Journal of Advanced Zoology
ISSN: 0253-7214
Volume 44 Issue S7 Year 2023 Page 1463:1472

# The Computational System To Classify Cyber Crime Offenses With Twitter Dataset Using Effdt Classification

**M. Nisha[1]*, Dr. J. Jebathangam[2]**

[1]*Department of computer science, Vels Institute of Science Technology and Advanced Studies Chennai, India. Email: Manikantnisha23@gmail.com
[2]Department of computer science, Vels Institute of Science Technology and Advanced Studies Chennai, India Email: jthangam.scs@velsunive.ac.in

***Corresponding author:** M. Nisha
*Department of computer science, Vels Institute of Science Technology and Advanced Studies Chennai, India Email: Manikantnisha23@gmail.com

***Article History***

*Received:*
*Revised:*
*Accepted*

***Abstract***

The Rapid growth of the Internet in the current decade enables the users to access the internet for day-to-day activities. People access the internet for many purposes: entertainment, Transactions, educational purposes and business. On the other hand cyber-crime has increased equally in terms of handling the massive data in the cloud using the access failures. Cyber-crimes are eventually increasing and reducing cyber-attacks for the data stored in the cloud. Existing framework and approaches fail to control the cybercrime attacks and thus many officers are increased because of the predictive control failure. The present study is focused on developing an effective computational method using a machine learning algorithm to analyze the cybercrime rate and to classify the cybercrimes. The system utilized Natural Language Processing (NLP) is used to process the text data. The particle Swarm Optimization algorithm is used to extract the features from the text stop. The main process involved here is the end sampled feed forward decision tree algorithm used to classify the text where any cyber assault are injected into the text. The main operation is to remove the read and features in the text and classifies the existing test text data Using SVM classifier and K nearest neighbor classifier in order to obtain the efficient classifier.



***Keywords— Natural language processor, cyber-attacks, particle Swarm Optimization, mission learning, Cybercrimes.***

## I. Introduction

The current trending Network Technology, cyber-attacks can happen at any instant that seriously affect the regular activity of the people who frequently utilize social media. According to the world economic forum in 2019 the first incident of cyber-attack has been detected. Cyber-attacks cannot be treated immediately. Cyber-attacks systems are systematic methodology that enter into the system and take the control of the system, then perform resilient attacks on the system causing the system to fail at certain tasks. Frequency of Cyber-attack occurs and increases rapidly in the current scenario. A timely detection of Cyber threats are important in order to safeguard the system from losing the data and credential information without the knowledge of the user.

Cyber-attacks in recent days enter wire social media websites such as Facebook, Twitter etc [3]. People often spend more time on Twitter and Facebook, other micro blogging websites, public events, personal use and business purposes. People start posting information such as advertisements, personal, command and public opinion more often on social media websites. This website is helpful for the hackers to learn about the user interest, pattern of usage of the system, social media, common interest and even more that directly impact the information laws about the user [4].

Sometimes these kinds of micro blogging websites have a greater impact on the public and make the public pay attention towards the individual. People often post the interest in public websites graph the hackers of same sin are you and able to contact those people with similar interest [5].

Cyber security attacks can come from the skin of links and third party websites that are connected by the user during the long term usage of social media websites. People who often click the particular link and spend more time on the particular website without the knowledge of a user, a backend system can be generated and that application can grab the information from the system without the knowledge of the user. Peoples are attracted through the advertisements happening in the social media websites to stop these websites have back and vulnerable attacks in our used hidden in the link to stop people click the link without the prior knowledge of the website and able to which the user information such as email id password and give the confirmation of the email ID for downloading certain type of documents etc. The increasing amount of complexity in detecting cyber security attacks in recent days are developed using machine learning algorithms. Data mining techniques helpful in understanding the pattern of data involved in cyber-attacks. Attacks malwares, spywares create much impact on the system data. Many experts are developers in recent days to capture those patterns of cyber-attacks and malware attacks. Research mechanism, focused on detecting cyber security problems through robust methodology. Text analysis and data mining techniques are widely used in cyber security detection systems such as Malware attack detection, DDoS attack detection, Cybercrime problems are increasing day by day on social media. One of the most challenging task in internet is to take the of the world websites. If the websites are able to detect the vulnerable entry of any malicious websites while using the social media sites then those web sites are more intelligent than manual interference. Analysis of websites from cyber-attack entries through a robust systematic methodology is time the study area of research.

