NLP-based Supervised Learning Algorithm for Cyber Insurance Policy Pattern Prediction

B. Subba Reddy[1], V. Bhargavi.[2], S. Samhitha[2], Y. Anjana[2], V. Saivaishnavi[2]

[1,2]Department of Information Technology

[1,2]Malla Reddy Engineering College for Women (A), Maisammaguda, Medchal, Telangana.

Abstract

Cyber insurance has become increasingly important in today's digital landscape, as organizations face the growing threat of cyberattacks and data breaches. Accurate prediction of cyber insurance policy patterns can help insurance companies assess risk, set appropriate premiums, and develop effective coverage strategies. In this study, we propose a methodology that combines TF-IDF (Term Frequency-Inverse Document Frequency) feature extraction and multinomial naive Bayes classifier to predict cyber insurance policy patterns. We leverage the TF-IDF algorithm to represent policy documents as numerical feature vectors, which capture the importance of terms in the documents. The multinomial naive Bayes classifier is then employed to classify the policy patterns based on the extracted features.

Keywords: Cyber insurance, Policy prediction, supervised learning.

1. Introduction

The increased dependency of modern society in digital services has led organizations in significant investments for administrative and technical countermeasures to prevent accidental or malicious cybersecurity incidents. Nonetheless, the realization of modern cyberattacks and cybersecurity incidents that result in severe impacts have made evident that organizational cybersecurity management cannot rely solely on risk mitigation measures; instead, cyber insurance rises as a necessary complement to organizational safeguards. Exemplary cybersecurity attacks that resulted with critical severity include WannaCry and Not Petya in 2017, which affected thousands of companies in multiple regions and industries. Another example is the ransomware attack that affected governmental organizations in the USA (i.e., Departments of Defense, Homeland Security, State, Treasury, Energy and Commerce, as well as several others) [1]. Advanced cyber threats of high severity that prevail today include crypto jacking, malware, supply-chain attacks, ransomware, business email compromise and others [2, 3].

Information security management is widely accepted as a risk-based process. Following risk assessment, organizations can decide how to manage risks by choosing amongst four strategies: risk modification, risk retention, risk avoidance and risk sharing. The latter strategy (a.k.a., risk sharing) pertains that the organization shares the risk with an external party that can most effectively manage the particular risk depending on risk evaluation. Risk sharing can be implemented using insurance to support the consequences of an incident, or by sub-contracting to prevent the risk from materializing. Cyber insurance market and practices are rapidly growing and expected to further develop. Despite the strong motivation that organizations must employ insurance as a cybersecurity strategy for specific threats, as the same time they are reluctant to do so. Given that the cyber insurance domain is going through foundational development, transformation and shaping, in this paper we aimed to investigate and present an overall view of the domain today, including drivers, obstacles, practices and involved processes, status and involved stakeholders. Similar literature reviews [4] have been performed by few researchers, including and. Our work not only updates the insights that these works provide, but it is also complementary to their findings. In [5], the authors also highlight practical challenges, including how to assess cyber risks during the underwriting process and how to calculate and receive appropriate compensation. Our purpose in this article is to bring together the insights given by past literature analysis works and perform an extensive and updated literature review on the current cybersecurity insurance research and practice toward drafting the current landscape and providing insights on future directions. In addition to the approach taken by past literature analyses, in this paper we place special attention to recent industrial survey reports and statistical data, to reveal the most prevailing and up to date information regarding the cyber insurance market. Our findings deriving from the analysis of existing literature highlight the distinguishing characteristics of the cybersecurity insurance market, including the dominance of large clients, the complex and lengthy underwriting process compared to other insurance products and the imbalance between demand and capacity. Our analysis also presents the available types of insurance policies and the typically insurable cybersecurity risks. Nonetheless, insurance policy risk coverage limits are not as clear as in other insurance products; for example, an incident might be detected after some time, which makes challenging for the insured to receive coverage. Finally, our work aggregates information related to the underwriting and claims management process, which can be informative for organizations who wish to consider the option of insurance for cybersecurity. Further, this literature review recognizes research and practical directions for further development of the field toward addressing identified challenges and obstacles.

2. Literature Survey

Kure et al. [6] proposed a novel integrated cyber security risk management (i-CSRM) framework that responds to that challenge by supporting systematic identification of critical assets through the use of a decision support mechanism built on fuzzy set theory, by predicting risk types through machine learning techniques, and by assessing the effectiveness of existing controls. The framework is composed of a language, a process, and it is supported by an automated tool. The paper also reported on the evaluation of our work to a real case study of a critical infrastructure. The results revealed that using the fuzzy set theory in assessing assets' criticality, our work supports stakeholders towards an effective risk management by assessing each asset's criticality. Furthermore, the results have demonstrated the machine learning classifiers’ exemplary performance to predict different risk types including denial of service, cyber espionage and crimeware.

