International Journal of Engineering Trends and Technology
ISSN: 2231 – 5381 /doi:10.14445/22315381/IJETT-V69I11P206

Volume 69 Issue 11, 47-52, November, 2021
©2021 Seventh Sense Research Group®

SQLI Detection Based on LDA Topic Model

Nilesh Yadav[1], Dr. Narendra Shekokar[2]

[1]*Research Scholar, Department of Computer Engineering, DJSCE, Vile Parle, India.*
[2]*Professor, Department of Computer Engineering, DJSCE, Vile Parle, India*

[1]nileshyadav2004@gmail.com, [2] narendra.shekokar@djsce.ac.in

Abstract - *Structured Query Language Injection (SQLI) is the topmost dangerous web application vulnerability in all web attacks, and this causes serious harm to the entire web system. Due to the heterogeneous nature of this attack, its detection remains a challenging problem. Researchers started using the Machine Learning (ML) based approach to mitigate this attack, but ML-based techniques heavily depend on the accuracy of feature extraction. To get more useful reduced features and improve accuracy, consider the semantic consistency and proper probability distribution of the words. The proper reduced dimensions improve the text classification process. Therefore, this paper uses a topic modeling-based Latent Dirichlet Allocation concept as a dimensionality reduction technique to acquire informative features. It helps to grab the more useful features by considering the semantic co-occurrence between the observed words from logs. This topic-modeling concept can act here as an efficient feature reduction technique and extracts the more valuable features from the most dangerous vulnerability logs. The paper explores the efficient detection of SQLI. The ECML/PKDD-2007 HTTP traffic logs experiments used supervised ML techniques and evaluated the results using accuracy matrix, performance time, and ROC curve.*

Keywords — *Attack, SQLI, Latent Dirichlet Allocation, Dimension Reduction, ECML.*

I. INTRODUCTION

Due to heterogeneous attack vectors and hidden attack structures, SQLI is the topmost web vulnerability [1] in web attacks. How to defend SQLI attack effectively becomes the most challenging task in front of web security developers. ML classifiers can classify the SQLI weblogs, but efficiently reducing the features and creating the final reduced dataset with labels are challenging tasks as the ML model's accuracy primarily depends on the feature-engineering concept. To improve text classification, we always need enlightening methods to characterize the useful features from words [2], [3]. ML techniques cannot use latent information in the documents to generate better classifiers. Therefore, we required the expressive method, which could play an essential role in knowledge extraction. If we think of words as the smallest informative measure of a document, the Topic Model concept can analyze the connections between words and documents. Latent Dirichlet allocation (LDA) is a topic modeling technique that uses a probabilistic model to find the semantic co-occurrence between observed words and latent, i.e., hidden topics from data collection. The topics produced by the LDA model from the documents can be used as dimensions/features by the classifier in text classification. Combining topic information and word information may result in a better feature set and improve a text classifier's accuracy.

In our previous experimentation [6] on ECML/PKDD-2007 HTTP traffic logs, we have used the ECML Dataset, which is generated and pre-processed. Initially, "Term-Frequency_Inverse Document-Frequency_Ngram with singular value decomposition," i.e., Combined methodology (TF-IDF_Ngram with SVD) [6] is used for SQLI detection. During feature reduction, this technique does not take into account the semantic information of texts. Here, the topic modeling-based 'LDA' approach is used as a feature extraction technique. The LDA transformation with data labeling is done separately for Malicious and non-malicious data logs. The collective outcome of this is used further with classifiers for SQLI detection. We are using the ECML dataset for further experimentation. Supervised classifiers are used for evaluation.

In the paper representation, Section 1 is introductory. The 2nd section discusses related work. Section 3 explains the TFIDF-Ngram with the LDA approach as an enhanced design. The mathematical explanation of the LDA algorithm is presented in Section 4. The experimental results are presented in Section 5. Finally, we concluded the paper

II. RELATED WORK

A brief overview of several research papers and other resources, which were analyzed, are given below. Paper [7] proposed a generic methodology based on topic modeling and data mining concept. Here, the authors evaluated the teaching-learning process by modeling university open-ended survey questions. The author [8] analyzed the research project data available for Thailand country and used the topic-modeling concept to reveal the research themes and trends in this data. In [9] [17] [18], investigated text categorization using topic-modeling technique. The researchers used a Topic model-based LDA approach in the classification process. A. Terko, D. Donko [10] did the text classification. The data labeling is done using KL divergence on the outcome of the LDA model and finally transformed into a feature vector. These vectors are the inputs to various classification techniques for the classification of scientific publication data.

