CSE-431

Name: GM Mohaiminuzzaman Apurbo
ID: 20301100
Section: 01

Paper Title

Network Intrusion Detection using Natural Language Processing and Ensemble Machine Learning

Paper Link

https://ieeexplore.ieee.org/document/9308268

1. Summary

1.1 Motivation

In this paper, an intrusion detection system has been proposed, NLPIDS, which is going to be used for converting natural language HTTP requests into vector forms. These vectors will then be used to train various supervised and ensemble-based machine learning models. The primary contribution of this paper is to identify aberrant, potentially malicious, traffic after analyzing natural language-based network traffic using NLP techniques.

1.2 Contribution

The author has added both Machine Learning (ML) and Natural Language Processing (NLP) to the IDS because these techniques work as catalysts in enhancing the detection accuracy of the IDSs. This module has the ability to add the pattern or source of any attack that has not yet been observed to the database of known assaults. It is also discussed how NLP and ML are useful techniques for detecting several types of cyber attacks such as long-mining, phishing attacks, etc. The suggested approach builds vector spaces from text corpuses using natural language processing techniques, which are then used to train machine learning models to identify anomalies. The author chooses two NLP techniques: Word2Vec and Doc2Vec as well as an open-source program: Gensim. The use of logistic regression (LR), support vector machines (SVM), na¨ıve Bayes with Gaussian function (NB), decision tree (DT), and neural networks (NN) have been discussed as well. A workflow has also been shown which consists of data pre-processing phase, natural language processing phase, and ensemble based machine learning

phase. After undergoing the experiments, it has been concluded that Ens NN, Ens DT, and Ens SVM are the three models which provide 100% accuracy in finding the anomalous data containing the highest values of sensitivity, specificity, and F1-score.

1.3 Methodology

The methodology of the research provides an overview of the use of the two significant natural language tools: Word2Vec and Doc2Vec. Three vital phases have also been followed.

- Conversion of HTTP requests to Corpus: A preprocessing tool is used to extract pertinent information from the HTTP requests and translate them into English. For each HTTP request, these terms produce one statement in natural language. Sentences are combined to form paragraphs, which are then saved in documents. A corpus is the collection of all these documents.
- Construction of Vector Space Model: the documents in the corpus are then converted into vector space models using gensim. A single corpus of M documents yields a matrix of M x N, where N is the fixed length of the feature vectors. Later, the combined matrix is split into training which are stored in CSV files.
- Ensemble Machine Learning: it is a two-stage framework where supervised classifiers are used for classifying the data. These individual classifiers' classification outputs are fed into an additional classifier known as the ensemble classifier.

1.4 Conclusion

One of the most effective technologies in fields like speech recognition, sentiment analysis, question-answering, anomaly detection, etc. is natural language processing, or NLP. The evaluation metrics values for different classifier models and the corresponding graph depicts the values for F1-score, accuracy, precision and sensitivity. It is found that the accuracy values for the models vary from 96.34% for LR-based ensemble classifier, Ens LR, to 99.95% for three ensemble

models, Ens NN, Ens DT, and Ens SVM. The best precision value of 99.96% is given by four ensemble models, Ens NB, Ens NN, Ens DT, and Ens SVM. After comparing performance of all the eleven models, the best performing model has been retained to be deployed as part of NLPIDS for real-time intrusion detection. To conclude, it can be said that the addition of the NLP and ensemble ML techniques into IDS have resulted in gaining more accurate experimental results.

2. Limitations

2.1 First Limitation:

- Limited dataset: The study utilizes the HTTP DATASET CSIC 2010 for validation, which may not represent all possible network intrusion scenarios. A broader and more diverse dataset would strengthen the findings and generalize the results.

2.2 Second Limitation:

- Lack of scalability analysis: The scalability of the proposed NLPIDS is not addressed in the paper. It is important to evaluate how the system performs in large-scale network environments with high traffic volumes and diverse data sources.

3. Synthesis

The paper proposes a network intrusion detection system (NLPIDS) that utilizes natural language processing (NLP) and ensemble-based machine learning to detect network intrusions. The system converts natural language HTTP requests into vectors using NLP techniques and trains machine learning models to identify anomalous traffic. The NLPIDS is capable of detecting unknown or zero-day attacks and unseen malicious traffic, unlike traditional intrusion detection techniques. The proposed method is compared to existing methods using the HTTP DATASET CSIC 2010, and it demonstrates better performance with a higher F1-score (0.999) and fewer false alarms (0.007). The document discusses related works in the field, the advantages of using NLP and machine learning, and

provides background information on NLP tools, methodology, experimental setup, evaluation metrics, and results. Overall, the paper presents a novel approach to network intrusion detection that exhibits promising accuracy and a low false alarm rate .