Detection of various behavioral patterns of the individuals and groups utilizing the social media websites are helpful to detect the cyber-attacks, cyber probes globally. A proposed Framework is used to detect cyber security attacks happening in the publicly available networks using data mining techniques. In social media networks are allowed information to host to publicly many cyber criminals are involving in the analysis of user information. The process of that system is focused on creating a robust algorithm that continuously checks malicious website entry in social media networks. Suspicious words in the form of links, detected in the social media text can be analyzed by the proposed system.
The percent of paper is formulated as detailed background study in the Section II. Further the detailed methodology, system architecture is described in section IV. The paper is concluded with the presented results and discussion, the Conclusion and future work are discussed and further references are described.

## II. RELATED WORK

***Deepak Soni et al.*** The semantic analysis techniques utilized in the present system detects the cybercrime patterns that recently occurred in the system. Logistic regression algorithm is utilized by the author to assess the popularity of public certain political parties and their discussions were the social media network [7].

***Rashid Kamal et al.*** the author a model to perform sentimental analysis from the live tweets collected from the Twitter website. The other classified the tweeter sentiments using Hadoop* software. The classification of tweets differentiate the sentiment polarity is positive, negative, neutral etc.[8]

***M. Nimbarte et al.*** [9] The Other developed classification algorithm in order to detect negative words present in the database. Classification is utilized compared with the pretrained negative words stored in the database. the similar approach is used to detect the negative word impact in the particular website.

***Nisha Tanwani et al.*** [10] The author designed the framework collect data from the Facebook website related to academics, some of the student groups and classify the sentiment polarity mining in terms of text written by

the students and other people's in the website. The present system attitude 86% positive words, 32% neutral words, 12% negative commands provided with the syllabus, learning environment and teaching feedbacks.

***M. Hood et al***.[11] The order presented a system in which the relationship between the bulimic X availability on the social media websites, provided with recommendation against cyber-crime are detected. Gender specific languages are detected from the given text also the present it is too much of 63% accuracy on analyzing the personal information that impact the other person. Extract text extraction Framework is developed here to analyze the positive and negative emotions impacted in the written text.

***Vinita S. et al.,*** the author content based feature extraction technique to detect the building detection in the social media websites. represented system analyze the background of the particular social media user on how many friends they are connect, the dubbed, comments multimedia upload along with the content that does not create any impact on the social security.

***Le Sceller et al.*** [13] the author a automatic self-learning framework to detect the cyber security threats present in the keywords are see data of stop the present an algorithm to detect known trips that are trained by the cyber security detection systems further it analyze the cyber events that are able to enter into the properly working websites.

***Bose et al.*** [14] the author machine learning algorithms approach on cyber detection in Twitter. The approach can extract the Tweeter terms and extract the real meaning of the Twitter terms in order to classify the keyword as positive negative and neutral. This keywords are analyzed in depth in order to find out the Entity are owned the particular user and how the cyber impact disturb the user performance on ranking.

***Ji et al.*** [15] the author in supervised learning methodology to make multi-tasking learning based model for detecting cyber truck event. the presented approach detect the complex problems arise in publicly available networks.

***Le Sceller et al.*** [16] the author presented automatic keyword centric self-analysis network for detecting jio location, through Twitter streams. the author describe the Framework in three different types of network. first method is similar to the present system, can streaming of waiter words using IOC extraction.

***Liao et al.*** [17] the author presented automated detection of IoT extraction technique using Twitter network service. The advantage is compared with the percentage system and many state of art approaches on Peter extraction an analysis to stop

***Sabottke et al.*** [18] the author proposed a twitter based exploratory data analysis system using support vector machine classifiers. The present detector is used to extract vulnerable attacks related to the information presented in the threats. An interesting feature of the presented system is to detect the abnormal pattern using an adverse network. Malicious devices are capable of detecting the real world scenario and for the purpose of making separate text based database services for vulnerable attack detection. The social media platform uses large amounts of diversifying users, the publicly available website without any time constraints. More data are stored in the cloud in daily basis.

***Machine learning techniques***

Machine learning technique is typically used as a branch of artificial intelligence, closely related to statistical computations, data mining and analysis and particularly focused on Computers to learn about the data patterns. The measuring models typically comprise 2 sets of rules and statistical methodologies to measure the uniqueness present in the pattern of the data. Interestingly, they are helpful to understand, by the behavior of the hackers. Machine learning algorithms play a major role in cyber security attack detection systems.

