Detection of Malicious Social Bots Using Learning Automata with URL Features in Twitter Network

Peddireddy Likhitha

Department of Computer Science, Sri Venkateswara University, Tirupati

Abstract— *Malicious social bots generate fake tweets and automate their social relationships either by pretending like a follower or by creating multiple fake accounts with malicious activities. Moreover, malicious social bots post shortened malicious URLs in the tweet in order to redirect the requests of online social networking participants to some malicious servers. Hence, distinguishing malicious social bots from legitimate users is one of the most important tasks in the Twitter network. To detect malicious social bots, extracting URL-based features (such as URL redirection, frequency of shared URLs, and spam content in URL) consumes less amount of time in comparison with social graph-based features (which rely on the social interactions of users). Furthermore, malicious social bots cannot easily manipulate URL redirection chains. In this article, a learning automata-based malicious social bot detection (LA-MSBD) algorithm is proposed by integrating a trust computation model with URL-based features for identifying trustworthy participants (users) in the Twitter network. The proposed trust computation model contains two parameters, namely, direct trust and indirect trust. Moreover, the direct trust is derived from Bayes' theorem, and the indirect trust is derived from the Dempster– Shafer theory (DST) to determine the trustworthiness of each participant accurately. Experimentation has been performed on two Twitter data sets, and the results illustrate that the proposed algorithm achieves improvement in precision, recall, F-measure, and accuracy compared with existing approaches for MSBD.*

Keywords: Learning automata (LA), malicious social bots, online social networks (OSNs), trust.

I. INTRODUCTION

Malicious social bot is a software program that pretends to be a real user in online social networks (OSNs) . Moreover, malicious social bots perform several malicious attacks, such as spread social spam content, generate fake identities, manipulate online ratings, and perform phishing attacks. In Twitter, when a participant (user) wants to share a tweet containing URL(s) with the neighbouring participants (i.e., followers or followees), the participant adapts URL shortened service (i.e., bit.ly in order to reduce the length of URL (because a tweet is restricted up to 140 characters). Moreover, a malicious social bot may post shortened phishing URLs in the tweet.

1.1 Data Mining:

Data mining is a process used by companies to turn raw data into useful information. By using software to look for patterns in large batches of data, businesses can learn more about their customers to develop more effective marketing strategies, increase sales and decrease costs. Data mining depends on effective data collection, warehousing, and computer processing. Data mining processes are used to build machine learning models that power applications including search engine technology and website recommendation programs.

Data mining is the act of automatically searching for large stores of information to find trends and patterns that go beyond simple analysis procedures. Data mining utilizes complex mathematical algorithms for data segments and evaluates the probability of future events. Data Mining is also called Knowledge Discovery of Data (KDD). Data Mining is a process used by organizations to extract specific data from huge databases to solve business problems. It primarily turns raw data into useful information. Data Mining is similar to Data Science carried out by a person, in a specific situation, on a particular data set, with an objective. This process includes various types of services such as text mining, web mining, audio and video mining, pictorial data mining, and social media mining. It is done through software that is simple or highly specific. By outsourcing data mining, all the work can be done faster with low operation costs. Specialized firms can also use new technologies to collect data that is impossible to locate manually. There are tonnes of information available on various platforms, but very little knowledge is accessible. The biggest challenge is to analyse the data to extract important information that can be used to solve a problem or for company development. There are many powerful instruments and techniques available to mine data and find better insight from it.

1.2 Example of Data Mining

Grocery stores are well-known users of data mining techniques. Many supermarkets offer free loyalty cards to customers that give them access to reduced prices not available to non-members. The cards make it easy for stores to track who is buying what, when they are buying it and at what price. After analyzing the data, stores can then use this data to offer customers coupons targeted to their buying habits and decide when to put items on sale or when to sell them at full price.

Data mining can be a cause for concern when a company uses only selected information, which is not representative of the overall sample group, to prove a certain hypothesis.

1.3 Types of Data Mining

1.3.1 Relational Database:

A relational database is a collection of multiple data sets formally organized by tables, records, and columns from which data can be accessed in various ways without having to recognize the database tables. Tables convey and share information, which facilitates data searchability, reporting, and organization.

