely or partially exposed to the public makes us unique targets for different types of attacks, the worst of which could be identity theft. Identity theft occurs when someone uses an individual's information for personal gain or purpose. In recent years, online identity theft has been a major problem, affecting millions of people worldwide. Victims of identity theft may suffer different types of consequences, such as losing time and money, going to jail, having their public image ruined, or damaging their relationships with friends and loved ones.

Today, the vast majorities of product networks do not verify the accounts of regular users and have very weak privacy and security policies. In fact, most product network applications default their settings to minimal privacy, making them a perfect platform for fraud and abuse. Product networking services have facilitated identity theft and impersonation attacks for both serious and naive attackers. The issues involving product networking, such as privacy, online bullying, misuse, and trolling, are often used by false profiles on product networking sites. False profiles are profiles that are not genuine. The false Facebook profiles are often involved in malicious and unwanted activities, causing problems for product network users. Individuals create fake profiles for product engineering, online impersonation to defame a person, promoting and campaigning for a person or a group of people. Facebook has its own security system to protect user credentials from spamming and phishing, known as Facebook Immune System (FIS). However, FIS has not been able to detect fake profiles created on Facebook by users to a larger extent.

II. ARTIFICIAL INTELLIGENCE

Defining AI?

Artificial intelligence (AI) refers to the ability of a computer program or machine to think and learn, as well as a field of study that aims to make computers intelligent. As machines become more capable, mental abilities once thought to require intelligence are no longer included in the definition. AI is a computer science area that focuses on creating machines that can perform tasks and react like humans, such as face recognition, learning, planning, decision-making, etc.

Simply put, artificial intelligence is the use of computer programming to simulate human thought and action by analyzing data and surroundings, anticipating and solving problems, and learning or self-teaching to adapt to a variety of tasks.

III. MACHINE LEARNING

What is machine learning?

Machine learning is a rapidly growing technology that enables computers to learn automatically from past data. It uses various algorithms to build mathematical models and make predictions based on historical information. Machine learning has a variety of applications, including image recognition, speech recognition, email filtering, Facebook auto-tagging, and recommender systems.

Machine learning is a subset of artificial intelligence that focuses on developing algorithms that allow computers to learn from data and past experiences independently. The term "machine learning" was first introduced by Arthur Samuel in 1959.

IV. CLASSIFICATION OF MACHINE LEARNING

1. At a broad level, machine learning can be classified into three types:
2. Supervised learning
3. Unsupervised learning
4. Reinforcement learning

1. SUPERVISED LEARNING

Supervised learning is a type of machine learning that involves providing labelled sample data to the machine learning system to train it, and subsequently predict the output. The system creates a model using labelled data to comprehend the datasets and learn about each data point. Once the training and processing are complete, we test the model by providing a sample data to check whether it accurately predicts the output. The objective of supervised learning is to map input data to the output data. Supervised learning can be further classified into two categories of algorithms:

- Classification
- Regression

2. UNSUPERVISED LEARNING

Unsupervised learning is a type of machine learning in which a machine learns without any supervision from humans. In unsupervised learning, the algorithm is provided with a set of data that has not been labeled or categorized. The algorithm then needs to act on that data without any supervision to find meaningful patterns or relationships within the data. The ultimate goal of unsupervised learning is to restructure the input data into new features or groups of objects with similar patterns. There are two main categories of unsupervised learning algorithms:

- Clustering
- Association

3. REINFORCEMENT LEARNING

Reinforcement learning is a machine learning method that rewards desired behaviors and punishes undesired ones. The method involves an agent that can perceive and interpret its environment, take actions, and learn through trial and error.

V. LITERATURE SURVEY

The consumption of news through has become increasingly popular in recent years due to its fast dissemination, low cost and easy access. However, the quality of news is often considered lower than traditional news outlets, resulting in the spread of large amounts of fake news. Detecting fake news is crucial and has been gaining more attention due to its detrimental effects on individuals and society.

The current methods of detecting fake news only from the content are not satisfactory. Therefore, it is suggested that user product engagements should be incorporated as auxiliary information to improve fake news detection. This requires a deeper understanding of the correlation between user profiles and fake news.