## III. METHODOLOGY

Proposed Framework considered as multiple websites for analysis. Some of the commonly available websites used by the public more often are Twitter, Facebook, YouTube, Wikipedia talk. Cyber security detection system detects various groups present in the publicly available websites of. Students group, children group, adults group are separated by the algorithm to detect the vulnerable act. The main intention of the proposed

system used to determine the vulnerable act such as bullying is present in the given websites. The block diagram shows the proposed methodology where the feature selection is done using particle Swarm Optimization technique. The analysis of text is developed using a natural language processor to stop the proposed algorithm named Feed forward decision tree (EFFDT) is developed here.

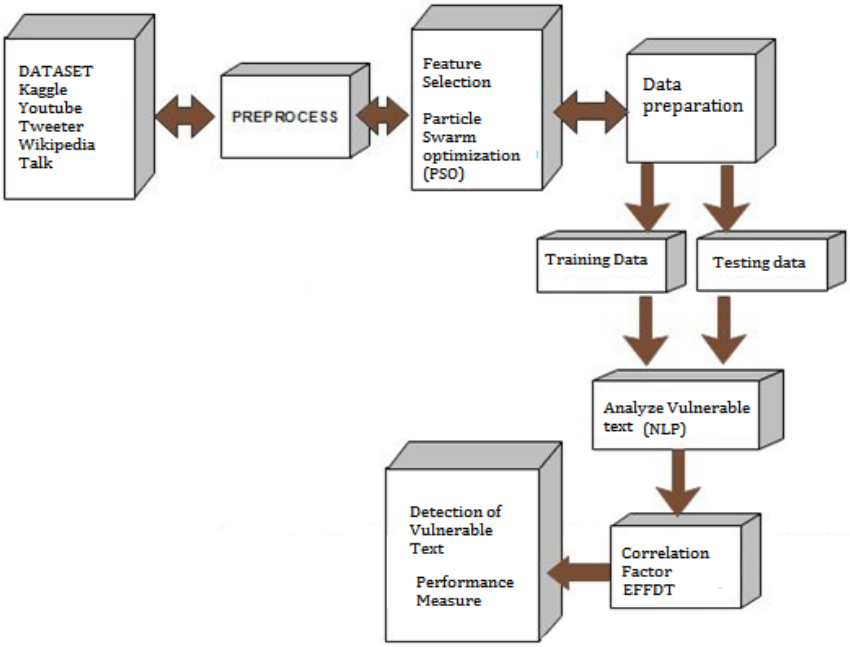


**Fig 1** Block diagram of proposed method of EFFDT
**Fig 1** shows the proposed method of EFFDT algorithm working process.

***Text mining process***
The necessary steps involved in text mining process are data retrieval, information grabbing, natural language analysis (NLP_based) are text processing, text transformation, feature selection, text mining methods, and evaluation frameworks are developed.

***Data preprocess***
The foremost step of analysis is the preprocessing technique used to clean the given raw data set. The preprocessing technique involves reading the raw data from the given file, removing the junk values present in a, Tokenization, stage segmentation, planning the unstructured text into words by removing the blank spaces from the given paragraph. The arrangement of words such as, 'a', 'is', 'of' etc is performed. Further the segmented text is applied to text transformation.

***Text Transform***
After selection of the data set, the data is transformed into a useful variable after the garbage text to removal process. The data is stored in a temporary variable for further processing. The processed data is ready to make further pattern analysis. These text data are compared with bags of words.

***Data encapsulation***
In the proposed process, the Data extraction techniques are developed using TFIDF technique. It will evaluate the bigrams, unigrams and trigrams related to Cyber-crimes. The Other form of data processing techniques for term frequency extraction is defined as the Count of a particular word occurrence frequency and finds out the importance of each word in the document through the term recurrent. Reverse document occurences is the down scales of the given word in the particular text are paragraph and how many times it is being repeated in the report. The word embedding process trained the different words of the given data set and compared with the collection of data set already stored in the network and form avoid space between the repeated data set. The values of each data are stored in the matrix format.

***Feature extraction techniques***

Feature extraction technique is the algorithm of gathering unique information as unique and important human visual value used for learning machine learning. Support system is focus on detecting various statistical features such as mean contrast median and Varience, standard deviation, skewness, kurtosis etc.