Albasheer et al. [7] reviewed the state-of-the-art cyber-attack prediction based on NIDS Intrusion Alert, its models, and limitations. The taxonomy of intrusion alert correlation (AC) is introduced, which includes similarity-based, statistical-based, knowledge-based, and hybrid-based approaches. Moreover, the classification of alert correlation components was also introduced. Alert Correlation Datasets and future research directions are highlighted. The AC received raw alerts to identify the association between different alerts, linking each alert to its related contextual information and predicting a forthcoming alert/attack. It provides a timely, concise, and high-level view of the network security situation. This review can serve as a benchmark for researchers and industries for Network Intrusion Detection Systems’ future progress and development.

Tsohou et al. [8] examined the relevant literature on cybersecurity insurance, research and practice, in order to draft the current landscape and present the trends. This has led to an increase of cyberattacks, as a direct consequence of the increase of the attack surface but subsequently also led to an increased necessity for the protection of information systems. Toward the protection of information systems, cyber insurance is considered as a strategy for risk management, where necessary. Cyber insurance is emerging as an important tool to protect organizations against cyberattack-related losses.

Zhao et al. [9] proposed CTP-DHGL, a novel Cyber Threat Prediction model based on Dynamic Heterogeneous Graph Learning, to predict the potential cyber threats by investigating public security-related data (e.g., CVE details, ExploitDB). Particularly, we first characterize the interactive relationships among different types of cyber threat objects with a heterogeneous graph. This work then formalized cyber threat prediction as a dynamic link prediction task on the heterogeneous graph and propose an end-to-end dynamic heterogeneous graph embedding method to learn the dynamic evolutionary patterns of the graph. As a result, CTP-DHGL can infer potential link relationships based on the evolving graph embedding sequences learned from previous snapshots to infer stealthy cyber threats. The experimental results on real-world datasets verify that CTP-DHGL outperforms the baseline models in learning the evolutionary patterns of cyber threats and predicting potential cyber risks.

Husák et al. [10] studied the both methods based on discrete models, such as attack graphs, Bayesian networks, and Markov models, and continuous models, such as time series and grey models, are surveyed, compared, and contrasted. This work further discussed machine learning and data mining approaches, that have gained a lot of attention recently and appears promising for such a constantly changing environment, which is cyber security. The survey also focused on the practical usability of the methods and problems related to their evaluation.

Singh et al. [11] first look at the soft spots and threats faced by the insurance companies, and the impacts of these threats. This work finds that both management and technology measures are necessary to tackle the threat. This work then come up with a five-pronged recommendation framework on how insurance companies can strengthen their security infrastructure.

Hwang et al. [12] studied the latent Dirichlet allocation is applied to extract text-document-based technical topics for the symmetrical thesis and patent information to identify security convergence fields and technologies for cyber safety. In addition, it elucidates cyber security convergence fields and technology trends by applying a dynamic topic model and long short-term memory, which are useful for analyzing technological changes and predicting trends. Based on these results, cyber security administrators, system operators, and developers can effectively identify and respond to trends in related technologies to reduce threats, and companies and experts developing cyber security solutions can present a new security approach.

Sarker et al. [13] presented an Intrusion Detection Tree (“IntruDTree”) machine-learning-based security model that first considers the ranking of security features according to their importance and then build a tree-based generalized intrusion detection model based on the selected important features. This model is not only effective in terms of prediction accuracy for unseen test cases but also minimizes the computational complexity of the model by reducing the feature dimensions. Finally, the effectiveness of our IntruDTree model was examined by conducting experiments on cybersecurity datasets and computing the precision, recall, fscore, accuracy, and ROC values to evaluate. This work also compared the outcome results of IntruDTree model with several traditional popular machine learning methods such as the naive Bayes classifier, logistic regression, support vector machines, and k-nearest neighbor, to analyze the effectiveness of the resulting security model.

Lu et al. [14] established a kind of network safety situation forecast model based on Grey Wolf Optimization (GWO) algorithm to optimize support vector machine (SVM) parameters and solves the problem of support vector machine (SVM) parameter optimization. It overcomes the problems of neural network training and local optimization, which makes it more generalized, also effectively improve the prediction effect of SVM. The simulation experiments indicated that this model has improved the accuracy of prediction and shows the general tendency of the network security situation.