In this paper, real-world datasets have been constructed to measure users' trust levels on fake news, and representative groups of both "experienced" users who can recognize fake news items as false and "naïve" users who are more likely to believe fake news have been selected. A comparative analysis has been performed on explicit and implicit profile features of these user groups, which reveals their potential to differentiate fake news.

As professional networks like LinkedIn continue to grow, it becomes increasingly valuable for individuals to have their profiles noticed. However, this also leads to an increased risk of unethical behaviour, such as creating fake profiles. Fake profiles can harm the reputation of the network and lead to wasted time and effort in building connections based on false information. Unfortunately, identifying fake profiles can be difficult, and most methods require data that is not publicly available for LinkedIn profiles. To address this challenge, our research has identified the minimum amount of profile data necessary to identify fake profiles on LinkedIn and developed a data mining approach to do so. Our method is highly accurate, achieving 87% accuracy and 89% True Negative Rate even with limited profile data. This is comparable to results obtained from larger data sets and more extensive profile information, and provides an improvement of approximately 14% accuracy compared to similar approaches.

Product networking platforms, such as Twitter and Facebook, have experienced tremendous growth in the past decade and have attracted millions of users. They have become a preferred means of communication, but this popularity has also attracted various malicious entities, such as spammers. The growing number of users has also created the problem of fake accounts. False and fake identities are often involved in malicious activities, such as spreading abuse, misinformation, spamming, and artificially inflating the number of users to promote and sway public opinion. Detecting these fake identities is important to protect genuine users from malicious intent. To address this issue, we propose using a feature-based approach to identify fake profiles. We have used twenty-four features to efficiently identify fake accounts. To verify our classification results, we used three classification algorithms. Experimental results show that our model was able to achieve 87.9% accuracy using the Random Forest algorithm. Therefore, our proposed approach is efficient in detecting fake profiles.

The method aims to safeguard user privacy in an online product network. It involves selecting negative examples of fake profiles and positive examples of legitimate profiles from the existing user database of the product network. Next, a predefined set of features is extracted for each selected fake and legitimate profile. This is done by dividing the friends or followers of the chosen examples into communities and analyzing the relationships between each node within and between the communities. Classifiers that can detect other fake profiles based on their features are created and trained using supervised learning.

The creation of fake profiles is a significant problem and can lead to a range of malicious activities. This paper focuses on identifying fake profiles and the different methods used to detect them. Approaches to identifying fake profiles can be categorized into those that analyze profile data and those that analyze individual accounts. Fake profile creation on product networks is particularly harmful, and it is important to detect these profiles before the user is notified. There have been various algorithms and methods proposed for detecting fake profiles in the literature. This paper also highlights the role of fake identities in advanced persistent threats. To accurately identify fake and genuine profiles, we will evaluate the effectiveness of three supervised machine learning algorithms: Random Forest (RF), Decision Tree (DT-J48), and Naïve Bayes (NB).

VI. EXISTING SYSTEM

There are several challenges associated with implementing a fraud detection system, and one of the biggest problems is the lack of both experimental results in literature and real-world data for academic researchers to conduct experiments on. The sensitive financial data associated with fraud must be kept confidential to ensure customer privacy, which limits the availability of data for research purposes. To ensure accurate detection, a fraud detection system should possess certain properties. It must be capable of handling skewed distributions since a very small percentage of all credit card transactions are fraudulent. The system should also be able to handle noise, which refers to errors present in the data, such as incorrect dates. Another problem is overlapping data, where some transactions may resemble fraudulent transactions when they are actually legitimate, and vice versa.

A fraud detection system should be able to adapt to new types of fraud as successful fraud techniques become less effective over time due to their increasing familiarity. The system should also rely on good metrics for evaluating the classifier system; overall accuracy is not a suitable metric for skewed distributions since even with high accuracy, most fraudulent transactions can still be misclassified.