1.3.2 Data warehouses:

A Data Warehouse is the technology that collects the data from various sources within the organization to provide meaningful business insights. The huge amount of data comes from multiple places such as Marketing and Finance. The extracted data is utilized for analytical purposes and helps in decision- making for a business organization. The data warehouse is designed for the analysis of data rather than transaction processing.

1.3.3 Data Repositories:

The Data Repository generally refers to a destination for data storage. However, many IT professionals utilize the term more clearly to refer to a specific kind of setup within an IT structure. For example, a group of databases, where an organization has kept various kinds of information.

1.3.4 Object-Relational Database:

A combination of an object-oriented database model and relational database model is called an object-relational model. It supports Classes, Objects, Inheritance, etc.

One of the primary objectives of the Object-relational data model is to close the gap between the Relational database and the object-oriented model practices frequently utilized in many programming languages, for example, C++, Java, C#, and so on.

1.3.5 Transactional Database:

A transactional database refers to a database management system (DBMS) that has the potential to undo a database transaction if it is not performed appropriately. Even though this was a unique capability a very long while back, today, most of the relational database systems support transactional database activities.

1.4 Advantages of Data Mining

- The Data Mining technique enables organizations to obtain knowledge-based data.
- Data mining enables organizations to make lucrative modifications in operation and production.
- Compared with other statistical data applications, data mining is a cost-efficient.
- Data Mining helps the decision-making process of an organization.
- It Facilitates the automated discovery of hidden patterns as well as the prediction of trends and behaviours.
- It can be induced in the new system as well as the existing platforms.
- It is a quick process that makes it easy for new users to analyse enormous amounts of data in a short time.

1.5 Data Mining Applications

Data Mining is primarily used by organizations with intense consumer demands- Retail, Communication, Financial, marketing company, determine price, consumer preferences, product positioning, and impact on sales, customer satisfaction, and corporate profits. Data mining enables a retailer to use point-of-sale records of customer purchases to develop products and promotions that help the organization to attract the customer.

1.5.1 Data Mining in Healthcare:

Data mining in healthcare has excellent potential to improve the health system. It uses data and analytics for better insights and to identify best practices that will enhance health care services and reduce costs. Analysts use data mining approaches such as Machine learning, Multi-dimensional database, Data visualization, Soft computing, and statistics. Data Mining can be used to forecast patients in each category. The procedures ensure that the patients get intensive care at the right place and at the right time. Data mining also enables healthcare insurers to recognize fraud and abuse.

1.5.2 Data Mining in Market Basket Analysis:

Market basket analysis is a modelling method based on a hypothesis. If you buy a specific group of products, then you are more likely to buy another group of products. This technique may enable the retailer to understand the purchase behaviour of a buyer. This data may assist the retailer in understanding the requirements of the buyer and altering the store's layout accordingly. Using a different analytical comparison of results between various stores, between customers in different demographic groups can be done.

1.5.3 Data mining in Education:

Education data mining is a newly emerging field, concerned with developing techniques that explore knowledge from the data generated from educational Environments. EDM objectives are recognized as affirming student's future learning behaviour, studying the impact of educational support, and promoting learning science. An organization can use data mining to make

precise decisions and also to predict the results of the student. With the results, the institution can concentrate on what to teach and how to teach.

1.5.4 Data Mining in Manufacturing Engineering:

Knowledge is the best asset possessed by a manufacturing company. Data mining tools can be beneficial to find patterns in a complex manufacturing process. Data mining can be used in system-level designing to obtain the relationships between product architecture, product portfolio, and data needs of the customers. It can also be used to forecast the product development period, cost, and expectations among the other tasks.

1.5.5 Data Mining in CRM (Customer Relationship Management):

Customer Relationship Management (CRM) is all about obtaining and holding Customers, also enhancing customer loyalty and implementing customer-oriented strategies. To get a decent relationship with the customer, a business organization needs to collect data and analyse the data. With data mining technologies, the collected data can be used for analytics.

1.5.6 Data Mining in Fraud detection:

Billions of dollars are lost to the action of frauds. Traditional methods of fraud detection are a little bit time consuming and sophisticated. Data mining provides meaningful patterns and turning data into information. An ideal fraud detection system should protect the data of all the users. Supervised methods consist of a collection of sample records, and these records are classified as fraudulent or non-fraudulent. A model is constructed using this data, and the technique is made to identify whether the document is fraudulent or not.

