IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 4, April 2022

DOI: 10.17148/IJARCCE.2022.114177

SECURE CONNECT

Reshma Farhin J[1], Neenupriya K[2], Pavithra M[3], Saalai Ezhilarasi R[4]

[1]Assistant Professor, Department Of CSE, Adhiyamaan College of Engineering, Hosur, India.

[1,2,3]UG Scholar, Department Of CSE, Adhiyamaan College of Engineering, Hosur, India.

Abstract: Information mining of open-source knowledge on the Web has become an undeniably significant point over a wide scope of spaces, for example, business, law requirement, military, and online protection. Text mining endeavours use characteristic language handling to change unstructured web content into organized structures that can drive different machine learning applications and information ordering administrations. For instance, applications or text mining in online protection have created a scope of danger insight benefits that serve the IT business. In any case, a less contemplated issue is that of computerizing the recognizable proof of semantic irregularities among different content information sources. In this paper, we present Secure connect, another irregularity checking framework for recognizing semantic irregularities inside the network safety space. In particular, we inspect the issue of recognizing specialized irregularities that emerge in the utilitarian portrayals of open-source malware danger detailing data. Our assessment, utilizing a huge number of relations determined from online malware danger reports, shows the capacity of secure connect to recognize the presence of irregularities.

Keywords: secure connect, Framework, semantic irregularities, malware danger, text mining, tweet analysis.

I.INTRODUCTION

Social media networking sites are more popular over Internet. The Internet users spend more amount of time on social media sites like Twitter, Facebook, Instagram and LinkedIn etc. The social media networking users share their ideas, opinions, information and make new friends. Social networking sites provide large amount of valuable information to the users. This large amount of information in social media attracts spammers to misuse information. These spammers create fake accounts and spread irrelevant information to the genuine users. The spam message information may be advertisements, malicious links to disturb the natural users. This spam data in social media is a very serious problem. Spam detection in social media networking sites is critical process. To extract spam messages in social media various spam detection methodologies are developed by researchers. In this paper we proposed an ensemble methodology for identification spam on Twitter social media network. In this methodology we used Decision tree induction algorithm, Naïve bayes algorithm and bloom filter to construct a model. As part of this approach, we compare the classification results of any two classification algorithms, if both classifiers predict the same result, then we finalize the class of tweet under investigation. If the predicted classes of both classification algorithms differ, then we use the prediction of third algorithm as the final class label of tweet.

II.EXISTING SYSTEM

Existing plans can't be straightforwardly applied to screening security explicit data for the accompanying three reasons. To start with, these current plans centre on checking the rightness between effectively organized (arranged) information (e.g., stature of mountain, writers of a book), and they don't have any significant bearing to assessing unstructured information. Tragically, the opensource data used to drive CTI administrations depends on unstructured web text, written in characteristic language. In this way, a few explicit language preparing strategies, for example, named element acknowledgment (NER) and connection extraction (RE), for the network protection space are expected to separate organized qualities for security data.

Disadvantages Of Existing System:

- Focus on verifying the correctness between already-structured (formatted) data (e.g., height of mountain, authors of a book), and they do not apply to evaluating unstructured data
- The open-source information used to drive CTI services relies on unstructured web text, written in natural language. Therefore, some specific language processing techniques, such as named entity recognition (NER) and relation extraction (RE) they are required.
- Security information extracted from unstructured texts still requires additional formalization. If we extract data (e.g., nouns and verbs) from unstructured texts, many different shapes of terms represent the same meaning.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 4, April 2022

DOI: 10.17148/IJARCCE.2022.114177

III. PROPOSED SYSTEM

This paper proposes an orderly way to deal with identifying irregularities in security data from various freely accessible sources. Our framework, Secure Connect, deconstructs this issue into three fundamental assignments:

- The extraction of organized information from unstructured content,
- An information refinement measure on the removed information, and
- The definition of semantic associations between malware monikers.

To begin with, the diagram constructor, comprising of an element tagger, connection developer, the diagram constructor fabricates malware diagrams, which comprise just of relations that contain the equivalent malware name. Second, the information formatter plays out an information refinement measure on the information inside the malware diagrams. The information separated from unstructured content is communicated with different states of terms to pass on comparable implications.