VII. PROPOSED SYSTEM

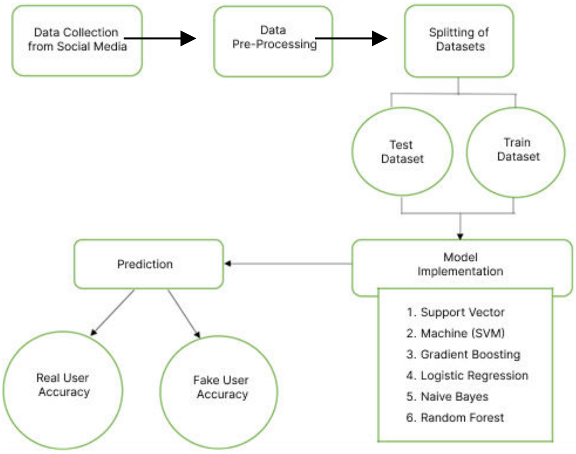
After conducting a thorough literature survey, it has been found that there are multiple ways to detect fake profiles on product networks. Two of the most common approaches are Machine Learning and Block chain. In order to identify the users who pose a threat in product networks, we need to classify their profiles. By doing so, we can distinguish between genuine and fake profiles. There are different classification methods that have been traditionally used for detecting fake profiles, but the accuracy rate needs to be improved.

The machine learning and natural language processing system that can detect false profiles on online product networks. To increase the accuracy of detecting fake profiles, we have incorporated five algorithms, namely Support Vector Machine (SVM), Random Forest classifier, Gradient Boost classifier, Naïve Bayes, and Logistic Regression algorithm. The system provides accuracy values, classification reports, and confusion matrices in the final prediction. Our proposed system helps in evaluating the best model for improving the detection accuracy rate of fake profiles.

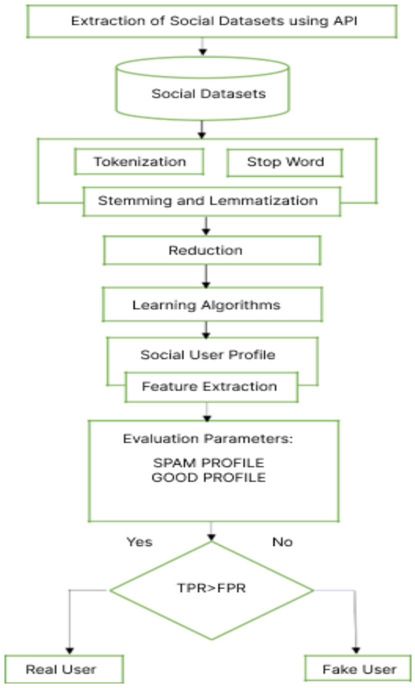
VIII. SYSTEM ARCHITECTURE

To analyze, who are encouraging threats in product network we need to classify the product networks profiles of the users. From the classification, we can get the genuine profiles and fake profiles on the product networks. Traditionally, we have different classification methods for detecting the fake profiles on the product networks. But we need to improve the accuracy rate of the fake profile detection in the product networks and the work flow contains the following modules:

- Support Vector Machine (SVM)
- Naïve Bayes
- Random Forest
- Correlation Heat map
- Confusion Matrix



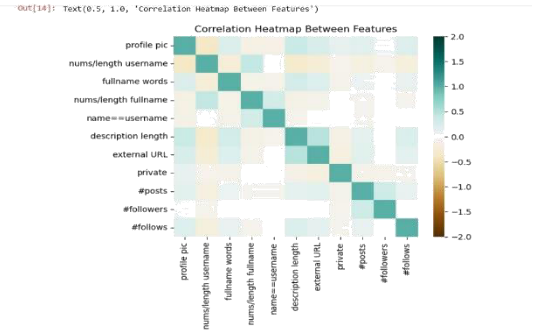
System architecture



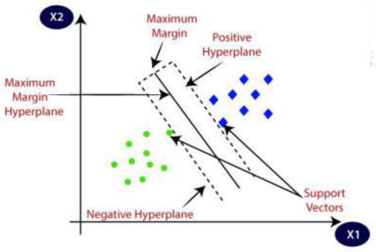
Data Flow (DF) Diagram

IX. MODULES

SUPPORT VECTOR MACHINE (SVM)



```
Out[14]: Text(0.5, 1.0, 'Correlation Heatmap Between Features')
```



Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called hyper plane.SVM chooses the extreme points/vectors that help in creating the hyper plane. These extreme cases are called as support vectors, and hencealgorithm is termed as Support Vector Machine. Consider the below diagram in which there are two different categories that are classified using a decision boundary or hyper plane:

HYPER PLANE

There can be multiple lines/decision boundaries to segregate the classes in n-dimensional space. This best boundary is known as the hyper plane of SVM. The dimensions of the hyper plane depend on the features present in the dataset, which means if there are 2 features (as shown in image), then hyper plane will be a straight line. And if there are 3 features, then hyper plane will be a 2-dimension plane. We always create a hyper plane that has a maximum margin.