***Feature selection process***

Proposed model is based on the Particle Swarm Optimization (PSO) technique used to make the feature selection from the given text. PSO is a meta-heuristic global Optimization algorithm used for detecting the best intelligent text after analyzing various patterns from the given training data set. The particle Swarm Optimization algorithm is based on the biological inspiration of spam that separates the different food naturally. The behavior of the swarm is used here to segregate the data into different scattered patterns before considering for analysis. The swarm based intelligence technique used for global optimization. The basic algorithm consists of spam particles and the positions of each from particles that is nothing but the instant sample recorded from each node.

Further changes it position by using Three Types of techniques. It keeps its inertia before making any decision to stop it updates the condition of the present state before its optimal position is updated. It will update the condition with respect to most of the other optimal positions of spam are undergoing.

$$v_{id}^{k+1} = v_{id}^{k} + c_1 r_1^k \left(pbest_{id}^{k} - x_{id}^{k}\right) + c_2 r_2^k \left(gbest_{id}^{k} - x_{id}^{k}\right) \quad (1)$$

$$x_{id}^{k+1} = x_{id}^{k} + v_{id}^{k+1} \quad (2)$$

In these two equations, $v^k_{id}$ and $x^k_{id}$ denotes the speed of the particle of i and its k times and the d-dimension quantity of its position, $pbest^k_{id}$ represents the d-dimension quantity of the individual i at its most optimist position at its k times. $gbest^k_{id}$ is the optimist position of the d-dimension quantity of the swarm. Fast computation of C1 and C2 is used to regulate the moving particles in the vector space. The whole system is randomized and further achieves a particular class of pattern that is helpful to create a model.

**Classifications**

Classification process is used to detect the suspicious messages present in the given text data. The classification is mainly involved in two steps. Data learning is the step used to analyze the training data and to build a model before making the prediction. The testing data is the second set of work used to classify the pattern of test data provided to the ensemble feed forward decision tree algorithm for making the suspicious pattern of text. The proposed methodology is validated by calculating the accuracy of the presented model.

**EFFDT ALGORITHM**

The proposed EFFDT algorithm is multilayer feed forward neural network architecture consists of three layers. The input layer denotes to read the neural network input training and test data to stop the neural network Toolbox is configured up to 10 hidden layers. the transfer function used for making the analysis need to be selected before the input is fetched into the neural network toolbox of a store Based on the complexity of the input data set the hidden layers of can be altered.

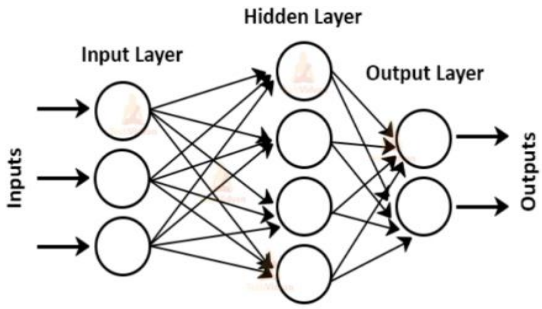


**Fig 2.** General structure of Feed forward networks.

The proposed architecture consists of three layer neural network model with

- 15 nodes arranged in the first layer.
- 1 to 25 nodes formulated as hidden layer.

- 1-20 nodes present in the adjacent proceeding layer
- The nodes present in output layer connect the declared nodes.
- The weights of the assessment node is updated at every iterations.

Input data after the feature extraction change state, output nodes provider 0 are other 1 respectively. In case of malicious pattern present in a network the presented neural network architecture signed Out data one. in case of no malicious activity present in the network then the data zero is sent out to stop The percentage feed forward neural network architecture utilizes sigmoid as a transfer function to stop At every iteration of the training pattern the weights are calculated and updated. The output layer of the end simple feed forward decision tree algorithm consists of two layers. The output of the network is used to fetch the presence of malicious attack are not.

**Input layer**
The purpose of input layer is to form a connectivity between 15 input nodes, and unique features setting at hidden nodes arranged in the collected dataset.

**Functional Hidden layers**
The goal of hidden layer is to make a connection between the input layer and classification layer. The layer act as functional connectivity between input layer and the expected classified output layer.
The proposed system based on two hidden layers

- The first one with 20 numbers of neural nodes.
- The second contains 25 numbers of neural nodes.
- These nodes are arranged with transfer functions such as sigmoid, Relu for training.