[11], introduces new multi-label datasets by concatenating the topics of label sets and text features to improve text classification performance. To diagnose

Parkinson's disease using these hybrid features, a new data called the Parkinson's dataset was created. I. Deliu and K. Franke [12] used the classifiers with the topic model very interestingly. They help Cyber Threat Intelligence. Initially, ML Classifiers filter the malicious posts, and then LDA is used for topic estimation from these posts. Turkish tweet emotions are classified using the n-stage LDA [13]. In this study, the classical LDA model is compared to the two-stage, three-Stage LDA model. In [14], the Part of Speech (POS) based feature selection method is used, which has positively impacted the LDA technique for the sentiment classification. The genetic algorithm-based hybrid approach is used to discover the optimal weights for topics, and these topics are extracted using LDA in [15] research. Online reviews of game apps were used as the data set here. Authors [16] created the short text dataset from the Sina news website. They employed the LDA to measure the topic similarity and utilized it as a distance matrix of the KNN algorithm to classify short texts. [19] used multilevel LDA to produce a low-dimensional data representation, allowing clustering techniques that cannot run on high-dimensional feature space datasets. Here, the researchers used the datasets ‘Friend Feed social network’ and ‘Reuters-21578’.

However, several approaches used the Topic Modeling technique for cataloging, but no one has analyzed the SQL injection logs separately for feature reduction and attack detection accuracy improvement. Here, LDA is used to maintain the semantic consistency between the words and create latent topics, acting as reduced features. In this paper, we evaluated a novel approach on filtered ECML SQLI logs.

III. IMPROVED COMBINED APPROACH

The existing statistical dimension reduction techniques like SVD are mathematical feature extraction techniques. While dimension reduction, the algebraic solutions are not using the logical information between the words. When grouping words into one scheme as a feature, these statistical techniques do not consider their semantic consistency. To address this problem, we explored the LDA technique and validated it in this study. For SQLI detection, we developed a novel Improved Combined method.

In the ‘Improved Combined’ approach. Initially, SQLI logs are separated and pre-processed from ECML-PKDD_2007 HTTP Request web-logs. Feature creation (TF-IDF-Ngram), feature extraction using LDA, and labeling of data is done separately for malicious (SQLI requests) and non-malicious (normal request) data. Furthermore, concatenate the outcome to create the final reduced dataset. Two supervised ML algorithms are used to classify this final dataset. Evaluated results using accuracy matrix, performance time, and ROC curve.

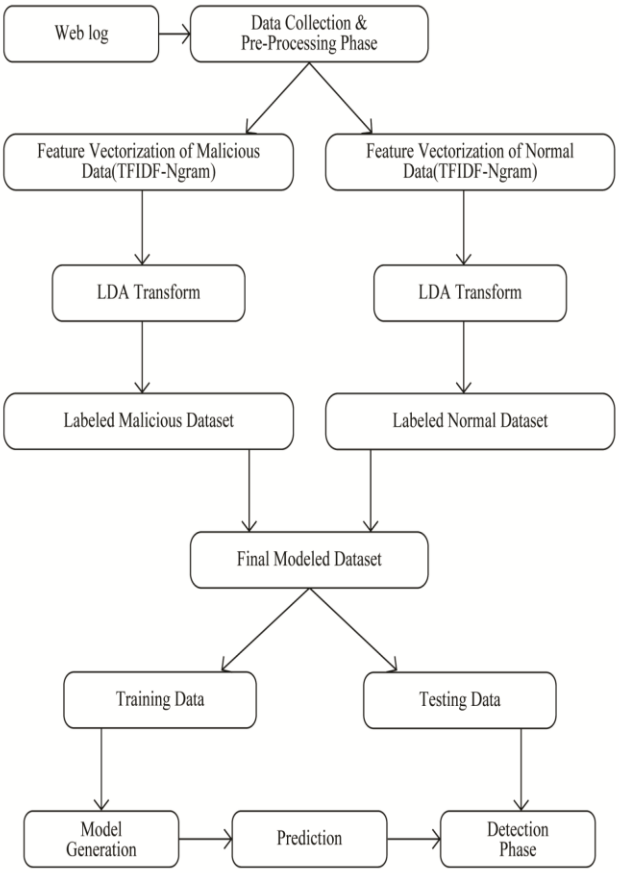

Web log
Data Collection & Pre-Processing Phase
Feature Vectorization of Malicious Data(TFIDF-Ngram)
Feature Vectorization of Normal Data(TFIDF-Ngram)
LDA Transform
LDA Transform
Labeled Malicious Dataset
Labeled Normal Dataset
Final Modeled Dataset
Training Data
Testing Data
Model Generation
Prediction
Detection Phase

Fig. 1: The Process flow of the Improved Combined System.

