Grenze International Journal of Engineering and Technology, June Issue

Cybersecurity Risk Prediction

Abha Kiran Rajpoot[1], Abhimanyu Rana[2], Shivam Dwivedi[3] and Siddharth Singh[4]
[1]Professor, Galgotias University, School of Computer Science & Engineering, Greater Noida, India
[2-4]School of Computer Applications and Technology, Galgotias University, Greater Noida, India
Email: asr.codez@gmail.com, luckyshiv420@gmail.com, siddharth10sing@gmail.com

***Abstract*— Predicting Cyberattacks with the help of Machine Learning.**
Learning (PCAML) is a system that predicts cyber threats using a range of data sets and artificial intelligence. PCAML models have high precision and low false positives through the analysis of system logs and user behaviour based on logistic regression, Support Vector Machines, Multinomial Naive Bayes, and TFIDF. Having its proactive defence strategies and early warning features, this technology can revolutionize the process of cybersecurity. PCAML has immense potentials of enhancing security within organisations and lower the risk of cyber attacks even though challenges such as accuracy and evolving threats exist.

***Key Words*— PCAML, Machine learning, forecast cyber attacks, TFIDF, Mitigating risks.**

I. INTRODUCTION

Machine learning models like TF-IDF, Logistic Regression, SVM, and Multinomial Naive Bayes can aid cybersecurity by analysing previous data and trends and making specific predictions about hazardous activity. TF-IDF recognizes the key indicators of cyber danger by deriving key textual features. Logistic regression provides understandable information of the risk factors as it probabilistically categorizes attacks. Such an overarching strategy enhances a security posture and this is a significant advancement in predicting cyber threats.

A. Algorithm usage and Selection

Multinomial naive bayes involves the use of probabilities in text classification in predicting cyberattacks. SVM is ideal where the boundary of the decisions is hard because it can receive both linear and non-linear data. The logistic regression is a simple way of binary classification tasks. TF-IDF is beneficial in forecasting cyber attacks by determining the words relevance on the basis of frequency. Collectively, these programs enhance the security of the cyber environment by accurately identifying and anticipating attacks, enhancing digital infrastructure, and minimising the risks arising out of hate speech. SVM works with both linear data and non-linear data, whereas Multinomial Naive Bayes employs probability as a method of text classification. TF-IDF is useful in predicting cyber threats, as it quantifies the relevance of terms in terms of their frequency, whereas the Logistic Regression offers a simple approach.

II. EXISTING SYSTEM

The way to locate trends in cyber-security data breach involves the current method that utilizes machine learning techniques.

Grenze ID: 01.GIJET.12.2.238_20

Algorithms such as the logistic regression, decision trees, SVMs, and neural networks are implemented with the use of Django, Scrapy and Beautiful Soup.
These solutions make the implementation of security algorithms to be easily addressed. In the case of web scraping activities, the use of the Scrapy and Beautiful Soup in conjunction with Django and its powerful framework ensures complete data collection. By effectively identifying trends indicative of cyber-security issues with a combination of numerous machine learning algorithms, the system can enhance general detention and response efforts against potential data breaches.
The issue is that the system fails to adapt to emerging cyber threats because of its use of outdated practices.
When real-time information is not taken into consideration, its capability to respond to an impending danger is compromised.
Scalability issues and data management in large datasets make them ineffective to process them effectively. To overcome these disadvantages, there is a need to change to modern technology and methods. To enhance system efficacy and adaptability, it is necessary to have real-time data integration, scalable infrastructure, simplified data administration, and simplified architecture.

III. PROPOSED SYSTEM

The proposed solution applies machine learning to predict data breaches. By combining the support vectors machines, tf-idf, logistic regression, and multinomial naive bayes algorithms in Jupyter notebook and PyCharm, the proposed system will be able to efficiently examine various sets of data to identify any pattern that indicates the possibility of data breaches. Since it has the ability to foresee different attacks, proactive mitigation techniques are guaranteed, enhancing cybersecurity protection and minimising the effects of negative actions on organisational resources.
The proposed system comprises of 4 module

- User login
- Cyber bully prediction
- phishing URL detection
- malicious attack

The user login provides each user with a specific login page that displays the user projected forecasts on the user whether the usage of a word has an influence on the mental health of a person. This will ensure that individuals are able to have safe use of social media sites.
Phishing URL identifies the existence or non-existence of a malicious URL.A malicious attack identifies the infiltration or the absence of an unauthorised source into the network.