**Output layer**
It produces the result (normal are suspected messages). All nodes in the input layer have full connect with all nodes in the next hidden layer, and so on in all the rest layers.

Calculate the fetched inputs and outputs of the k formulated as output layer neuron

$$net_j^h = \sum_{i=1}^{N+1} W_{ji} x_i \quad (3)$$

The net inputs and outputs of the j hidden layer neurons can be calculated as follows

$$y_j = f(net_j^h) \quad (4)$$

$$net_k^o = \sum_{j=1}^{J+1} V_{kj} y_j \quad (5)$$

$$Z_k = f(net_k^o) \quad (6)$$

At every iterations, update the bias weights in the output layer (for all k, j pairs)

$$v_{kj} \leftarrow v_{kj} + c\lambda (d_k - Z_k) Z_k (1 - Z_k) y_j \quad (7)$$

At every iterations Update the weights in the hidden layer

$$w_{ji} \leftarrow w_{ji} + c\lambda^2 y_j (1 - y_j) x_i \left( \sum_{k=1}^{k} (d_k - z_k) z_k (1 - z_k) v_{kj} \right) \quad (8)$$

Update the performance error

$$E \leftarrow E + \sum_{k=1}^{k} (d_k - z_k)^2 \quad (9)$$

Also, rehash from Step 1 until all info designs have been introduced (one age). In the event that E is beneath some predefined resistance level, stop. In any case, reset E = 0, and rehash from Step 1 for one more epoch. Now when the total information is arranged into two gatherings and to diminish the phony problem rate and for the recognition of the digital wrongdoing we pay attenuation on the datasets having a place with class1.We

train the model utilizing the grouping C4.5 calculation and the bootstrapping method. We utilize the bootstrapping procedure for the plan of the choice tree and along these lines result is feed into the C4.5 classifier. The precision got is 70% and 0.3025 goes under mistake rate.

***Data collection***

The data set is collected from multiple sources such as a publicly available network Twitter, Facebook, twitter, Wikipedia Talked.

## IV. RESULTS AND DISCUSSION

Results are validated using statistical measures such as accuracy, precision, recall, F1 score estimation with 80% of training data, 20% of testing data is collected. The classification algorithm detects the malicious patterns from the given text and formulates the performance measure. Confusion matrix is formulated using true positive value estimation, false positive rate, true negative value, false negative value. The values are calculated by the formula below.

### *a. Recall (TPR)*

The recall rate shows the positive values that occurred during the classification process.

$$TPR = \frac{TP\ frequency}{TP\ frequency + FN\ frequency} \quad (10)$$

### b. Precision

The precision determines the amount of achievable positive values obtained from the given pattern.

$$\boldsymbol{Precision} = \frac{TP\ frequency}{TP\ frequency + FP\ frequency} \quad (11)$$

### c. Accuracy

Accuracy is determined by the set of positive values with respect to all possible values.

$$\boldsymbol{Accuracy} = \frac{TP\ frequency + TN\ frequency}{TP\ frequency + FN\ frequency + TN\ frequency + FP\ frequency} \quad (12)$$

### d. F- measure

F1score is measured as the test accuracy and defined as the weighted average of precision and recall.

Table 1 Performance outcomes of proposed method and existing method

| Algorithms | Accuracy | Precision | Recall | F1Score | Error rate |
|---|---|---|---|---|---|
| Knn | 61.28 | 69.84 | 76.17 | 72.87 | 0.3872 |
| SVM | 68.26 | 68.26 | 100 | 81.14 | 0.3187 |
| EFFDT | 70 | 70.98 | 94.02 | 81.96 | 0.3025 |

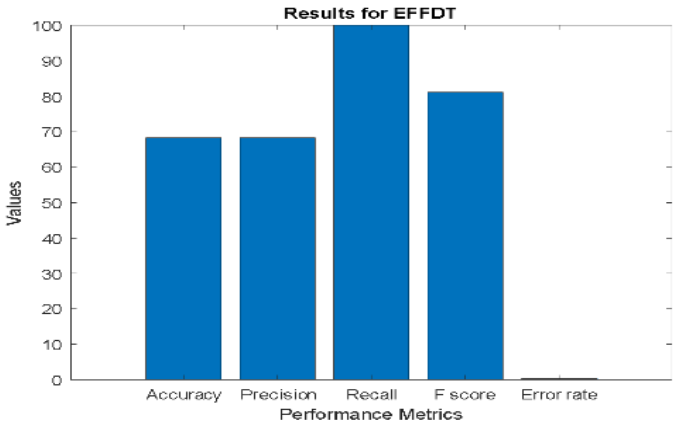


**Fig 3** performance metrics using EFFDT

Fig.3 shows the Performance metrics of EFFDT model.