Advantages:

- We present Secure Connect that derives structured relations from unstructured text about the main features of malware.
- Secure Connect addresses the inconsistency between any forms of entities, structured or unstructured text, by inferring common terms for different words with similar or equivalent meaning.
- Evaluate Secure Connect against 470K security reports and find many inconsistencies in the four main features of malware.

IV. ARCHITECTURE DESIGN

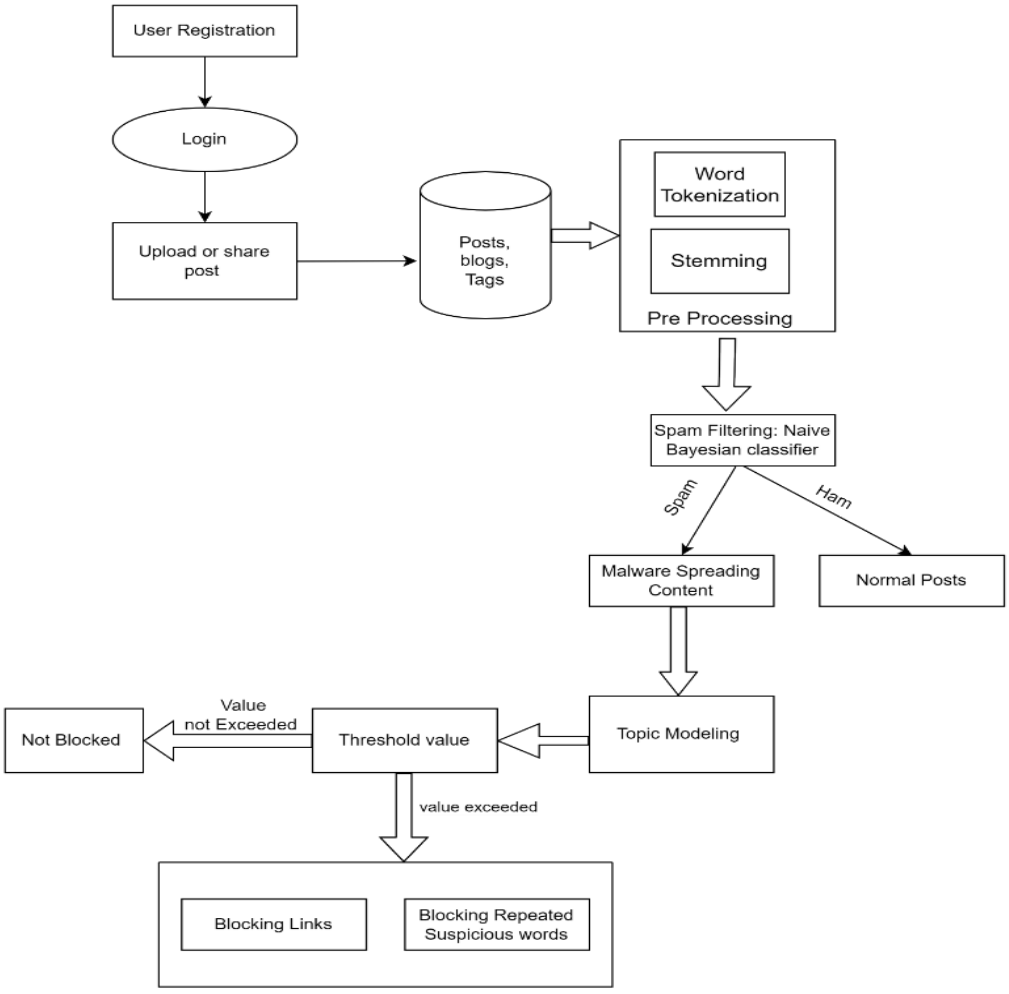

Figure 1: Architecture diagram

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 4, April 2022

DOI: 10.17148/IJARCCE.2022.114177

The framework of our spam detection methodology is shown in Figure. In our proposed methodology we used different steps to detect spam in Twitter data. The previous spam detection analysts used various spam detection methodologies. Each approach uses its own data set and features for classification of data. Various spam detection approaches used different kind of features like user-based features, content-based features, network- based features and location- based features etc. Initially we train and test the classifiers individually with individual features on Twitter dataset. Next, we train and test the ensemble approach on individual features and combination of user based and content -based features. Later we did the same experiments on Twitter data set with cross validation approach. Ensemble cross validation approach has outperformed compare to other performed approaches.

