E-ISSN: 2707-6644
P-ISSN: 2707-6636
IJCPDM 2020; 1(2): 01-06
Received: 02-05-2020
Accepted: 04-06-2020

Katru Roshini
Dept. of Computer Science, Sri Venkateswara University, Tirupati, Andhra Pradesh, India

Corresponding Author:
Katru Roshini
Dept. of Computer Science, Sri Venkateswara University, Tirupati, Andhra Pradesh, India

Analysis of evasion attack defense methods in text classified training dataset

Katru Roshini

DOI: https://doi.org/10.33545/27076636.2020.v1.i2a.9

Abstract
Classification algorithms built different kind of feature representations based on training datasets. The major threat on training datasets are, they affected by various attacks. The unstructured training datasets are faced the challenges when they convert into structured datasets. The tiny text perturbation in the original training dataset will cause misclassification and incorrect predictions in machine learning. The different classification algorithms measurements help to detect the evasion attack on training dataset. To compare different defense methods helps the way of mitigating training dataset attacks. The experimental results prove that the text classifier training dataset secured from the evasion attack.

Keywords: evasion attack defense, training dataset

Introduction
The training datasets faced the challenging problems due to manipulate data and the modified datasets widely used in learning algorithms. The training datasets becomes susceptible to different kind of attacks before they feed as input to the ML classifier algorithms. The intelligent and technological adversaries download the training dataset from the trusted data sources and act their performance on that dataset for making dataset manipulation [1]. Adversary make a small change in real training dataset will cause loss something like classification performance, prediction or accuracy [2, 3]. The adversaries training datasets manipulations will change ML algorithm's results time to time. Let the classification function F for the training dataset x, the adversarial perturbation ▲x on the dataset, the resulting training dataset x′ as a new attacked dataset [4].

$$x' = x + \blacktriangle x$$

$$F(x) \neq F(x')$$

The text training dataset classification measurement calculated as before and after manipulation of training dataset variation is ▲x= x′ - x. The small modifications of text character change the meaning of the original text data. The manipulated text data train in Machine learning classifier which makes misclassification and produce wrong decisions.
To minimize data corruption of training dataset in a research area is to increase the robustness of classification results in machine learning models [5]. The attacker generated malicious datasets in the training datasets, the ML classifications results for the particular training datasets shows different result. Hence we can identify the researchers using affected training dataset for their research purpose. The defense methods prevent the training dataset from the attacker. This paper focus on compare different text classification training dataset defense mechanism and regeneration algorithm to create a new training dataset with labels from original text files.

Background
The evasion attacker change feature of the training dataset with the modification of limited or unlimited input malicious data [6]. The training datasets text file words are manipulated with misspelled words or malicious words by the attacker and they formed illegitimate text data to evade simple machine learning classifiers SVM, Naive Bays [7]. The training dataset file samples collected from any data source by the adversary and retrained with the malicious data.

The potential attack on training dataset is defined various categories are shown in figure1.

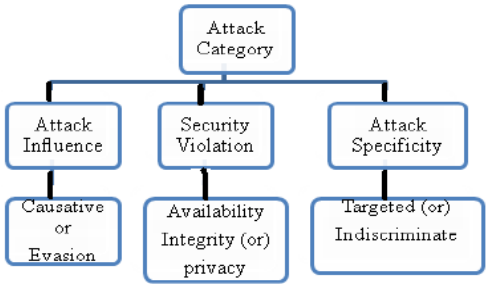


Fig 1: Training Dataset Attack Categories

The attack influence is a method of obtaining training dataset from producing data source and manipulate for own use. The security violation is to manipulate the training dataset for their expected result. The attack specificity is the target of attack on specific data in the dataset. These attack categories have different attack models like Causative, Evasion, Availability, etc.

a. Evasion malicious samples

An attacker has the technical knowledge of modifying the training datasets with malicious data which is correctly classified by ML algorithms. One of the techniques genetic programming [8] automatically generate such type of attack.

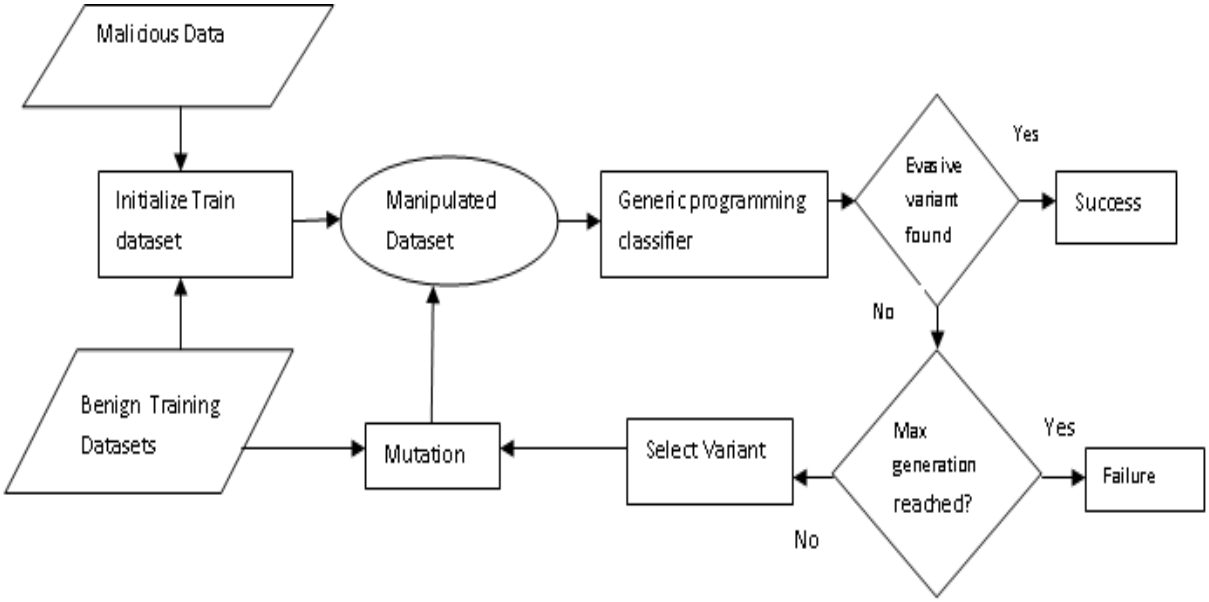


Fig 2: Generic programming evasion attack

The figure2 illustrate malicious data mixed with original training dataset and form initialize dataset. The dataset manipulated and pass through the generic programming for check the evasion attack is enough or not. If the evasion attack not found, a subset variant sent the training dataset for mutation. This process continues until the evasion attacked dataset reached.

b. PDF Malware

PDF malware attack means to steal information such as account number, algorithms and trade secrets [9] from the document files. Document format structures are support different Portable Document Format files. It has the basic format shown in figure3.

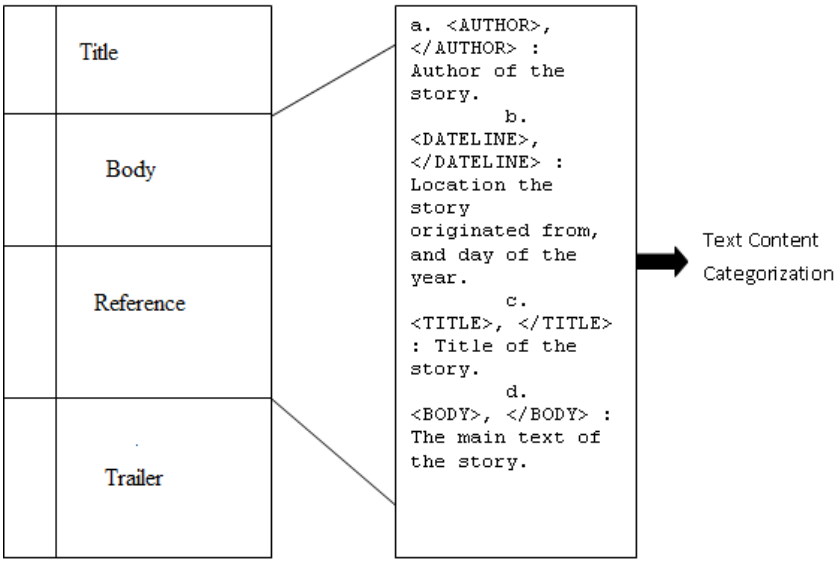


Fig 3: PDF file structure

The Reuters text data collection consists of the parts: Title, Body, Reference table and Trailer. The logical structural path PDF files make a uniform classification. Structural paths are merged and inherit contents reduced the tree structure to improve evasion robustness. The evasion attack PDF malwares are happen either insert or delete or replacement of text in the structured files [10]. Hot Flip method is the adversarial example [11] to inert or deletes sequence of characters in the text classified training dataset and confuse the dataset classification performance task. The PDF files are formed in logical structure which helps to extracting keywords. The keywords are modified in a limited way to form an evasion attack in spam filtering. The ML classification techniques recognize the PDF malware based on the variation of the different performance [19].

Problem Setting

To analyze evasion attack problem, we used the unsupervised training dataset Reuters which collected from UCI Repository. The Reuters data collection has 21 document files. The files contain number of category set Exchanges, Orgs, People, Places and Topics. The topics include the categories coconut, gold, inventories, etc. There are 69 number of categories repeated 13332 document files of training dataset. The category words are extracted and mean words calculated from Reuters training dataset document file.

a. Training Dataset Frame work

In the training dataset attacker directly manipulated original text classified dataset input to evasion attacked text related to produce misclassification [12]. Retraining with adversarial example in the ML algorithm helps to reduce adversarial frame work risk and build robustness of evasion attack model training datasets [13]. The frame works structure to reduce adversary evasion attack cost and support to change the training dataset structure as the adversarial need. Unstructured formatted text training datasets are challenge for the classification and adversarial frame work attack [14]. The researchers wants to perform classification, they should change the unstructured training dataset into structured format.

Class	No of Doc	Mean word
earn	3964	104.4
acq	2369	150.1
money-fx	717	219
grain	582	223.6
crude	578	247.3
trade	485	294.3
interest	478	198
wheat	283	225.6
ship	286	203.6
corn	237	259.1
money-su	174	170.5
dlr	175	203.4
sugar	162	247.2

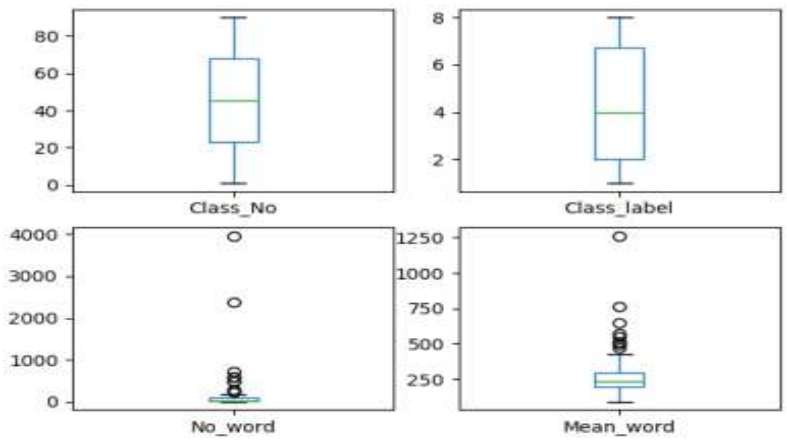
ocoa el-salvadorusauruguay C T f0704¬reute u f BC-BAI
out the week in the Bahia cocoa zone, alleviating th
superior+ certificates. In view of the lower quality
95 dlrs for March/April, 785 dlrs for May, 753 dlrs
rth America Inc said they plan to form a venture to
Inc's Texas Commerce Bank-Houston said it filed an
nder pressure to act quickly on its proposed equity
SEC approval is taking longer than expected and mar
industry through the quarter, the initial shock reac
rtment reported the farmer-owned reserve national fi
ures show crop registrations of grains, oilseeds and
prev 3.7, Feb 21.1, Mar nil, Apr 2.0, May 9.0, Jun 1
tnership said it filed a registration statement with
Inc said it lowered the debt and preferred stock ra
d its board of directors approved a two-for-one stoc
Computer Terminal Systems Inc said it has completed the sale of 200,000 share
e company said the moves were part of its reorganization plan and would help
Shr 34 cts vs 1.19 dlrs Net 807,000 vs 2,858,000 Assets 510.2 mln vs 479.7 ml

Fig 4: Frame work of Text Dataset

The above figure4 shows that the Reuters text formatted training dataset collected from UCI repository and convert structured csv dataset format.

b. Evasion Attack text Classification

The adversary selects the original text classified training dataset from any data source and they mixed misspelled words in the training dataset for misclassification [15]. The training dataset denoted T build to evasion attacked dataset T′. To select n sample malicious data to attack the dataset T cause misclassification and satisfy the adversary goal. The evasion attack of modified text classified training dataset categories will change the classification score of histogram is shown in Figure5.



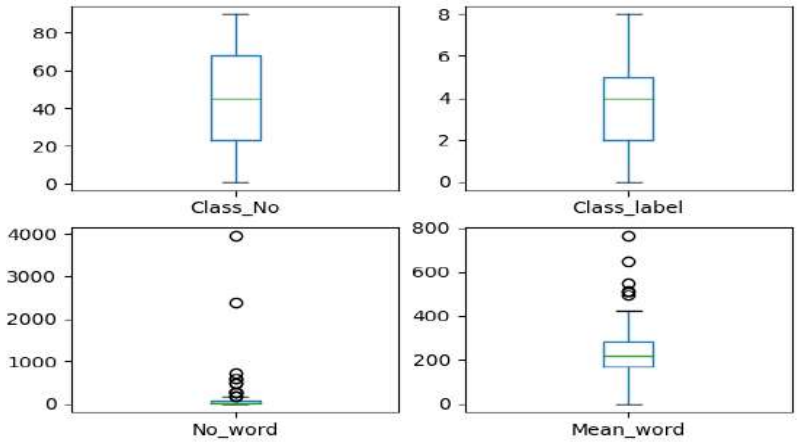


Fig 5: Histogram of the dataset before and after evasion attack

The above figure stated the category of different topic text data in the Reuters training dataset affected after evasion attack. The class label and mean word shows the different in the histogram.

Experiments

The experiments stated the different ML algorithm measurement and defensive method to prevent text classified training dataset from evasion attack.

a. Training dataset measurement

The Reuters Training dataset train in the various ML algorithms used for finds its accuracy; it helps to which algorithm is suitable for the dataset. After the evasion attack the accuracy measurement shows the different for decreasing measurement value.

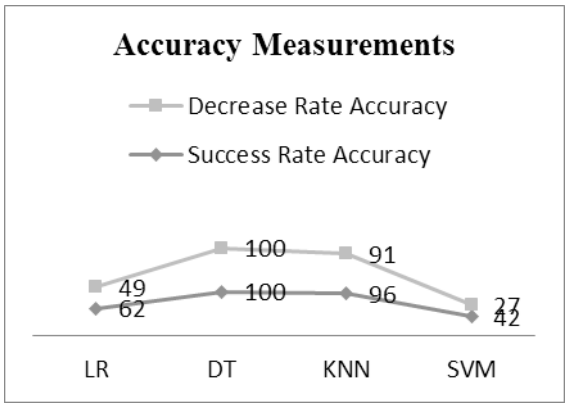


Fig 6: Accuracy Measurement Comparison

The k-nearest neighbor (KNN) suitable for the selected text training datasets and its accuracy measurement of 96%. The evasion attack decreased accuracy as 91%. So, we can easily detect the researcher choose the attacked dataset for evaluation. The defense models helpful to prevent training dataset from the evasion attack.

b. Defense Model

The defense methods concentrate on proactive arm race [18] models. The first step followed to identify the attack category. Second step to provide the security of the training dataset and protect machine learning classifier performance. The learning algorithms evaluate with different values of parameters in each class provides higher level security against text classification evasion attack [20]. The public availability of training dataset and unlabelled text datasets are not need security, they improve their security strength when they convert as labeled data [21]. The way of convert the Reuters text training dataset into labeled dataset using the Regeneration algorithm and output is shown below:

Input ----> Load Unstructured Training Dataset
Output ---> New genenerated formatted Training Dataset
1. Keyword = New extracting word
2. For line in text:
3. If Keyword in line:
4. Write the keyword
5. Count = Count +1
6. If Count>0
7. Print No of time the Keyword Found
8. If not found:
9. Goto New keyword
10. End for

Algorithm 1: Regeneration algorithm

The Regeneration algorithm used to accept the keyword from the user and extract the keywords from the text file. The extracted words are saved under some labeled parameters. The counting words are saved under the parameter word count. The output of the Regeneration algorithm is shown in figure7.

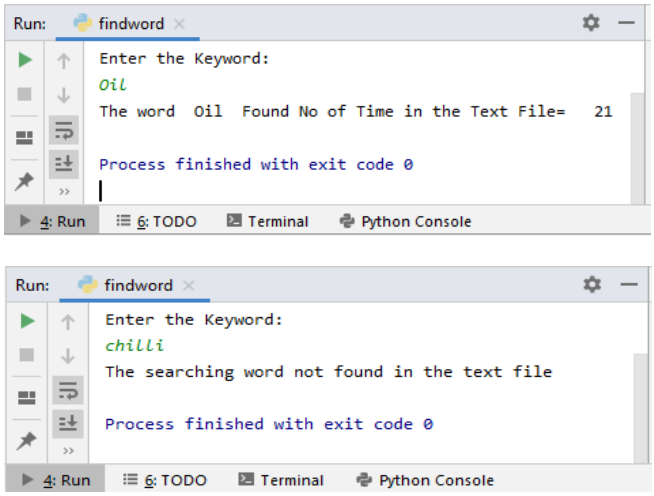


Fig 7: Output of Regeneration algorithm

The parameters Class_No, Class_Name, No_Documents, Mean_word formed as labels and the parameter's values are retrieved from the text formatted documents using the Regeneration algorithm. The values are changed into structural format and use in ML algorithm.

Comparision of Evasion Attack Defense Methods

The defense algorithms against evasion attack to avoid worst case attack on training datasets and secure ML algorithms performance. The different kind evasion attack defense models are listed in the below table1.

Table 1: Evasion Attack Defense Methods.

Defense Methods	Advantage	Disadvantage
Secdefender [16]	Perfect resilient solution against learning system knowledged attacker.	Defender has no knowledge about the attack.
DeepWordBug [4]	Four different transformer functions Substitution, Insertion, Deletion, Swap to change the attack words and form original dataset.	The prediction accuracy of training dataset decreases for editing word distance limited.
Defensive distillation [17, 22]	Trained as usual. Soft class labels probabilities are compared to hard class labels.	It evaluated on DNN architecture.
SVMPW [18]	Best detection model against evasion attack.	Absence of to increase performance of SVM.

The proposed algorithm regeneration created original text classified training dataset from the unstructured dataset. The new generated dataset train on the machine learning, it gives best classification, prediction, accuracy performance. Comparing our proposed defense algorithm with other defense algorithm models, the regeneration algorithm produced good result and trustable training dataset.

Conclusion

In this paper provide how to generate trusted training dataset from unstructured formatted training dataset and secure the originality of machine learning result performance. The users can identify the evasion attacked training dataset through ML algorithm measurement difference. The evasion attack detected, the defense method regenerated original dataset from the unstructured dataset. So the users train learning algorithms with original training datasets and prevent the ML algorithm's performance. The future work implemented by to analyze and retrain the security algorithm with new collected training dataset. In future, the research work will extent to handle how the Defensive distillation defense method work on text classified training dataset.

References

1. Fei Zhang, Patrick Chan PK, Battista Biggio, Daniel Yeung S, Fabio Roli," Adversarial Feature Selection against Evasion Attacks", Article April 2015, https://www.researchgate.net/publication/275355520.
2. Takeru Miyato, Andrew Dai M, Ian Goodfellow. "Adversarial Training Methods for Semi-Supervised Text Classification", Published as a conference paper at ICLR, 2017.
3. Matthias Hein, Maksym Andriushchenko, "Formal Guarantees on the Robustness of a Classifier against Adversarial Manipulation", 31st Conference on Neural Information Processing Systems, Long Beach, CA, USA. NIPS, 2017.
4. Ji Gao, Jack Lanchantin, Mary Lou Soffa, Yanjun Qi. "Black-box Generation of Adversarial Text Sequences to Evade Deep Learning Classifiers", arXiv:1801.04354v5 [cs.CL] 23 May 2018.
5. Alhussein Fawzi, Omar Fawzi, Pascal Frossard. "Analysis of classifiers robustness to adversarial perturbations" arXiv:1502.02590v4 [cs.LG] 28 Mar 2016.
6. Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Srndi, Pavel Laskov *et al.* "Evasion attacks against machine learning at test time", arXiv:1708.06131v1 [cs.CR] 21 Aug 2017.
7. Bogdan Kulynych, Jamie Hayes, Nikita Samarin, Carmela Troncoso. "Evading Classifiers in Discrete Domains with Provable Optimality Guarantees", arXiv:1810.10939v3 [cs.LG] 1 Jul 2019.
8. Weilin Xu, Yanjun Qi, David Evans. "Automatically Evading Classifiers", In Network and Distributed System Security Symposium 2016 (NDSS), San Diego, February 2016.
9. Aru Okereke Eze, Chiaghana Chukwunonso E, "Malware Analysis and Mitigation in Information Preservation", IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661,p-ISSN: 2278-8727, 20, 4, Ver. I (Jul - Aug 2018), PP 53-62, DOI: 10.9790/0661-2004015362.
10. LiangTong, Bo LI, Chen Hajaj, Chaowei Xiao. "Improving Robustness of ML Classifiers against Realizable Evasion Attacks Using Conserved Features", Proceedings of the 28th USENIX Security Symposium. August 14–16, 2019 • Santa Clara, CA, USA 978-1-939133-06-9.
11. Javid Ebrahimi, Anyi Rao, Daniel Lowd, Dejing Dou. "HotFlip: White-Box Adversarial Examples for Text Classification", arXiv:1712.06751v2 [cs.CL] May 2018.https://www.researchgate.net/publication/321936350.
12. Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, Michael K. Reiter, "A General Frame work for Adversarial Examples with Objectives", ACM Transactions on Privacy and Security, 22(3), Article16. June 2019, https://doi.org/10.1145/3317611.
13. Bo Li, Yevgeniy Vorobeychik, Xinyun Chen. "A General Retraining Frame work for Scalable Adversarial Classification", Workshop on Adversarial Training, Barcelona, Spain. NIPS, 2016.
14. Suja Mary D, Suriakala M. "Evasion Attack on Text Classified Training Datasets", International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249-8958, Volume-8 Issue-6S, August, 2019.
15. Yi Shi, Yalin Sagduyu E. "Evasion and Causative Attacks with Adversarial Deep Learning", Conference Paper · October 2017 DOI: 10.1109/MILCOM.2017.8170807.

16. Lingwei Chen, Yanfang Ye, Thirimachos Bourlai. "Adversarial Machine Learning in Malware Detection: Arms Race between Evasion Attack and Defense", European Intelligence and Security Informatics Conference (EISIC), IEEE, 18 Dec 2017. doi: 10.1109/EISIC.2017.21.
17. Yonghong Huang, Utkarsh Verma, Celeste Fralick, Gabriel Infante-Lopez, Brajesh Kumar, Carl Woodward. "Malware Evasion Attack and Defense", IEEE/IFIP International Conference on Dependable Systems and Networks Workshops, 2019.
18. Zeinab Khorshidpour, Sattar Hashemi, Ali Hamzeh. "Learning a Secure Classifier against Evasion Attack", IEEE 16th International Conference on Data Mining Workshops, 2016.
19. Muni Prashneel Gounder, Mohammed Farik. "New Ways to Fight Malware", International Journal of Scientific & Technology Research Volume 6, Issue 06, June, 2017.
20. Paolo Russu, Ambra demontis, Battista Biggio, Giorgio Fumera, Fabio Roli. "Secure Kernel Machines against Evasion Attacks", AISec'16, Vienna, Austria, October 28, 2016.
21. Nicolas Papernot, Patrick McDaniel, Arunesh Sinha, Michael P. Wellman, "SoK: Security and Privacy in Machine Learning", IEEE European Symposium on Security and Privacy, 2018.
22. Nicolas Papernot, Patrick McDaniel, Xi Wu§, Somesh Jha§, Ananthram Swami. "Distillation as a Defense to Adversarial Perturbations against Deep Neural Networks", 37th IEEE Symposium on Security & Privacy, San Jose, CA, IEEE, 2016.