IV. ALGORITHM DESCRIPTION

Algorithm used:

- Svm
- Multinomial naive bayes
- Logistic regression

Multinomial Naive Bayes makes use of the Multinomial distribution and the conditional independence assumption to compute probabilities of features.
It employs bag-of-words representation and frequency-based procedure as a means of finding prior and posterior probability. Also, a smoothing technique can enhance the categorisation accuracy of texts by effective sparse data distributions.
In SVM pre-processing, data is TF-IDF vectorised and then model building is done. A train-test split is used to guarantee strong evaluation. A linear kernel is used to evaluate the model, and it is easier to interpret it. Custom input prediction also enhances its usability in a range of situations.
Logistic regression is based on the sigmoid function to do threshold classification and on linear combination of features to estimate probability.
In binary classification problems, regularisation guarantees stability of the model and optimisation method like gradient descent is applied.
TF-IDF (Term Frequency-Inverse Document Frequency) uses text and inverse document frequency as vectors to do the vectorisation. In the case of classification applications, normalisation ensures a similar representation, and this ensures effective processing and interpretation of the textual data.

V. SYSTEM ARCHITECTURE

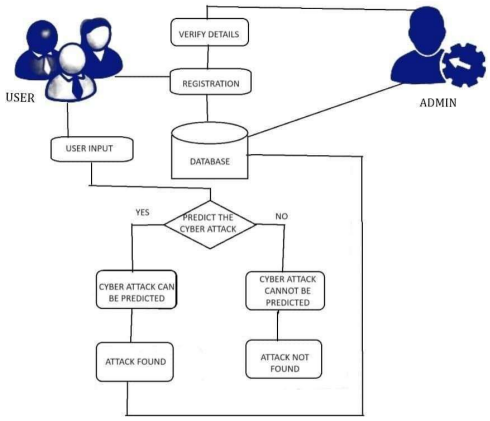

Fig 1. Architecture Diagram

VI. IMPLEMENTATION

A. Phishing URL

The issue of combating phishing continues to be one of the biggest cybersecurity challenges among any business in the entire world. It needs a powerful system that will effectively fight this menace, and the first step towards this is pre-processing of data carefully.

To ensure that the phishing site URLs will be accurate and consistent in subsequent analysis, this step involves purification and homogenisation of the phishing site URLs

- Text tokenisation and stemming algorithms are then applied to reduce the URLs to their root forms and meaningful tokens which allows extraction of useful features.

Fig 2. Phishing URL

- The system predicts the activity of phishing by extracting distinctive features using advanced feature extraction methods, which enhances the predictive capability of the system. Based on the step of the model training, machine learning is employed to learn the labelled phishing and the genuine URLs after the extraction of features, which enables the system to distinguish the two groups successfully.

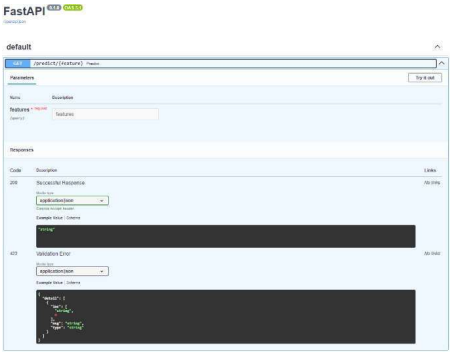

Fig 3. Phishing URL given as input

- It is important that model evaluation will be used to identify whether the trained model will be effective in real-life scenarios by evaluating its performance and generalisation ability. The model is sustained to be easily integrated into the operating environment in case the evaluation results are positive. The algorithm proceeds to the prediction stage in which it identifies potential phishing appropriately after analysing real-time incoming URLs.
- In order to ensure the effectiveness and reliability of the system, the model testing of phishing URLs detection with machine learning algorithms involves several essential steps. To get relevant variables such as domain age, URL length, the presence of HTTPS, and the lexical features, a diversified set of URLs consisting of legal and phishing ones is preprocessed and collected first. This dataset is then divided into training and testing set to attain robustness, typically by means of such techniques as the k-fold cross-validation. The effectiveness of various machine learning algorithms, such as Random Forest, Support Vector Machines, and Gradient Boosting, that are trained on the training set, are evaluated using such metrics as accuracy, precision, recall, F1-score, and ROC-AUC.
- Furthermore, it is also assessed through methods like receiver operating characteristic (ROC) curves and confusion matrices which measure the ability of the model to distinguish between phishing and authentic URLs at different threshold levels. To ensure the model selected is able to make conclusions outside the training data, it is lastly further confirmed using a different dataset.

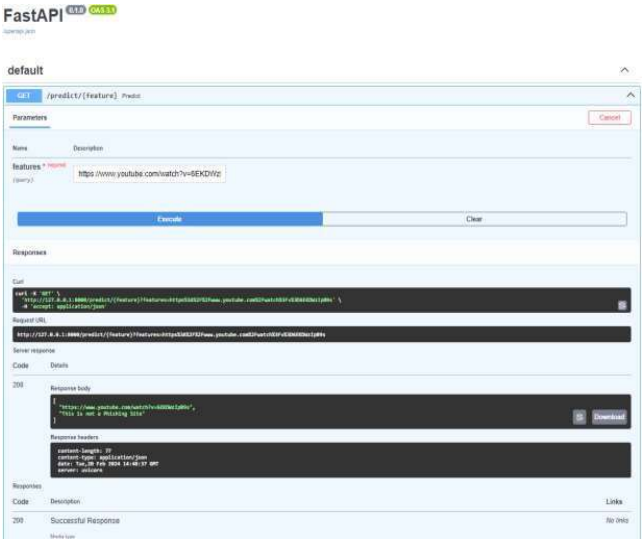

Fig 4. URL is predicted whether Phishing or not

B. Cyber Bully Prediction

An approach which is systematic is necessary in addressing the issue of cyberbullying and the first step is the importation of relevant datasets which contain instances of cyberbullying. The data is then handled carefully through pretreatment procedures to clean and standardise them which ensures uniformity and quality data to be analysed. Categorical variables are transformed into numerical values through feature encoding methods after preprocessing and this facilitates their inclusion in machine learning models.

In cyberbullying prediction, machine learning methods are necessary to identify and curb the detrimental online behaviour. The initial step is to collect a substantial amount of textual data on numerous online sources like social media posts, comments and messages among other forms of communication. This data is then taken through the preprocessing techniques to find relevant content, including the sentiment analysis, linguistic patterns, and contextual data. This data is then divided into training and testing data sets by employing such techniques as stratified sampling to ensure that cases of cyberbullying are evenly represented.

To deduce any pattern that is indicative of cyberbullying behaviour, various machine learning algorithms such as Support Vector Machines, Naive Bayes and Recurrent Neural Networks are trained on the labelled dataset. Performance of the model is measured using metrics such as accuracy, precision, recall, and F1-score, and the results of prediction can be analysed using techniques such as confusion matrices to compare the performance of the model across two or more classes.

Fig 5. Cyber bully prediction

Additionally, such methods as feature relevance.

Step 1

To uncover patterns and associations that are indicative of cyberbullying behaviour, a set of machine learning methods is applied on the preprocessed and encoded data at the model training stage. Once the results of an evaluation are satisfactory, a comprehensive review is done to determine the effectiveness of the model and its generalisation to new data.

The instances of incoming text data undergo preprocessing and vectorisation at the prediction stage before being fed into the trained model to be analyzed.

Step 2

In order to find patterns and associations suggestive of cyberbullying behaviour, the preprocessed and encoded data is fed into a variety of machine learning techniques during the model training phase. To assess the model's effectiveness and capacity for generalisation to new data, a thorough review is carried out. after the evaluation's results are satisfactory.

Before being fed into the trained model for analysis, incoming text data instances go through preprocessing and vectorisation during the prediction phase.

Step 3

The variables of predicting cyberbullying behaviour can be explained by analysis.

Step4

Finally, the prediction of the model provides valuable information on how potential cases of cyberbullying may unfold and as a result, the stakeholders could make timely measures to curb and mitigate the behaviour.

C. User login

One of the major elements of user login in cyberattack prediction is the analysis of the behaviour pattern to identify abnormalities that indicate potential breaches. Machine learning algorithms enhance cybersecurity by detecting suspicious action. Multi-factor authentication and continuous monitoring on the sensitive data and systems ensure that the systems will not be easily compromised by attackers.

Fig 6. given text is predicted whether cyber bully

D. Malicious Attack

With SVM, Multinomial Naive Bayes, TF-IDF and Logistic Regression to predict malicious assaults, the following steps are commonly involved:

Data Collection

Gather the relevant information using such sources as system events, network logs, or other data related to cyber security.

Data Preprocessing

Preprocess and clean the data such as activities to handle missing values, noise elimination, and TF-IDF transformation of text data into numerical values.

Feature Extraction

Take pertinent features—like network traffic patterns, system log entries, or text-based features obtained from TF-IDF—out of the preprocessed data. Model Training: Using the preprocessed data and extracted features, train separate machine learning models for each algorithm (SVM, Multinomial Naive Bayes, TF-IDF, and Logistic Regression).

Model Evaluation

Evaluation metrics such as accuracy, precision, recall, or F1-score are useful to know the extent to which the trained models can predict malicious assaults.

Ensemble or Fusion

As a technique to enhance the global accuracy of prediction, there is an optional strategy of combining the output of multiple models with techniques such as ensemble learning or model fusion

Deployment and Monitoring

Apply the acquired models to predict dangerous attacks either in real time or batch in a manufacturing environment. Monitor the performance of track models over time and update models accordingly to the changing threats

VII. Results and Discussion

Good results were demonstrated in the use of TF-IDF to forecast phishing URLs and cyberbullying. Major features such as “login, password and words of disparagement having high weights and suggestive of phishing behaviour and language of bullying were adequately identified with the assistance of TF-IDF. The model was good in terms of high accuracy, precision, recall and F1-score, with minimal misclassification. The results indicate the effectiveness of TF-IDF in feature representation and selection and provide valuable data concerning cyberbullying and phishing trends.

The SVM which is known to be effective in high-dimensional space demonstrated strong selectivity in distinguishing between phishing and genuine URLs. Like this, Multinomial Naive Bayes applied TF-IDF features and probability-based classification in the prediction of phishing websites. A competitive performance measure, particularly with respect to accuracy and precision, was achieved with Multinomial Naive Bayes, even though its assumptions about the independence of features are very simplistic.

Its probabilistic model delivered some results that were easily interpretable, and would provide insight into variables that are most likely to cause risk in phishing URLs. The logistic regression was effective in identifying significant features that are indicative of phishing behaviour because it can model the probability that a URL is in the phishing class that can be used to predict a threat in advance.

VIII. Conclusion

Overall, with SVM, Multinomial Naive Bayes, and Logistic Regression along with TF-IDF-based feature extraction, the predictive powers of the overall model were improved, providing a complete and sound way of predicting phishing URLs. Further research to enhance the performance of the forecasting can be dedicated to the combination of sophisticated ensemble learning models like Random Forests, Gradient Boosting, AdaBoost, and stacking models, which use a combination of multiple classifiers to achieve better generalization and minimise classification errors. Additionally, one can consider deep learning models that include Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) networks and transformer-based networks that can be automatically trained to detect complex patterns, semantic relationships, and sequential dependencies in URLs and webpage content. These models can learn high-level representations, without the manual feature engineering needed to perform this, and thus improve the accuracy of detection of previously unknown and advanced phishing schemes. Also, the integration of attention systems as well as composite systems combining classic machine training with deep learning may further enhance the power of the system in detecting the existence of implicit anomalies and contextual indicators. Live learning and adaptive models that are constantly updated with new phishing data may also be viewed to ensure that it remains resilient to new attack methods. Also, the feature space can be expanded with several features including domain reputation, information on the use of an SSL certificate, and behavioral features, which can give a more comprehensive picture of malicious activity. With such sophisticated approaches, future systems can be more scalable, resilient, and accurate which will eventually lead to creating more intelligent and active phishing detection systems that would be able to protect users in an ever-evolving cyber threat environment..

References

[1] Predicting criminal activity and cyberattacks by using machine learning algorithms. Balamurali Ramakrishnan, Kanishka M, Surendran R and Aravind Swaminathan.

[2] Cybersecurity Attack Forecasting Deep Learning-based approach November 2020.

[3] Cyber attack Detection Model (CADM) based on Machine Learning. Marzana Akter; Fahima Hossain;

[4] Big data analytics and statistical machine learning on social media, 2019.

[5] S.A knowledge base technique of intrusion detection modelling. M. Matthews, More, A. Joshi, T. Finin, IEEE Symposium on Security and Privacy Workshops, San Francisco, CA, USA, IEEE, 2012, pp. 75-81.