The complete experimental approach for designing the Improved Combined method is depicted in Fig. 1. The SQLI HTTP requests are extracted and subsequently pre-processed [6] to acquire filtered SQLI data from the available ECML weblog, which comprises a mixture of all web attack logs. The experimentations were conducted on these filtered ECML HTTP requests. We built the necessarily reduced dataset separately for malicious (SQLI requests) and non-malicious (regular request) data because the raw data could not be directly fed into the ML classifiers. The Term-Frequency_Inverse Document-Frequency (TF-IDF) with Ngram (3, 3) technique is used for the feature vectorization of the malicious log. How a pertinent word with request log in the collection of logs is measured by TF-IDF here and Ngram improves this measurement by considering the adjacent word sequence information.

Initially, the feature vector has many dimensions that likely could increase the complexity of the model and lead to more computational cost. The diverse dimension reduction technique should be used to resolve this issue. The main idea is to reduce high-dimensional space into a lower-dimensional subspace using the topic-modeling concept. The LDA maintains the semantic consistency, generates latent topics, which are fewer in number, and creates a strong informative reduced feature space. To create the low dimensional subspace, the bad request's feature vectorizer is fitted and transformed to become LDA transformed.

Similarly, as shown in Fig. 1, we have processed non-malicious data and created the reduced normal dataset. The dataset items are labeled programmatically and separately. Further concatenated these outcomes to create the final reduced dataset. This final reduced informative feature set helped improve the classification model's overall detection accuracy and performance. Two robust and supervised ML algorithms, i.e., SVM & RF, are used for the SQLI detection. The outputs are evaluated using confusion matrices, graphs, and ROC curves.

A probabilistic graphical model or generative model using the plate notation with mathematical expressions of the LDA algorithm is explained in the next section.

IV. LATENT DIRICHLET ALLOCATION

The document is a collection of words that we can see, and the topics have various words, but the topics are latent [20] [21]. LDA is one of the most popular topic modeling methods [22]. LDA finds the topics, which belong to a particular document. Based on the words present in the document, LDA chooses a topic mixture according to Dirichlet distribution. It will be used to group related words into a logical scheme, acting as a reduced dimension.

Fig. 2 is the graphical representation of LDA [23]. As shown in this diagram, the shaded nodes are the observed variables. Wd n is the only shaded part, and the rest all are hidden random variables. Where,

D: a collection of corpus/documents.
K: Number of topics.
N: Total unique words.

All variables and details are shown in Fig. 2. In this generative model, for eta (η) (Dirichlet higher hyper-parameter), draw beta i (β_i) for each topic (for all K topics). For each document d, find the topic proportions θ_d. Then for Dirichlet Distribution over α, calculate the topic proportions for this particular document. After that, for each observed word, Draw $Z_{d,n}$ from a multinomial distribution over θ_d and draw a word from this.

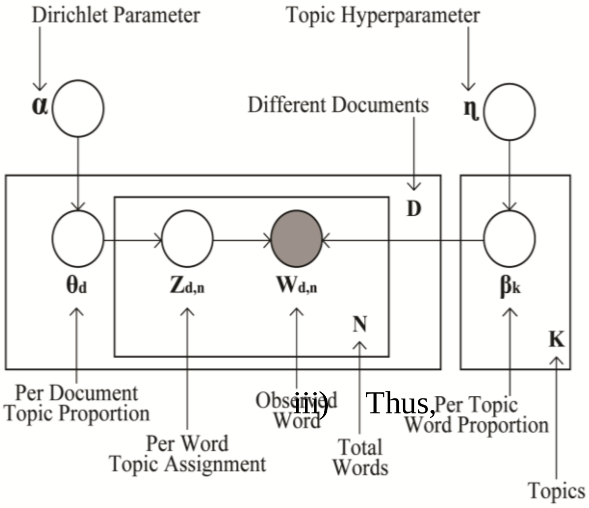


Fig. 2: The Graphical Representation of LDA

The steps of this procedure are as follows:-

1) Draw each topic $\boldsymbol{\beta_i}$ ~ Dir ($\boldsymbol{\eta}$), for i € {1, 2, 3.....k}

2) For Each Document:

   a) Draw the topic proportions $\boldsymbol{\theta_d}$ ~ Dir($\boldsymbol{\alpha}$).

   b) For each word :

      1. Draw $\mathbf{Z_{d,n}}$ ~ Multi($\boldsymbol{\theta_d}$).

      2. Draw $W_{d,n}$ ~ Multi($\boldsymbol{\beta_{Zd,n}}$).

To distribute the words of the documents to several hidden themes, the Dirichlet distribution effect is utilized. The main goal here is to deduce the following values.