Dataset

Our research is conducted based on users perspective and content perspective. We get all the tweets from normal users instead of crawling public tweets. We randomly pickup 25 normal users from our Twitter and crawl tweets of the publishers they follow.

Labelling

Tweets Once the Twitter data was gathered, the next task was to develop a collection of tweets labelled into spam and ham groups. These categories could be used to train and test our classifier.

Feature selection

In the proposed spam detection method 11 features are identified. The feature set is classified into two categories namely user- based features and content -based features User Based Features: User based features are used to describe the behaviour of users in twitter. These features are based on user relationships and properties of user accounts in twitter dataset. Generally, in social media networks users can develop their own social networks with other users. In social network one user follows other users and allows other users to follow him. Spammers want to follow many profiles to spread misinformation to them, so they try to follow large number of users to spread misinformation. Generally, we consider, the number of users following is more than number of users following him, such user account is considered as spam account. Here we are using different user- based features to construct a model.

Tweet Frequency

Generally, the tweet frequency of spammers is greater than genuine twitter user.

Spam words

we use specific spam words and count their occurrence in tweets of users. The spammers use this spam words and spread misinformation to the users.

V. MODULES

The modules present in this project,

- Detecting spam in trending topic
- Fake user identification
- Ham learning algorithm
- Fake content-based spammer detection

Detecting spam in trending topic

- The collection of tweets with respect to trending topics on social media. After storing the tweets in a particular file format, the tweets are subsequently analysed. Labelling of spam is performed to check through all datasets that are available to detect the malignant social media.
- Feature extraction separates the characteristics construct based on the language model that uses tweets as a tool and helps in determining whether the tweets are fake or not.
- The classification of data set is performed by shortlisting the set of tweets that is described by the set of features provided to the classifier to instruct the model and to acquire the knowledge for spam detection.
- The spam detection uses the classification technique to accept tweets as the input and classify the spam and non-spam examined the degree to which the trending affairs in twitter are exploited by spammers.
- Although numerous methods to detect the spam have been proposed, the research on determining the effects of spam on twitter trending topics has attained only limited attention of the researchers.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 4, April 2022

DOI: 10.17148/IJARCCE.2022.114177

Fake user identification

- A categorization method is proposed to detect spam accounts on social media. The dataset used in the study was collected manually.
- The classification is performed by analysing user-name, profile, number of friends and followers, content of tweets, description of account, and number of tweets.
- The dataset comprised 501 fake and 499 real accounts, where 16 features from the information that were obtained from the social website were identified. Two experiments were performed for classifying fake accounts.
- The first experiment uses the naïve bayes learning algorithm on the social website dataset including all aspects without discretization, whereas the second experiment uses the naïve bayes learning algorithm on the social website dataset after the discretization.
- Proposed a hybrid technique that utilizes user-based, content-based, and graph-based characteristics for spammer profiles detection. A model is proposed to differentiate between the non-spam and spam profiles using three characteristics.
- The proposed technique was analysed using social website dataset with users and approximately tweets. The goal is to attain higher efficiency and preciseness by integrating all these characteristics. User-based features are established because of relationship and properties of user accounts.
- It is essential to append user-based features for the spam detection model. As these features are related to user accounts, all attributes, which were linked to user accounts, were identified.

Ham learning algorithm

- Ham is wordnet that is not spam. In other words, non-spam or good positive words. It should be considered a shorter, snappier synonym for "non-spam. Its usage is particularly common among anti-spam software developers and not widely known elsewhere in general it is probably better to use the term non-spam, instead.
- Content attributes have the property of the wordings of tweets that are posted by the users which gather features that are relevant to the way users write tweets. On the other hand, user behaviour attributes gather particular features of the behaviour of users in the context of the posting frequency, interaction, and impact on social website. The following attributes are considered as user characteristics, which include the total number of followers and following, account age, number of tags, fraction of followers per followings, number of times users replied, number of tweets received, average, maximum, minimum, and median time among user tweets, and daily and weekly tweets

Fake content-based spammer detection

- Performed an in-depth characterization of the components that are affected by the rapidly growing malicious content.
- It was observed that a large number of people with high social profiles were responsible for circulating fake news.
- To recognize the fake accounts, the authors selected the accounts that were built immediately after the boston blast and were later banned by social media due to violation of terms and conditions.
- This dataset is known as the largest dataset of boston blast. The authors performed the fake content categorization through temporal analysis where temporal distribution of tweets is calculated based on the number of tweets posted per hour ours.
- Fake tweet user accounts were analysed by the activities performed by user accounts from where the spam tweets were generated. It was observed that most of the fake tweets were shared by people with followers.