SUPPORT VECTORS

The data points or vectors that are the closest to the hyper plane and which affect the position of the hyper plane are termed as Support Vector. Since these vectors support the hyper plane, hence called a Support vector.

NAIVE BAYES

Naïve Bayes algorithm is a supervised learning algorithm, which is based on Bayes theorem and used for solving classification problems. It is mainly used in text classification that includes a high-dimensional training dataset. Naïve Bayes Classifier is one of the simple and most effective Classification algorithms which helps in building the fast machine learning models that can make quick predictions. It is a probabilistic classifier, which means it predicts on the basis of the probability of an object. Some popular examples of Naïve Bayes Algorithm are spam filtration, Sentimental analysis, and classifying articles.

RANDOM FOREST

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. Before understanding the working of the random forest, we must look into the ensemble technique. Ensemble simply means combining multiple models. Thus, a collection of models is used to make predictions rather than an individual model. Ensemble uses two types of methods:

1. BAGGING.
2. BOOSTING.

CORRELATION HEAT MAP

Correlation heatmaps are a type of plot that visualize the strength of relationships between numerical variables. Correlation plots are used to understand which variables are related to each other and the strength of this relationship. A correlation plot typically contains a number of numerical variables, with each variable represented by a column. The rows represent the relationship between each pair of variables. The values in the cells indicate the strength of the relationship, with positive values indicating a positive relationship and negative values indicating a negative relationship.

CONFUSION MATRIX

A confusion matrix is used to measure the performance of a classifier in depth. In this simple guide to Confusion Matrix, we will get to understand and learn confusion matrices better. A confusion matrix presents a table layout of the different outcomes of the prediction and results of a classification problem and helps visualize its outcomes.

X. CONCLUSION AND FUTURE ENHANCEMENT

In the project, we proposed machine learning algorithms along with natural language processing techniques. By using these techniques, we can easily detect the fake profiles from the sites. In this project we took the dataset to identify the fake profiles. 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 project.

For future development, there are still some challenges to be addressed in using sophisticated machine learning algorithms can enhance the detection capabilities of block chain-based systems. There are several ways to enhance the proposed block chain solution for detecting and removing fake profiles. One approach is to utilize advanced machine learning models, including Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), to improve the accuracy of fake profile detection. By embracing advancements in machine learning, data analytics, real-time monitoring, community participation, and privacy-preserving technologies, the proposed block chain solution can evolve into a more robust framework for combating fake profiles and enhancing user trust and security in online interactions.

REFERENCES

1. Romanov, Aleksei, Alexander Semenov, Oleksiy Mazhelis, and Jari Veijalainen. "Detection of fake profiles in - Literature review." In International Conference on Web Information Systems and Technologies, vol. 2, pp. 363-369. SCITEPRESS, 2018.
2. Adikari, Shalinda, and Kaushik Dutta. "Identifying fake profiles in linkedin." arXiv preprint arXiv:2006.01381 (2020).
3. Kaubiyal, Jyoti, and Ankit Kumar Jain. "A featurebased approach to detect fake profiles in Twitter." In Proceedings of the 3rd International Conference on Big Data and Internet of Things, pp. 135-139. 2019.
4. Elovici, Yuval, F. I. R. E. Michael, and Gilad Katz. "Method for detecting spammers and fake profiles in product networks." U.S. Patent 9,659,185, issued May 23, 2019
5. Elyusufi, Y. and Elyusufi, Z., 2019, October. Product networks fake profiles detection using machine learning algorithms. In The Proceedings of the Third International Conference on Smart City Applications (pp. 30-40). Springer, Cham.
6. Ozbay, F.A. and Alatas, B., 2020. Fake news detection within online using supervised artificial intelligence algorithms. Physica A: Statistical Mechanics and its Applications, 540.

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING