Mathematical Statistician and Engineering Applications
ISSN: 2094-0343
2326-9865

Darkintellect: An Approach to Detect Cyber Threat Using Machine Learning Techniques on Open-Source Information

Prof. Dr. Rushali Deshmukh [#1]

Associate Professor

Department of Computer Engineering, Jayawant Shikshan Prasarak Mandal's Rajarshi Shahu College of Engineering, Pune, India

Sudarshan Shinde [#2], Badal Yadav [#3], Amit Pathak [#4], Ashiq Shetty [#5]

Student

Department of Computer Engineering, Jayawant Shikshan Prasarak Mandal's Rajarshi Shahu College of Engineering, Pune, India

Article Info
Page Number: 1431-1439
Publication Issue:
Vol. 71 No. 4 (2022)

Article History
Article Received: 25 March 2022
Revised: 30 April 2022
Accepted: 15 June 2022
Publication: 19 August 2022

Abstract

Advances in information technology have led to a significant increase in cybercrime, security challenges, intruders and hackers. Cyberspace has a wealth of data that cybersecurity experts may utilize to develop threat intelligence, which will eventually aid in the prevention of cyberattacks and the protection of a company's network infrastructure. In contrast to the traditional random method of attack, cyber-attacks are now planned and carried out in a sophisticated manner targeting a specific target group, which is safe for the vast majority of netizens who have a keen awareness of the vast resources in cyberspace. The volume of Cyber Security literature available disseminated using social networking websites, particularly Twitter, has surged in recent days. A deep analysis of this data can aid in the development of a cyber threat situational awareness framework. We need scalable and efficient technology that can identify and summarize the information needed for a particular large data stream.

To Identify text linked to cyber threats, this paper recommends leveraging publicly available information from the Darknet platforms and Surface Web. With around 87 percent accuracy, our methodology can give law enforcement authorities and information security analyst with credible information that can be used to design control and prevention measures for cyber-attacks. We use machine learning techniques to assess the different sorts of online threats on social media in this research work. We discussed the algorithm based on the f-measurement value compared to accuracy and precision score.

Keywords: Cyber threats, Cyberspace, Dark Web, Dark Net, cyber-attacks.

1. Introduction

Humans have created a very complicated system called cyberspace which people use a lot on a daily basis and still lots of them have very little knowledge of it. We cannot run away from the fact of having security experts to conduct thorough analysis of specific sorts of attacks, such as identifying abnormalities in web traffic, viruses, and packet internet groper, among other things. On the other hand, studying social media data can help us observe new patterns of threat in cyberspace including Ddos, vulnerability and ransomware. Here we make an attempt to use a machine for predicting the threats in cyberspace in the proposed system. We collected Tweets with cyber threats and trained the dataset using machine learning methodologies which includes Support Vector Machine (SVM), Decision tree, Random Forest, XGBoost and AdaBoost algorithm to compare the accuracy and find the one with highest accuracy.

2. Related Work

Studies analyzing cyber threats for various purposes abound in the literature. G. Wang [1] presented a framework that extracts cyber risk and security data from Twitter data with a goal to identify the three types of threats and incidents. Their approach appeared to be effective as it was used extensively in both grammatical and contextual analysis and with dependency tree graphs. However, their architecture cannot be enabled to expand into more categories of constant threats and incidents.

[2] presents a paper framework to fetch and filter tweets using terms related to cybersecurity. Although the data filtration was useful for security analysts, there was no guarantee that the post completely had security-related content. Similarly, [3] proposed an automated method that uses the TF-IDF method to classify fresh unseen tweets either relevant or irrelevant after learning the features of cyber threat information from the categories of threats in the Common Vulnerabilities and Exposures database. For categorizing Cyber threat-relevant tweets, their classifier scored an F1-score of 0.643, which outperformed SVM, Multilayer Perceptron (MLP), and Convolutional Neural Network (CNN) baselines. Also, Rodriguez [2] Wang's [3] research analyzed emotions on hacker forums to predict cyber threats that were useful to cyber security analysts and users for preparing and planning for cyberattacks.

Another method proposed by [4], uses a combination of two CNN models with equivalent architectures, the first of which is used to evaluate relevancy of every tweet with respect to cybersecurity and the latter one was used to classify the relevant tweet in 8 different types. An average F1-score of 0.82 was achieved by the model. Main focus was on multi-source intelligence fusion methods and analytics to detect events and track cyber-threats, also used social network analysis to prioritize threat indicators, and monitor active threats.

There are many previous efforts that categorize tweets as cyber security relevant or irrelevant that use different methods.[12],[5],[8], [11] used keywords related to cyber threat to filter out useful tweets from the normal. Similarly [6] came up with a approach where different characteristics such as similar meaning words, locality, sexuality, age, tags, and sarcasm were considered for tweets relevancy. [6] discussed that by using this Random Forest Algorithm, we can Detect Cyber threats automatically with an efficiency of more than 80% and an

accuracy value of 0.80. Here there is a scope to improve accuracy so we can go with more algorithms.

With the same keyword-based tweet classification approach, [5] presented a 3-Dimensional framework that claims that Random Forest Classifier is better than Logistic Regression and Gradient Boost in terms of precision rate while a random forest classifier with a precision of 0.82 and a recall of 0.82 has the highest overall performance over an F1-score of 0.81. Data extracted with the help of a open source crawler, cybersecurity threat forum and python library TWeepy [10]. then used stop word remover and bag-of-words to get rid of noise (special characters, title description.) and correction of misspelled words respectively.

Unlike the prior discussed methods, [9] uses a Natural Language Processing (NLP) method doc2vec to distinguish between critical posts and noncritical posts on the dark web retrieved using Sixgill tool. A Stratified type of k-fold cross validation made the fraction of every label in the training data and evaluation data equal which resulted in less training time of MLP model with an accuracy of 90% which then dropped to 79.4% on unseen dataset.

3. Proposed Methodology

The Proposed system employs a three-dimensional approach to delivering information that may be utilized to warn cybersecurity specialists of prospective hazards as well as data that can be used to avoid cyber assaults before they occur.

A. Data set

While doing the literature survey, we came across a dataset of 21000 tweets formed by [4] with the help of a python package Tweepy. And another dataset from the Darknet forums extracted from CIC database [13]. This database included thread ID, topic title, name, postdate, and post text for postings made on Dark Net Market Platforms.

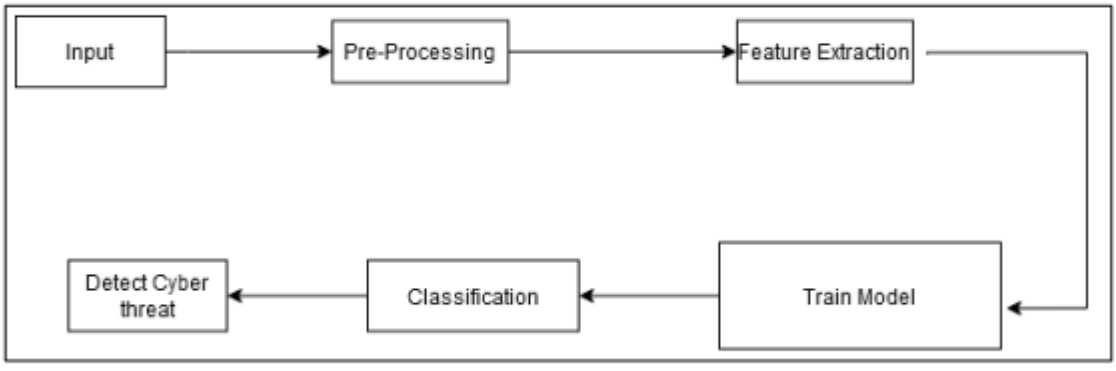


Fig. 1: Data Flow diagram

B. Data Preprocessing & Feature Extraction

The retrieved data from deep web forums often includes titles, descriptions, and special characters (%,!, *, &) that are eliminated using a stop word remover an NLP toolkit and serve

as noise to the classifier if not removed. Because forum postings are text data, they require NLP before being utilized as input for machine learning. Because forum entries are text data, NLP is required before they can be used as input for machine learning. As a first stage in machine learning, it is also important to extract feature values suitable for classification. In the proposed system Using Keras Term Frequency - Inverse Document Frequency (TFIDF) approach, concatenated texts from darknet forums and tweets were converted into a word embedding matrix with the help of TfidfVectorizer. We adopt the TF-IDF [14] approach to represent each text as a vector, which applies weights to the text words as follows. Let *t* represent a text in a database and *w* represent a word in the text. The weight of term *w* in text *t* is defined as

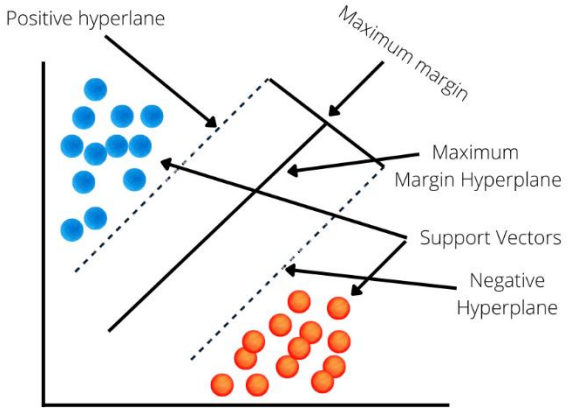
$$T\,F\text{–}IDF(w, t) = f(w, t) \times log(\frac{N}{n_w})$$

where *f(w, t)* is the number of the occurrences of term *w* in text *t*, *N* is the total number of the texts in the corpus and n_w is the number of the texts containing the term *w*.

Prior processing in NLP is crucial for correctly vectorizing a document and acquiring feature values. As preprocessing stages in the suggested methodology, we performed cleaning and stop-words processing. Cleaning refers to the removal of extraneous text characters such as numerals and parentheses.

C. Model Variation

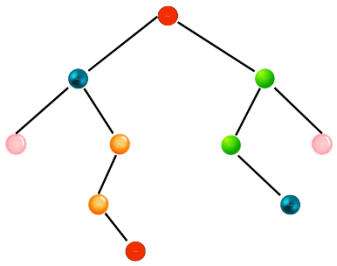
1) Support Vector Machine: The SVM algorithm is a supervised machine learning technique that may be used for both regression and classification. Despite the challenges with regression, classification is the best fit. It is inspired from the notion of dealing with the dual form of large-dimensional problems such that the classifier only needs a few support vectors to achieve the structural risk minimization principle.



2) Random Forest Classifier: In the original version of RF, using a bootstrap pattern randomly drawn from the original dataset, each tree was created using the CART approach and the criteria for splitting was Decrease Gini Impurity (DGI). During the creation of each

tree at each split, just a small number of randomly selected characteristics are assessed as candidates for splitting. From previous studies [6], the accuracy was about 80%, which could be increased to around 85% by our proposed system.

3) Decision Tree: A decision tree, as the name implies, makes decisions based on tree topologies, which is a typical decision-making method. The Decision Tree is used to develop a training model that will generally anticipate the category or degree of flexibility desired by reading basic decision rules based on past data (training data). Predicting the record category label in Decision Trees begins with tree support.



4) XGBoost: It is an ensemble Machine Learning approach based on decision tree in which boosting of gradients is done. XGBoost is quite effective when it comes to leveraging computer resources and processing speeds.[16]

5) AdaBoost: The basic idea behind Adaboosting technique is that we build a model on the training dataset first, then create a second model to correct the errors in the first model. This technique is repeated until the mistakes are reduced and the dataset is accurately predicted. This process lasts as long as the errors are minimized, and therefore the dataset is accurately estimated.

D. Performance Evaluation

We divided our data into train and test sets to analyze the effectiveness of our system for classifying text blocks related to cyber threats. 70% of our data was used to train our classifiers and 30% for test. In order to accurately predict cyber threats based on our indicators, the ultimate result of our trained model must be evaluated. For an in-depth comparison of the algorithms, the F1 score factor can be considered. In training our prediction model, we thoughtfully choose the appropriate metrics. Positive Class Precision, Recall, and F1 score are used to evaluate the performance of our suggested model. The following are the metrics' definitions.

$$Precision\ value = \frac{TN}{TP+FN} \quad Recall\ value = \frac{TP}{TP+FN} \quad F1\text{-}Score = \frac{2TP}{2TP+FP+FN}$$

where,

TP: True Positive

TN: True Negative

FP: False Positive

FN: False Negative.

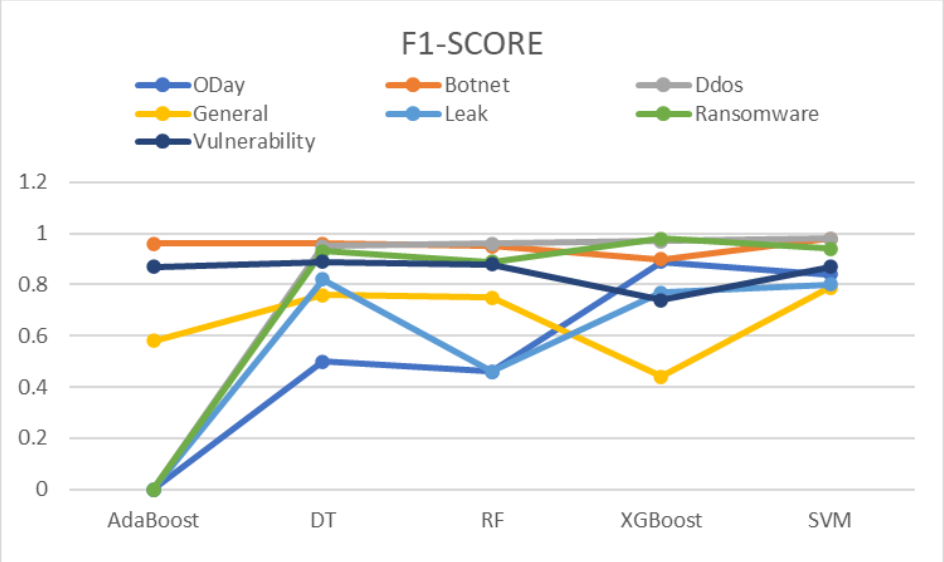


Fig 2. F1-scores for various algorithms based on attack type.