**Fig 4** performance metrics using SVM
Fig. 4. Shows the performance results obtained using of SVM algorithms.

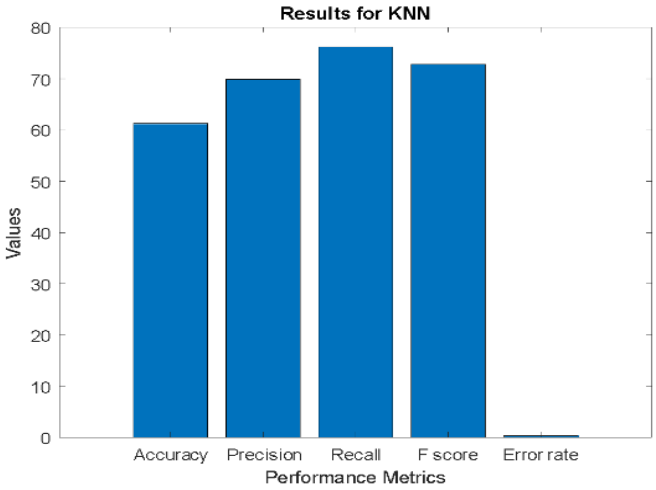


**Fig 5** performance results using KNN
Fig. 5. Shows the Performance results of KNN.

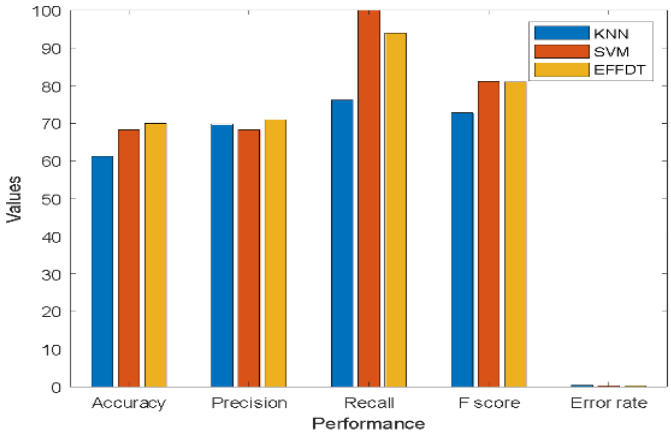


**Fig 6** performance Comparison of existing and proposed methods.
Fig 6 shows the performance Comparison of existing and proposed methods. The existing method such as KNN SVM is compared with the EFFDT model.

The outcomes as reflected in Fig 6 shows that among the three classifiers, EFFDT got the most elevated forecast exactness with 70% followed by SVM and KNN with 68.26%, 61.28% individually. Taking a gander at the blunder rates displayed in Table 1, the proposed classifier records the least mistakes which could decipher that the two classifiers have practically a similar normal expectation blunder.

## IV. CONCLUSION

The emerging growth of internet of things (IoT) and frequent use of social media activity by the users, cyber-attacks, bullying and Cyber threats are increasing. The presented approach utilized ensemble feed forward decision tree algorithms for capturing the vulnerable attack patterns from the website links. The present approach compares with existing methodologies such as SVM, KNN algorithm on detection of vulnerable patterns and formulates the efficient prediction scheme using EFFDT algorithm.