1.5.7 Data Mining in Lie Detection:

Apprehending a criminal is not a big deal, but bringing out the truth from him is a very challenging task. Law enforcement may use data mining techniques to investigate offenses, monitor suspected terrorist communications, etc. This technique includes text mining also, and it seeks meaningful patterns in data, which is usually unstructured text. The information collected from the previous investigations is compared, and a model for lie detection is constructed.

1.5.8 Data Mining Financial Banking:

The Digitalization of the banking system is supposed to generate an enormous amount of data with every new transaction. The data mining technique can help bankers by solving business-related problems in banking and finance by identifying trends, casualties, and correlations in business information and market costs that are not instantly evident to managers or executives because the data volume is too large or are produced too rapidly on the screen by experts. The manager may find these data for better targeting, acquiring, retaining, segmenting, and maintain a profitable customer.

1.6 Challenges of Implementation in Data mining

Although data mining is very powerful, it faces many challenges during its execution. Various challenges could be related to performance, data, methods, and techniques, etc. The process of data mining becomes effective when the challenges or problems are correctly recognized and adequately resolved.

1.6.1 Incomplete and noisy data:

The process of extracting useful data from large volumes of data is data mining. The data in the real-world is heterogeneous, incomplete, and noisy. Data in huge quantities will usually be inaccurate or unreliable. These problems may occur due to data measuring instrument or because of human errors. Suppose a retail chain collects phone numbers of customers who spend more than $ 500, and the accounting employees put the information into their system. The person may make a digit mistake when entering the phone number, which results in incorrect data. Even some customers may not be willing to disclose their phone numbers, which results in incomplete data. The data could get changed due to human or system error. All these consequences (noisy and incomplete data) make data mining challenging.

1.6.2 Data Distribution:

Real-worlds data is usually stored on various platforms in a distributed computing environment. It might be in a database, individual systems, or even on the internet. Practically, It is a quite tough task to make all the data to a centralized data repository mainly due to organizational and technical concerns. For example, various regional offices may have their servers

to store their data. It is not feasible to store, all the data from all the offices on a central server. Therefore, data mining requires the development of tools and algorithms that allow the mining of distributed data.

1.6.3 Complex Data:

Real-world data is heterogeneous, and it could be multimedia data, including audio and video, images, complex data, spatial data, time series, and so on. Managing these various types of data and extracting useful information is a tough task. Most of the time, new technologies, new tools, and methodologies would have to be refined to obtain specific information.

1.6.4 Performance:

The data mining system's performance relies primarily on the efficiency of algorithms and techniques used. If the designed algorithm and techniques are not up to the mark, then the efficiency of the data mining process will be affected adversely.

1.6.5 Data Privacy and Security:

Data mining usually leads to serious issues in terms of data security, governance, and privacy. For example, if a retailer analyses the details of the purchased items, then it reveals data about buying habits and preferences of the customers without their permission.

1.6.6 Data Visualization:

In data mining, data visualization is a very important process because it is the primary method that shows the output to the user in a presentable way. The extracted data should convey the exact meaning of what it intends to express. But many times, representing the information to the end-user in a precise and easy way is difficult. The input data and the output information being complicated, very efficient, and successful data visualization processes need to be implemented to make it successful.

II. PROPOSED SYSTEM

In this article, a learning automata-based malicious social bot detection (LA-MSBD) algorithm is proposed by integrating a trust computation model with URL-based features for identifying trustworthy participants (users) in the Twitter network. The proposed trust computation model contains two parameters, namely, direct trust and indirect trust. Moreover, the direct trust is derived from Bayes' theorem, and the indirect trust is derived from the Dempster– Shafer theory (DST) to determine the trustworthiness of each participant accurately. Experimentation has been performed on two Twitter data sets, and the results illustrate that the proposed algorithm achieves improvement in precision, recall, F-measure, and accuracy compared with existing approaches for MSBD.

Moreover, in our work, the belief values provided by multiple neighbouring participants are considered to be independent. The proposed LA-MSBD algorithm helps to detect malicious social bots accurately (in terms of precision, recall, F-measure, and accuracy) in Twitter. The major contributions are as follows.