VI. RESULTS

We have shown that our proposed solution is feasible and is much better classification result than other existing methodologies. One issue of our proposed approach is it takes more amount of time for model training. The feature extraction in our proposed solution is based on manual selection. The feature extraction process in our approach might be low adaptive and costive.

VII. CONCLUSION

In this paper, we have introduced an ensemble-based spam detection methodology for social networks. This methodology considers user-based features and content-based features and apply them into Naïve bayes algorithm for spam detection. These algorithms are implemented individually without cross validation and with cross validation. The ensemble approach is also implemented without cross validation and with cross validation.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 4, April 2022

DOI: 10.17148/IJARCCE.2022.114177

VIII. REFERENCES

[1] H. Jo, J. Kim, P. Porras, V. Yegneswaran and S. Shin, "GapFinder: Finding Inconsistency of Security Information From Unstructured Text," in IEEE Transactions on Information Forensics and Security, vol. 16, pp. 86-99, 2021, doi: 10.1109/TIFS.2020.3003570.

[2] J. Choi and C. Jeon, "Cost-Based Heterogeneous Learning Framework for Real Time Spam Detection in Social Networks With Expert Decisions," in IEEE Access, vol. 9, 103573-103587, 2021, doi: 10.1109/ACCESS.2021.3098799.

[3] N. Govil, K. Agarwal, A. Bansal and A. Varshney, "A Machine Learning based Spam Detection Mechanism," 2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC), 2020, pp. 954-957, doi: 10.1109/ICCMC48092.2020.ICCMC-000177.

[4] W. Z. Khan, M. K. Khan, F. T. Bin Muhaya, M. Y. Aalsalem and H. Chao, "A Comprehensive Study of Email Spam Botnet Detection," in IEEE Communications Surveys & Tutorials, vol. 17, no. 4, pp. 2271-2295, Fourthquarter 2015, doi: 10.1109/COMST.2015.2459015.

[5] E. Stai, E. Milaiou, V. Karyotis and S. Papavassiliou, "Temporal Dynamics of Information Diffusion in Twitter: Modeling and Experimentation," in IEEE Transactions on Computational Social Systems, vol. 5, no. 1, pp. 256-264, March 2018, doi: 10.1109/TCSS.2017.2784184.

[6] J. Song, S. Lee and J. Kim, "Inference Attack on Browsing History of Twitter Users Using Public Click Analytics and Twitter Metadata," in IEEE Transactions on Dependable and Secure Computing, vol. 13, no. 3, pp. 340-354, 1 May-June 2016, doi: 10.1109/TDSC.2014.2382577.

[7] 70 Chakkarwar, V., and Tamane, S. C. (2020). “Quick insight of research literature using topic modeling,” in Smart Trends in Computing and Communications. Smart Innovation, Systems and Technologies, Vol. 165, eds Y. D. Zhang, J. Mandal, C. So-In, and N. Thakur (Singapore: Springer), 189–197. doi: 10.1007/978-981-15-0077-0_20

[8] Nugroho, R., Paris, C., Nepal, S., Yang, J., and Zhao, W. (2020). A survey of recent methods on deriving topics from twitter: algorithm to evaluation. Knowl. Inf. Syst. 62, 2485–2519. doi: 10.1007/s10115-019-01429-z

[9] Yang, Y., Yao, Q., and Qu, H. (2017). VISTopic: a visual analytics system for making sense of large document collections using hierarchical topic modeling. Visual Inform. 1, 40–47. doi: 10.1016/j.visinf.2017.01.005

[10] Xu, A., Qi, T., and Dong, X. (2019). “Analysis of the douban online review of the mcu: based on LDA topic model,” in 2nd International Symposium on Big Data and Applied Statistics. Journal of Physics: Conference Series, Vol. 1437 (Dalian). doi: 10.1088/1742-6596/1437/1/012102.