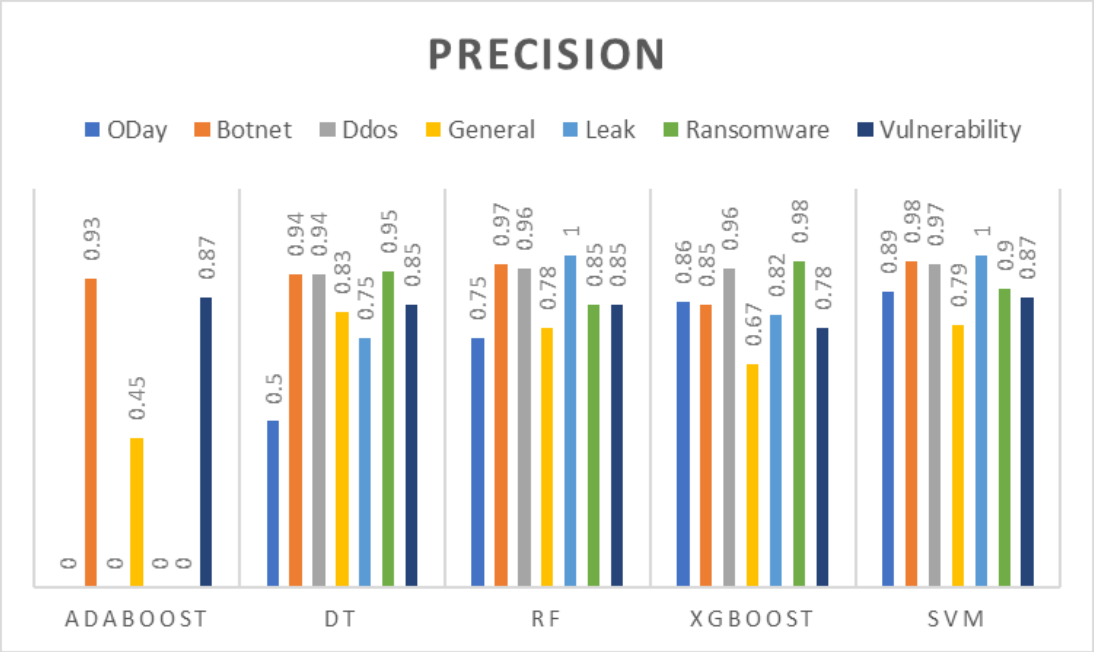


Fig 3. Precision of various algorithms based on the kind of attack

4. Discussion and Conclusion

Using different Machine learning algorithms, we were able to improve almost all accuracies and came to a conclusion that the Decision Tree Algorithm provides the best accuracy among all other algorithms, followed by SVM and XGBoost. SVM and Decision Tree outperformed all others, with a weighted average precision of roughly 0.87, which is higher than [15]'s 0.85. The AdaBoost algorithm, on the other hand, performs marginally worse, with an

accuracy of 65.87 percent. The Decision Tree with the Precision value of 0.87 and the Recall value of 0.87 has the best overall performance in terms of F1-score: 0.86.

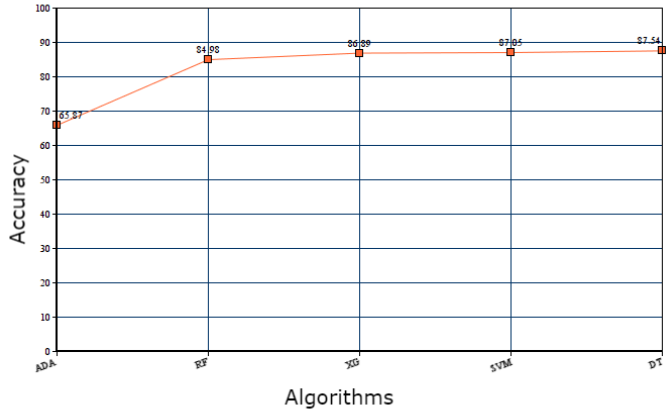


Fig 4. Accuracy Comparison Graph for algorithms used.

Comparison Table

#	Algorithm	Accuracy
1	RF	84.98
2	DT	87.54
3	XG	86.89
4	ADA	65.87
5	SVM	87.05

Fig 5. Accuracy Comparison Table for algorithms used.

In this study, we were able to provide a detailed overview of the various methods used to sort the relevance of tweets or posts related to cyber security. With around 85% of efficiency and 0.85 precision factors, Random Forest algorithm outperformed the results achieved by [5]. From fig 4, With 87.54 percent and 65.87 percent accuracy, respectively, Decision Tree was the best algorithm while AdaBoost was the worst. With a mean recall score of just under 0.67, the AdaBoost algorithm goal of threat type detection has a lot of potential for improvement.

5. Future Work

For future work, Use of a larger and more refined dataset from other social network services, Deep web, Surface web and other open-source intelligence sources to provide better learning to the system in order to increase accuracy. Using the RNN algorithm to check accuracy with

an expectation to get better results. With more parameter tuning Accuracy of AdaBoost may be increased by a considerable amount.