1) Analyse the malicious behaviour of a participant by considering URL-based features, such as URL redirection, the relative position of URL, frequency of shared URLs, and spam content in URL.
2) Evaluate the trustworthiness of tweets (posted by each participant) by using the Bayesian learning and Dempster–Shafer theory (DST).

3) Design of an LA-MSBD algorithm by integrating a trust model with a set of URL-based features.
4) Performance evaluation of the proposed LA-MSBD algorithm using two Twitter data sets, namely, The Fake Project data set and Social Honeypot data set in terms of precision, recall, F-measure, and accuracy for MSBD in the Twitter network.

2.1 Advantages:

- High security and more effective.
- Executes for a finite set of learning actions to update the action probability value and achieves the advantages of incremental learning.
- The performance and the accuracy is high.

2.2 Algorithms:

Ciphertext:

Ciphertext is also known as encrypted or encoded information because it contains a form of the original plaintext that is unreadable by a human or computer without the proper cipher to decrypt it. Decryption, the inverse of encryption, is the process of turning ciphertext into readable plaintext.

Techniques:

- Stop word
- Stemming

III. SYSTEM ARCHITECTURE

3.1 Admin

- Login with correct user name and the password
- Add URL and view all the added URL
- View all the malware site and the user search keyword
- Prediction: View all the prediction result and the performance metrices
- Graph—view the malicious and the non-malicious URLs
- Logout

3.2 Search:

- User can search through the keyword, related to that keyword all URLs is shown.
- If the user clicks the particular URLs, if its correct URLs means it gives result.
- If the URLs is not Correct means, the page shows this URLs is blocked because this is malicious page.

3.3 Attacker:

- Login the account
- Add the keyword and the unwanted URL it means Malicious.
- Logout.

IV. APPLICATION AND THE FUTURE ENHANCEMENT

This detection of malicious bots' method is using other applications such as what's up, face book etc.

Furthermore, as a future research challenge, we would like to investigate the dependence among the features and its impact on MSDA,vadditional behaviours of malicious social bots will be further considered and the proposed detection approach will be extended and optimized to identify specific intentions and purposes of a broader range of malicious social bots.

Home:


DETECTION OF MALICIOUS SOCIAL BOTS USING
LEARNING AUTOMATA WITH URL FEATURES IN
TWITTER NETWORK

Admin:

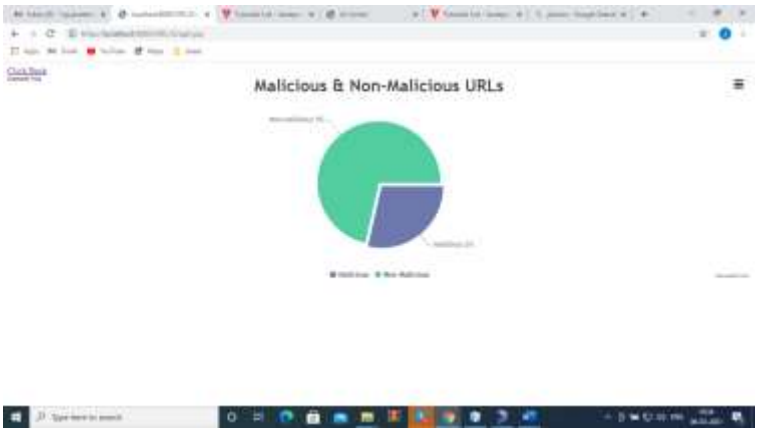

DETECTION OF MALICIOUS SOCIAL BOTS USING
LEARNING AUTOMATA WITH URL FEATURES IN
TWITTER NETWORK


DETECTION OF MALICIOUS SOCIAL BOTS USING
LEARNING AUTOMATA WITH URL FEATURES IN
TWITTER NETWORK

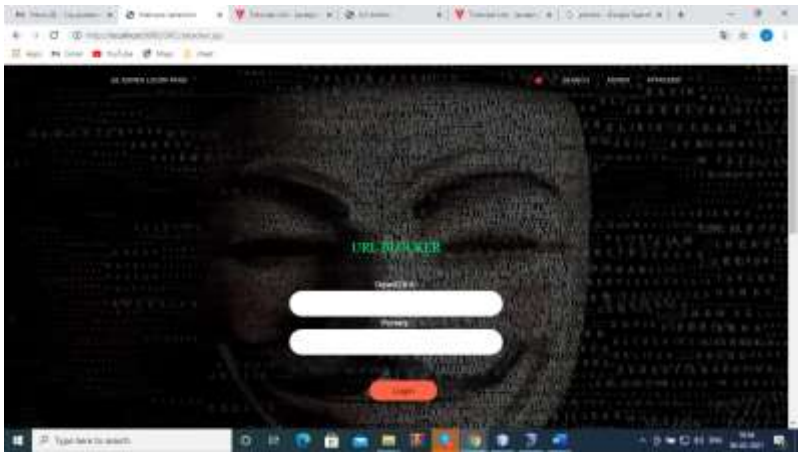

VIEW ATTACKER SITES
USER SEARCH AND ATTACKED SITES VIEWED HERE


MAXISONIX
USER IP AND USER SEARCHED KEYWORD VIEWED HERE


PREDICTION RESULT & PERFORMANCE METRICS
0.0
0.0
0.0
0.8


Malicious & Non-Malicious URLs

Attacker:


URL HACKER

Search

V. CONCLUSION

This article presents an LA-MSBD algorithm by integrating a trust computational model with a set of URL-based features for MSBD. In addition, we evaluate the trustworthiness of tweets (posted by each participant) by using the Bayesian learning and DST. Moreover, the proposed LA-MSBD algorithm executes a finite set of learning actions to update action probability value

(i.e., probability of a participant posting malicious URLs in the tweets). The proposed LA-MSBD algorithm achieves the advantages of incremental learning. Two Twitter data sets are used to evaluate the performance of our proposed LA-MSBD algorithm. The experimental results show that the proposed LA-MSBD algorithm achieves up to 7% improvement of accuracy compared with other existing algorithms. For the Fake Project and Social Honeypot data sets, the proposed LA-MSBD algorithm has achieved precisions of 95.37% and 91.77% for MSBD, respectively.

REFERENCES

[1] P. Shi, Z. Zhang, and K.-K.-R. Choo, "Detecting malicious social bots based on clickstream sequences," IEEE Access, vol. 7, pp. 28855–28862, 2019.

[2] M. Al-Janabi, E. D. Quincey, and P. Andras, "Using supervised machine learning algorithms to detect suspicious URLs in online social networks," in Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining, Jul. 2017, pp. 1104–1111.

[3] S. Lee and J. Kim, "WarningBird: A near real-time detection system for suspicious URLs in Twitter stream," IEEE Trans. Dependable Secure Comput., vol. 10, no. 3, pp. 183–195, May 2013.

[4] D. R. Patil and J. B. Patil, "Malicious URLs detection using decision tree classifiers and majority voting technique," Cybern. Inf. Technol., vol. 18, no. 1, pp. 11–29, Mar. 2018.

[5] H. Guo, S. Li, B. Li, Y. Ma, and X. Ren, "A new learning automatabased pruning method to train deep neural networks," IEEE Internet Things J., vol. 5, no. 5, pp. 3263–3269, Oct. 2018.

[6] A. A. Rahmanian, M. Ghobaei-Arani, and S. Tofighy, "A learning automata-based ensemble resource usage prediction algorithm for cloud computing environment," Future Gener. Comput. Syst., vol. 79, pp. 54–71, Feb. 2018.

[7] A. Moayedikia, K.-L. Ong, Y. L. Boo, and W. G. S. Yeoh, "Task assignment in microtask crowdsourcing platforms using learning automata," Eng. Appl. Artif. Intell., vol. 74, pp. 212–225, Sep. 2018.

[8] G. Lingam, R. R. Rout, and D. Somayajulu, "Learning automatabased trust model for user recommendations in online social networks," Comput. Electr. Eng., vol. 66, pp. 174–188, Feb. 2018.

[9] Manju, S. Chand, and B. Kumar, "Target coverage heuristic based on learning automata in wireless sensor networks," IET Wireless Sensor Syst., vol. 8, no. 3, pp. 109–115, Jun. 2018.

[10] N. Abokhodair, D. Yoo, and D. W. McDonald, "Dissecting a social botnet: Growth, content and influence in Twitter," in Proc. 18th ACM Conf. Comput. Supported Cooperat. Work Social CompuSt., 2015, pp. 839–851