- Per-word topic assignment $\mathbf{Z_{d,n}}$
- Per-document topic proportions $\boldsymbol{\theta_d}$
- Per-corpus topic distributions $\boldsymbol{\beta_k}$

A very ingenious Dirichlet distribution technique is used to formulate/calculate the per-document topic_distribution over the simplex:-

$$\mathbf{P(\theta /\alpha)} = [\,\Gamma(\textstyle\sum_i \alpha_i) / \prod_i \Gamma(\alpha_i)] * [\prod_i \theta_i^{\alpha i\,-1}] \;\ldots\ldots\ldots(1)$$

To infer the above parameter values, there are different ways of doing it. One of the approximate estimation techniques is the collapsed Gibbs Sampling method [24].

Markov_Chain_Monte_Carlo (MCMC) has a particular case called Gibbs Sampling [25]. Gibbs sampling considers each word token one at a time and estimates the likelihood of assigning the current word token to each topic based on the topic assignment of all other word tokens are known. Mathematically can be expressed as

$$\mathbf{P(\,Z_{i\,=\,j} \mid Z_{-i},\, w_i,\, d_i\,)} \ldots\ldots\ldots\ldots\ldots\ldots\ldots(2)$$

For the i^{th} token, the corresponding word id w_i and document id is d_i. Z_i = j is the probability that the i^{th} word should be assigned to topic j, and the minus sign in Z - i is everything other than this i. It should depend on two things.

i) How likely is topic j to be assigned to $d_{i,}$ i.e., matrix $\mathbf{C^{DT}}$**?**

ii) How likely is this word occurring in topic j is the word w_i for topic j, i.e., the matrix $\mathbf{C^{WT}}$**.**

Thus,

$$P(Z_{i\,=\,j} \mid Z_{-i},\, w_i,\, d_{i,.})\ \propto$$

$$[(C_{wij}{}^{WT}+\eta) / (\textstyle\sum_{w=1}^{W} C_{wj}{}^{WT}+W\eta)] * [(C_{dij}{}^{DT}+\alpha) / (\textstyle\sum_{j=1}^{T} C_{dj}{}^{DT}+T\alpha)] \ldots\ldots(3)$$

And Estimating θ & β

$$\boldsymbol{\beta_i^{(j)} = [(\,C_{ij}{}^{WT} + \eta\,) / (\textstyle\sum_{k=1}^{W} C_{kj}{}^{WT} + W\eta)]} \ldots\ldots\ldots\ldots(4)$$

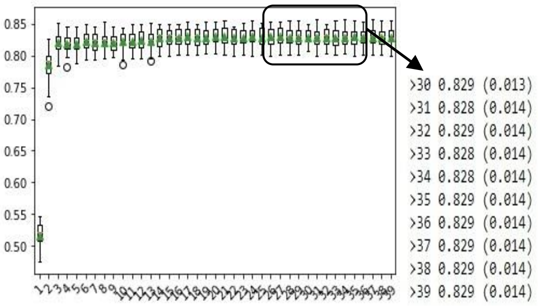
$$\boldsymbol{\theta_j^{(d)} = [(\,C_{dj}{}^{DT} + \alpha\,) / (\textstyle\sum_{k=1}^{T} C_{dk}{}^{DT} + T\alpha)]} \ldots\ldots\ldots\ldots(5)$$

These values will correspond to the distribution of sampling a new token of the word i from topic j and sampling a new token in document d from topic j.

V. EXPERIMENTS AND RESULT EVALUATION

Experimentation and analysis are explained in this section. We evaluated the performance of the improved method on an ECML pre-processed dataset [6]. Finally, ECML balanced dataset has 2274 SQLI HTTP requests and 2274 normal HTTP requests. Models were implemented in Python language using the Scikit library [26] [27]. In evaluating the models, we compared the 'Improved Combined' model (TF-IDF-Ngram & LDA) to the 'Combined' model (TF-IDF-Ngram & SVD). Linear SVM (LSVM) and Random Forest (RF), these two supervised classifiers, are used and, to get unbiased results, conducted the experimentation in 10-fold cross-validation.

A program is created to select the best number of components by calculating the explained_variance _ratio_ values for SVD. The number of features is selected based on the set threshold value. Compared the threshold value with the variance value and selected the number of components. This selected value is also verified by evaluating the model, which uses the mentioned ML classifiers. The model is executed using 10-fold cross-validation. Calculated the Mean and Standard Deviation scores. As shown in Fig. 3, this depicts the box chart diagram of the model using the SVM classifier. It is observed for the models that, after the particular number of components, especially 35 onwards, the models are giving the same constant score values. Therefore after a range of executions, finally, the optimum number of components selected here is = 37.



```
>30 0.829 (0.013)
>31 0.828 (0.014)
>32 0.829 (0.014)
>33 0.828 (0.014)
>34 0.828 (0.014)
>35 0.829 (0.014)
>36 0.829 (0.014)
>37 0.829 (0.014)
>38 0.829 (0.014)
>39 0.829 (0.014)
```

Fig. 3: Selection of Best Number of Components.

The outputs of the models are shown in result tables I and II, graphs 4 and 7, and curves 5,6,8,9. Accuracy matrix, performance time, and ROC curves are used to analyze it.

Table I. C.M. of 'Combined' Technique

Measure	**ML → Classifiers**	**RF**	**Linear SVM**
(True_Positive)TP		173	163
(True_Negative)TN		205	214
(False_Positive)FP		22	13
(False_Negative) FN		54	64
Accuracy (%)		**83.26**	**83.02**
Precision (%)		88.79	92.82
Recall (%)		76.21	71.59
F1_Score (%)		81.99	80.79
Train_time (s)		**0.190**	**0.0223**
Test_time (s)		**0.0037**	**0.0005**

Initially, TF_IDF-Ngram & SVD technique is used for reduced dataset preparation. Here for feature reduction, The SVD technique is used, and then the execution of both the mentioned classifiers. Table I shows the confusion Matrix (C.M.) of models on the ECML dataset using the combined technique. We observed that the training and testing time is more for the RF algorithm than the LSVM algorithm. The accuracy of the linear SVM (83.02 %) is somewhat less or equal to that of the RF (83.26%).

The data representation for given classifiers is shown in box and strip plots. The data visualization in relation to the classifiers is shown in Fig. 4. As we can observe, while using the SVD as a dimension reduction technique, the data is more spreader for both classifiers, so the accuracy is less.

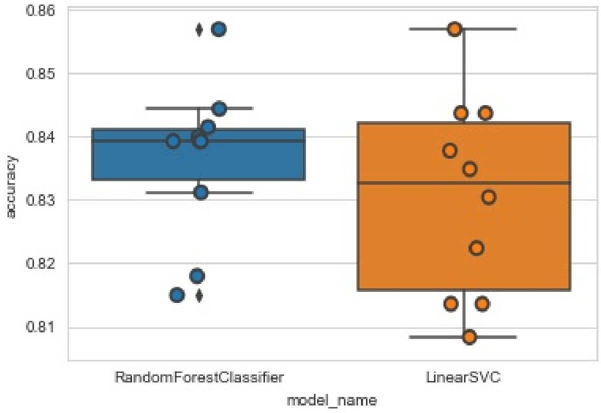


Fig 4: Accuracy Plot for 'Combined' Technique

Additionally, the average performance of the model using different classifiers is evaluated using Receiver Operating Characteristic (ROC) curve. It also helps to analyze the model separately for each scheme. The classification model is better when the area under this curve, i.e., AUC, is near to one and the curve is close to the top left corner.

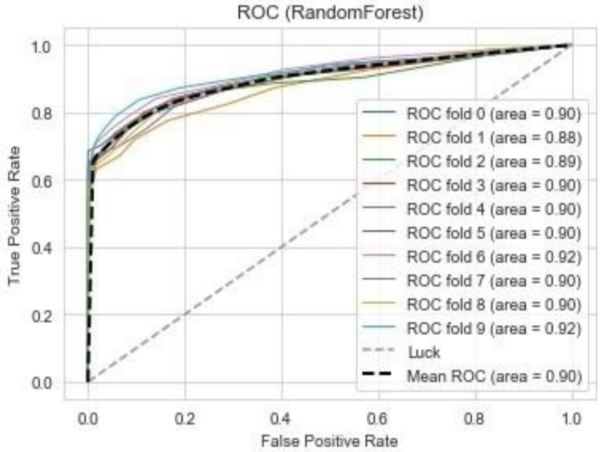


Fig. 5: ROC curve of Combined Scheme Using RF

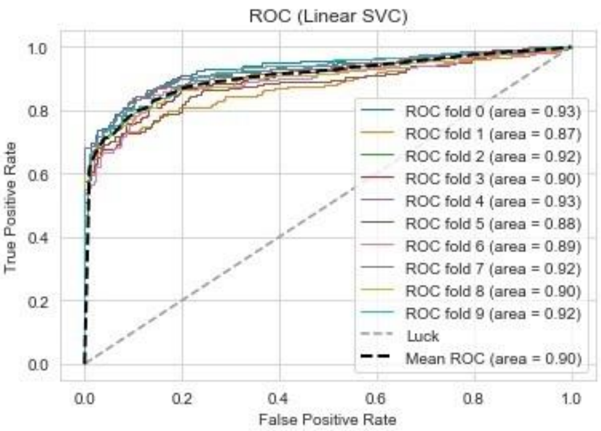


Fig. 6: ROC Curve of Combined Scheme Using LSVM

The two graphs 5, 6 are for the area of classification model using a ‘Combined’ scheme. Graphs 5 and 6 are for the RF and SVM algorithms used in the Combined scheme model. Ten-fold cross-validation is used, and the mean ROC value, i.e., the area, is 0.90 for both classifiers.

The CM shown in table II is for the Improved Combined scenario where along with TF-IDF_Ngram, the LDA scheme is used as a feature reduction.

The observed values show that the training and testing time is more for the RF algorithm than the linear SVM algorithm. The accuracy of the RF is more (99.85%) than the SVM (98.70%) because the RF learning process is boosted slightly due to semantic relations between words while building the trees.

Table II. C.M. of ‘Improved Combined' Technique

Measure	**ML → Classifiers**	**R.F.**	**Linear SVM**
(True_Positive)TP		226	224
True_Negative)TN		227	224
(False_Positive)FP		0	3
(False_Negative)FN		1	3
Accuracy (%)		**99.85**	**98.70**
Precision (%)		99.91	98.68
Recall (%)		99.85	98.72
F1_score (%)		99.84	98.70
Train_time(s)		**0.056**	**0.012**
Test_time(s)		**0.0040**	**0.0003**

Fig.7 depicts data visualization using the Improved Combined method in respect to ML classifiers. Here it is observed that, due to the LDA technique, the data is not that much spread & LDA gives more latent informative feature sets, so the accuracy is improved over here.

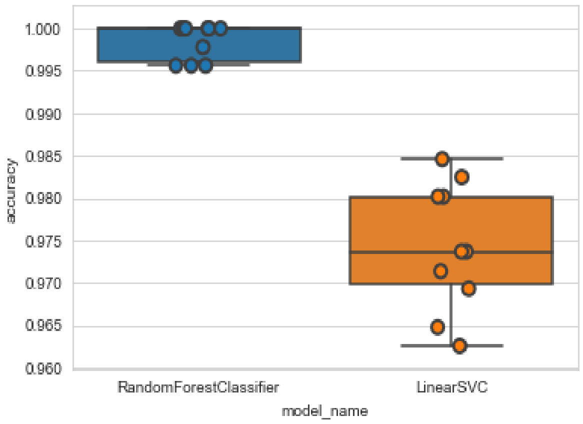


Fig. 7: Accuracy Plot for Improved Combined Technique

Similarly, we plotted graphs 8, 9 for the Improved Combined scheme using the RF and SVM algorithms, respectively. For both the classifiers, the mean ROC, i.e., the area, is 0.99.

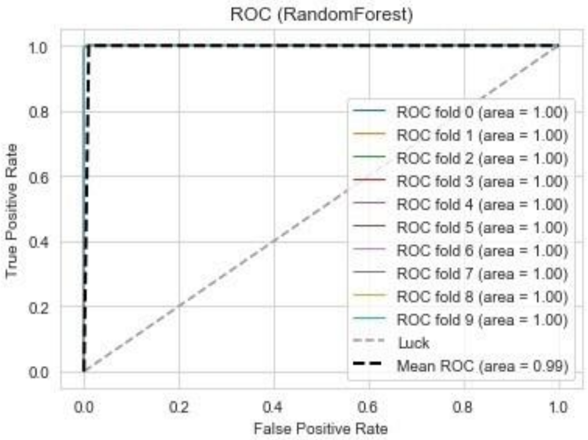


Fig. 8: ROC curve of Improved Combined Scheme for RF

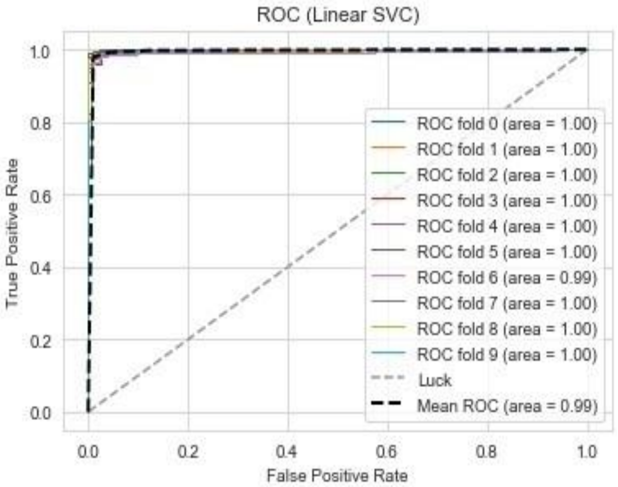


Fig. 9: ROC curve of Improved Combined scheme for LSVM

Overall, in the Improved Combined approach, both these classifiers' training and testing time are less than the Combined approach. The training and testing time is very less for SVM, and the accuracy for RF and SVM is more in the Improved scheme. AUC is higher (0.99) for the Improved scheme over the Combined scheme. A significant improvement in detection accuracy and performance is due to using the topic-modeling concept in SQLI detection. The 'latent topics' and 'discriminative word information' effectively help in the entire process.

VI. CONCLUSION

In this research, we tested a novel improved scheme on the SQLI ECML_PKDD_2007 dataset for SQLI detection. In SQLI's novel detection scheme, we used the TF_IDF_Ngram technique to vectorize or construct weblogs' features. The generated features are large in numbers, so the Latent Dirichlet Allocation (Topic Modeling) concept is used to reduce the feature space. The semantic consistency is maintained by LDA and generates latent topics/schemes, which are fewer in numbers, which acts as a reduced feature space. We achieved better results than the previous study on ECML logs because the LDA reduces the dimensionality. It created more informative features from the logs, which leads to less computational cost, and decreases the complexity of the model. The explained_variance_ratio_ of a dimension reduction technique & the mean scores of the models acted as guidelines to choose the best number of components. Overall results showed that both supervised ML classifiers' accuracy is significantly more than that of the Combined technique in the Improved Combined approach. The training time and testing time for both the classifiers in the Improved scheme is less. In addition, this scheme has the highest TPR and mean ROC, i.e., area. Hence, the LDA technique is more efficient than SVD for dimension reduction in SQLI detection.

REFERENCES

[1] OWASP Group., Top 10 Most Critical Web Application Security Vulnerabilities, (2021). [online]. Available: https// www.owasp.org/ index. php.

[2] X. Pan and H. Assal, Providing context for free text interpretation International Conference on Natural Language Processing and Knowledge Engineering, Proceedings. 2003, Beijing, China, (2003) 704-709 .

[3] Sebastiani, F.: Classification of text, automatic. In Brown, K., ed.: The Encyclopedia of Language and Linguistics, Volume 14, 2nd Edition. Elsevier Science Publishers, Amsterdam, (2006) 457–462.

[4] Knowledge Discovery in Databases: ECML/PKDD 200, 11th European Conference on Principles and Practice of Knowledge Discovery in Databases, Poland, (2007) 17-21.

[5] Gallagher, B., and Eliassi-Rad T., Classification of http attacks: a study on the ECML/PKDD discovery challenge, Technical Report No. LLNL-TR-414570. Lawrence Livermore National Laboratory, Livermore, CA, (2007) (2009).

[6] N. Yadav, Dr. N. Shekokar, Preprocessing HTTP Requests and Dimension Reduction Technique for SQLI Detection, Lecture Notes in Networks and Systems, Conference Proceedings of ICDLAIR2019, MNIT, Jaipur, India. Springer, (2021) 190-200.