Zhang et al. [15] reviewed Artificial Intelligence applications in cyber security areas and the vast literature on applying XAI in many fields including healthcare, financial services, and criminal justice, the surprising fact is that there are currently no survey research articles that concentrate on XAI applications in cyber security. Therefore, the motivation behind the survey is to bridge the research gap by presenting a detailed and up-to-date survey of XAI approaches applicable to issues in the cyber security field...

Mumtaz et al. [16] investigated on the importance of Cyber Security as well as the impact of COVID-19 on cyber security. The dataset (SCI as per the report of the Center for Strategic and International Studies (CSIS)) is divided into two subsets (pre-pandemic SCI and post-pandemic SCI). Data Mining (DM) techniques are used for feature extraction and well know ML classifiers such as Naïve Bayes (NB), Support Vector Machine (SVM), Logistic Regression (LR) and Random Forest (RF) for classification. A centralized classifier approach is used to maintain a single centralized dataset by taking inputs from six continents of the world. The results of the pre-pandemic and post-pandemic datasets are compared and finally conclude this paper with better accuracy and the prediction of which type of SCI can occur in which part of the world. It is concluded that SVM and RF are much better classifiers than others and Asia is predicted to be the most affected continent by SCI.

3. Proposed System

A cyber insurance policy, also known as cyber risk insurance or cyber liability insurance, is a type of insurance coverage designed to protect individuals, businesses, and organizations from financial losses and liabilities associated with cyber-related incidents and data breaches. With the increasing frequency and sophistication of cyber-attacks, cyber insurance has become an important risk management tool for many entities. To predict cyber insurance policy patterns using TF-IDF (Term Frequency-Inverse Document Frequency) and a Multinomial Naïve Bayes classifier, this work following steps:

Data Preprocessing: Gather a dataset of cyber insurance policy documents. Preprocess the data by removing any irrelevant information, such as headers, footers, or special characters. Tokenize the text into individual words or n-grams (sequences of adjacent words). Perform other preprocessing steps like stemming, lemmatization, or removing stop words based on your specific requirements.

TF-IDF Vectorization: Convert the pre-processed text documents into numerical features using the TF-IDF technique. TF-IDF assigns weights to words based on their frequency in a document (TF) and inverse frequency across all documents (IDF). This technique helps capture the importance of words within individual documents relative to the entire corpus.

Splitting the Dataset: Split your dataset into training and testing sets. The training set will be used to train the Multinomial Naïve Bayes classifier, while the testing set will be used to evaluate its performance.

Training the Classifier: Train a Multinomial Naïve Bayes classifier using the TF-IDF vectors from the training set. The Multinomial Naïve Bayes classifier is suitable for text classification tasks as it assumes that the features (TF-IDF values) are generated from a multinomial distribution.

Model Evaluation: Evaluate the trained classifier using the testing set. Calculate metrics such as accuracy, precision, recall, and F1-score to assess the performance of the classifier.

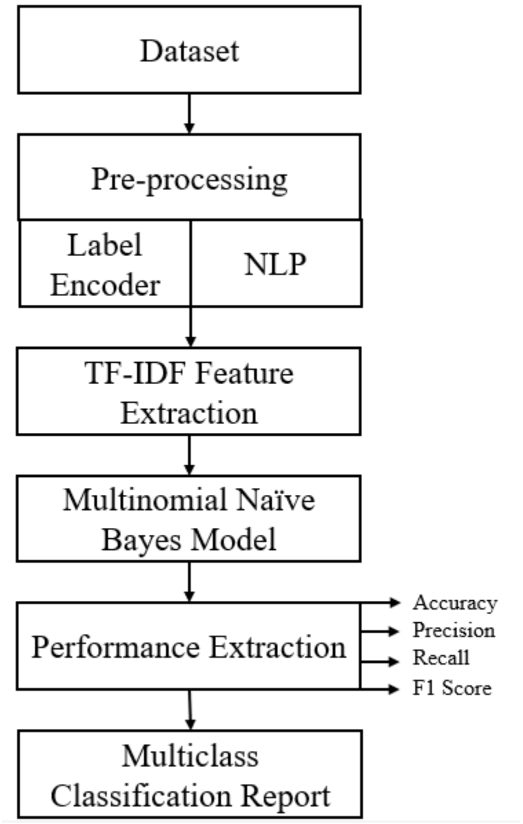


Fig. 1: Block diagram of proposed system.

3.1 Pre-processing

Data pre-processing is a process of preparing the raw data and making it suitable for a machine learning model. It is the first and crucial step while creating a machine learning model.