## REFERENCES


1. World Economic Forum. The Global Risks Report 2019. Available online: http://www3.weforum.org/docs/WEF_Global_Risks_Report_2019.pdf (accessed on 25 August 2020).
2. Satyapanich, T.; Ferraro, F.; Finin, T. CASIE: Extracting Cybersecurity Event Information from Text. Umbc Fac. Collect. 2020, 34, 8749–8757. [CrossRef]
3. Noor, U.; Anwar, Z.; Amjad, T.; Choo, K.K.R. A machine learning-based FinTech cyber threat attribution framework using high-level indicators of compromise. Future Gener. Comput. Syst. 2019, 96, 227–242. [CrossRef]
4. Yagcioglu, S.; Seyfioglu, M.S.; Citamak, B.; Bardak, B.; Guldamlasioglu, S.; Yuksel, A.; Tatli, E.I. Detecting Cybersecurity Events from Noisy Short Text. arXiv 2019, arXiv:1904.05054.
5. Mazoyer, B.; Cagé, J.; Hervé, N.; Hudelot, C. A French Corpus for Event Detection on Twitter. In Proceedings of the 12th Language Resources and Evaluation Conference, Marseille, France, 11–16 May 2020; pp. 6220–6227.
6. Da Costa Abreu, M.; Araujo De Souza, G. Automatic offensive language detection from Twitter data using machine learning and feature selection of metadata. In Proceedings of the IEEE World Congress on Computational Intelligence (IEEE WCCI), Glasgow, UK, 19–24 July 2020.
7. D. Soni, M. Sharma, and S. K. Khatri, "Political opinion mining using E-social network data," in Infocom Technologies and Unmanned Systems (Trends and Future Directions)(ICTUS), 2017 International Conference on, 2017, pp. 163-165.
8. R. Kamal, M. A. Shah, A. Hanif, and J. Ahmad, "Real-time opinion mining of Twitter data using spring XD and Hadoop," in Automation and Computing (ICAC), 2017 23rd International Conference on, 2017, pp. 1-4.
9. M. Nimbarte and M. O. Thakare, "User tracking using tweet segmentation and word," in Electronics, Communication and Aerospace Technology (ICECA), 2017 International conference of, 2017, pp. 664-668.
10. N. Tanwani, S. Kumar, A. H. Jalbani, S. Soomro, M. I. Channa, and Z. Nizamani, "Student opinion mining regarding educational system using facebook group," in Electrical Engineering and Computing Technologies (INTELLECT), 2017 First International Conference on Latest trends in, 2017, pp. 1-5.
11. M. Hood and A. L. Duffy, "Understanding the relationship between cyber-victimisation and cyber-bullying on Social Network Sites: The role of moderating factors," Personality and Individual Differences, 2017.
12. R. Pawar, Y. Agrawal, A. Joshi, R. Gorrepati, and R. R. Raje, "Cyberbullying Detection System with Multiple Server Configurations," in 2018 IEEE International Conference on Electro/Information Technology (EIT), 2018, pp. 0090-0095.
13. Le Sceller, Q.; Karbab, E.B.; Debbabi, M.; Iqbal, F. Sonar: Automatic detection of cyber security events over the twitter stream. In Proceedings of the 12th International Conference on Availability, Reliability and Security, Reggio Calabria, Italy, August 2017; pp. 1–11.
14. Bose, A.; Behzadan, V.; Aguirre, C.; Hsu, W.H. A novel approach for detection and ranking of trendy and emerging cyber threat events in twitter streams. In Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), Vancouver, BC, Canada, 27–30 August 2019; pp. 871–878.
15. Ji, T.; Zhang, X.; Self, N.; Fu, K.; Lu, C.T.; Ramakrishnan, N. Feature driven learning framework for cybersecurity event detection. In Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, Vancouver, BC, Canada, 27–30 August 2019; pp. 196–203.
16. Q. Le Sceller, E. B. Karbab, M. Debbabi, and F. Iqbal, "SONAR: Automatic Detection of Cyber Security Events over the Twitter Stream," in Proc. of the 12th International Conference on Availability, Reliability and Security (ARES). Association for Computing Machinery, 2017.

17. X. Liao, K. Yuan, X. Wang, Z. Li, L. Xing, and R. Beyah, “Acing the IOC Game: Toward Automatic Discovery and Analysis of Open-Source Cyber Threat Intelligence,” in Proc. of the 2016 ACM SIGSAC Conference on Computer and Communications Security (CCS). Association for Computing Machinery, 2016.
18. C. Sabottke, O. Suciu, and T. Dumitras, “Vulnerability Disclosure in the Age of Social Media: Exploiting Twitter for Predicting Real-World Exploits,” in Proc. of the 24th USENIX Security Symposium (USENIX Security 15). USENIX Association, 2015.
19. A. Attarwala, S. Dimitrov, and A. Obeidi, “How efficient is Twitter: Predicting 2012 U.S. presidential elections using Support Vector Machine via Twitter and comparing against Iowa Electronic Markets,” in Intelligent Systems Conference, 2017.
20. Yurnalita, “Cyberbullying on Social Networking Twitter (Trending Topic Semiotics Analysis),” Faculty of social and political sciences, University of Syiah Kuala, Banda Aceh Darussalam, 2016.