[7] D. Buenaño-Fernandez, M. Gonzalez, D. Gil, and S. Luján-Mora, Text Mining of Open-Ended Questions in Self-Assessment of University Teachers: An LDA Topic Modeling Approach in IEEE Access, 8 (2020) 35318-35330,.

[8] N. Sethasathien and P. Prasertsom, Research Topic Modeling: A Use Case for Data Analytics on Research Project Data, 1st International Conference on Big Data Analytics and Practices (IBDAP), Bangkok, Thailand, (2020) 1-6.

[9] L. Xia, D. Luo, C. Zhang, and Z. Wu, A Survey of Topic Models in Text Classification, 2nd International Conference on Artificial Intelligence and Big Data (ICAIBD), Chengdu, China, (2019) 244-250 .

[10] A. Terko, E. Žunić, and D. Đonko, NeurIPS Conference Papers Classification Based on Topic Modeling, International Conference on Information, Communication and Automation Technologies (ICAT), Sarajevo, Bosnia and Herzegovina, (2019) 1-5.

[11] W. Sun, X. Ran, X. Luo, and C. Wang, An Efficient Framework by Topic Model for Multi-label Text Classification, International Joint Conference on Neural Networks (IJCNN), Budapest, Hungary, (2019) 1-7.

[12] I. Deliu, C. Leichter and K. Franke, Collecting Cyber Threat Intelligence from Hacker Forums via a Two-Stage, Hybrid Process using Support Vector Machines and Latent Dirichlet Allocation, WA, USA, (2018) 5008-5013.

[13] Z. A. Guven, B. Diri, and T. Cakaloglu, Classification of New Titles by Two-Stage Latent Dirichlet Allocation 2018 Innovations in Intelligent Systems and Applications Conference (ASYU), Adana, (2018) 1-5.

[14] E. S. Usop, R. R. Isnanto, and R. Kusumaningrum, Part of speech features for sentiment classification based on Latent Dirichlet Allocation 4th International Conference on Information Technology, Computer, and Electrical Engineering (ICITACEE), Semarang, (2017) 31-34.

[15] C. Hsu and C. Chiu, A hybrid Latent Dirichlet Allocation approach for topic classification IEEE International Conference on Innovations in Intelligent Systems and Applications (INISTA), Gdynia, (2017) 312-315.

[16] Q. Chen, L. Yao and J. Yang, Short text classification based on LDA topic model International Conference on Audio, Language and Image Processing (ICALIP), Shanghai, (2016) 749-753.

[17] Y. Chen and S. Li, Using latent Dirichlet allocation to improve the text classification performance of support vector machine IEEE Congress on Evolutionary Computation (CEC), Vancouver, (2016) 1280-1286.

[18] Z. Li, W. Shang, and M. Yan, News text classification model based on a topic model, IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS), Okayama, (2016) 1-5.

[19] Inkpen, Diana & Razavi, Amir, Topic Classification using Latent Dirichlet Allocation at Multiple Levels 15th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2014), Nepal, (2014).

[20] D. M. Blei, A.Y. Ng, & M. I. Jordan, Journal of Machine Learning Res. 3 (2003) 993–1022.

[21] S. Thiyagarajan High user Experience by Providing Relevant News Articles using Topic Modelling, SSRG-International Journal of Engineering Trends and Technology (IJETT) – Volume 55, Number 1, January 2018. ISSN: 2231-5381.

[22] G. P. Paul, Sricharukesh, e.g., S. Vigneshkumar, G. Kannan, Advanced Scalable Algorithm for Community Question Answering Using Post Voting Prediction, SSRG International Journal of Computer Science and Engineering (SSRG-IJCSE) – Special Issue ICETSST – (2018) . ISSN: 2348 – 8387

[23] nptel.ac.in., Natural Language Processing Lec. 42, (2003) [online]. Available:https://nptel.ac.in/courses/106/105/106 105158/, last accessed 2021/1/15.

[24] Thomas L Griffiths and Mark Steyvers, Finding scientific topics Proceedings of the National Academy of Sciences, 101(l 1), (2004) 5228–5235.

[25] Stuart Geman and Donald Geman, Stochastic relaxation, Gibbs distributions, and the bayesian restoration of images, Pattern Analysis and Machine Intelligence, IEEE Transactions on, 6 (1984) 721–741.

[26] Scikit-learn.org scikit-learn: Machine Learning in Python, (2021). [online]. Available: https://scikit-learn.Org/stable/ index.html, last accessed 2021/4/27.

[27] Anaconda Distribution (2021), [online]. Available: https://www.anaconda.com/distribution.