References

[1] R. P. Khandpur, T. Ji, S. Jan, G. Wang, C.-T. Lu, and N. Ramakrishnan, "Crowdsourcing cybersecurity: Cyber attack detection using social media," in Proceedings of the 2017 ACM on Conference on Information and Knowledge Management. ACM, 2017, pp. 1049– 1057.

[2] Rodriguez, A., & Okamura, K. (2020). Social Media Data Mining for Proactive Cyber Defense. Journal of Information Processing, 28, 230- 238.

[3] Le, B. D., Wang, G., Nasim, M., & Babar, M. A. (2019, October). Gathering cyber threat intelligence from Twitter using novelty classification. In 2019 International Conference on Cyberworlds (CW) (pp. 316-323). IEEE.

[4] Behzadan, V., Aguirre, C., Bose, A., & Hsu, W. (2018, December). Corpus and deep learning classifier for collection of cyber threat indicators in twitter stream. In 2018 IEEE International Conference on Big Data (Big Data) (pp. 5002-5007). IEEE.

[5] V. Adewopo, B. Gonen and F. Adewopo, "Exploring Open Source Information for Cyber Threat Intelligence," 2020 IEEE International Conference on Big Data (Big Data), 2020, pp. 2232-2241, doi: 10.1109/BigData50022.2020.9378220.

[6] Arora, Twinkle & Sharma, Monika & Khatri, Sunil Kumar. (2019). Detection of Cyber Crime on Social Media using Random Forest Algorithm. 47-51. 10.1109/PEEIC47157.2019.8976474.

[7] Alves, F., Bettini, A., Ferreira, P., Bessani, A.,(2020, July) Processing tweets for cybersecurity threat awareness

[8] Ranade, P., Mittal, S., Joshi, A., Joshi, K. (2018, November). Using deep neural networks to translate multi-lingual threat intelligence. In 2018 IEEE International Conference on Intelligence and Security Informatics (ISI) (pp. 238-243). IEEE.

[9] Kadoguchi, M., Hayashi, S., Hashimoto, M., & Otsuka, A. (2019, July). Exploring the Dark Web for Cyber Threat Intelligence using Machine Learning. In 2019 IEEE International Conference on Intelligence and Security Informatics (ISI) (pp. 200-202). IEEE.

[10] Adewale, A.V., & Gonen, B. (2020). Scraping The Deep Web: A 3-Dimensional Framework For Cyber-Threat Intelligence.

[11] Mittal, Sudip & Das, Prajit & Mulwad, Varish & Joshi, Anupam & Finin, Tim. (2016). CyberTwitter: Using Twitter to generate alerts for Cybersecurity Threats and Vulnerabilities. 10.1109/ASONAM.2016.7752338.

[12] Robertson, J. & Diab, Ahmad & Marin, Ericsson & Nunes, Eric & Paliath, Vivin & Shakarian, Jana & Shakarian, Paulo. (2017). Darkweb cyber threat intelligence mining. 10.1017/9781316888513.

[13] P. -Y. Du et al., "Identifying, Collecting, and Presenting Hacker Community Data: Forums, IRC, Carding Shops, and DNMs," 2018 IEEE International Conference on Intelligence and Security Informatics (ISI), 2018, pp. 70-75

[14] Cisomang, et al. “Phone Numbers of Millions of Facebook Users Exposed Online.” CISO MAG — Cyber Security Magazine, 9 Sept. 2019, www.cisomag.com/unprotected-database-exposes-millionsof-facebook-users-contact-numbers

[15] Ketha, Simran & Balakrishna, Prathiksha & Ravi, Vinayakumar & Kp, Soman. (2020). Deep Learning Approach for Enhanced Cyber Threat Indicators in Twitter Stream. 10.1007/978-981-15-4825-3_11.

[16] XGBoost . “XGBoost Documentation — xgboost 0.81 documentation” https://xgboost.readthedocs.io/en/latest/