When creating a project, it is not always a case that we come across the clean and formatted data. And while doing any operation with data, it is mandatory to clean it and put in a formatted way. So, for this, we use data pre-processing task.

Why do we need Data Pre-processing?

A real-world data generally contains noises, missing values, and maybe in an unusable format which cannot be directly used for machine learning models. Data pre-processing is required tasks for cleaning the data and making it suitable for a machine learning model which also increases the accuracy and efficiency of a machine learning model.

- Getting the dataset
- Importing libraries
- Importing datasets
- Finding Missing Data
- Encoding Categorical Data
- Splitting dataset into training and test set
- Feature scaling

3.1.1 Splitting the Dataset into the Training set and Test set

In machine learning data pre-processing, we divide our dataset into a training set and test set. This is one of the crucial steps of data pre-processing as by doing this, we can enhance the performance of our machine learning model.

Supposeif we have given training to our machine learning model by a dataset and we test it by a completely different dataset. Then, it will create difficulties for our model to understand the correlations between the models.

If we train our model very well and its training accuracy is also very high, but we provide a new dataset to it, then it will decrease the performance. So we always try to make a machine learning model which performs well with the training set and also with the test dataset. Here, we can define these datasets as:

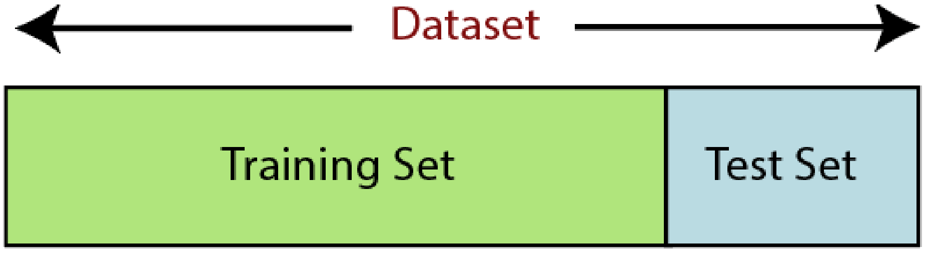


Fig. 2: Dataset splitting.

Training Set: A subset of dataset to train the machine learning model, and we already know the output.

Test set: A subset of dataset to test the machine learning model, and by using the test set, model predicts the output.

3.2 TF-IDF Feature Extraction

TF-IDF, short for Term Frequency-Inverse Document Frequency, is a commonly used technique in NLP to determine the significance of words in a document or corpus. To give some background context, a survey conducted in 2015 showed that 83% of text-based recommender systems in digital libraries use TF-IDF for extracting textual features. That's how popular the technique is. Essentially, it measures the importance of a word by comparing its frequency within a specific document with the frequency to its frequency in the entire corpus. The underlying assumption is that a word that occurs more frequently within a document but rarely in the corpus is particularly important in that document.

3.2.1 Mathematical formula for calculating TF-IDF

TF (Term Frequency) is determined by calculating the frequency of a word in a document and dividing it by the total number of words in the document.

TF = (Number of times the word appears in the document) / (Total number of words in the document)

IDF (Inverse Document Frequency), on the other hand, measures the importance of a word within the corpus as a whole. It is calculated as:

IDF = log((Total number of documents in the corpus) / (Number of documents containing the word))

3.3 Multinominal Naïve Bayes

What is the Multinomial Naive Bayes algorithm?

Multinomial Naive Bayes algorithm is a probabilistic learning method that is mostly used in Natural Language Processing (NLP). The algorithm is based on the Bayes theorem and predicts the tag of a text such as a piece of email or newspaper article. It calculates the probability of each tag for a given sample and then gives the tag with the highest probability as output.

Naive Bayes classifier is a collection of many algorithms where all the algorithms share one common principle, and that is each feature being classified is not related to any other feature. The presence or absence of a feature does not affect the presence or absence of the other feature.

How Multinomial Naive Bayes works?

Naive Bayes is a powerful algorithm that is used for text data analysis and with problems with multiple classes. To understand Naive Bayes theorem's working, it is important to understand the Bayes theorem concept first as it is based on the latter.

Bayes theorem, formulated by Thomas Bayes, calculates the probability of an event occurring based on the prior knowledge of conditions related to an event. It is based on the following formula:

$P(A|B) = P(A) * P(B|A)/P(B)$

Where we are calculating the probability of class A when predictor B is already provided.

$P(B)$ = prior probability of B

$P(A)$ = prior probability of class A

$P(B|A)$ = occurrence of predictor B given class A probability

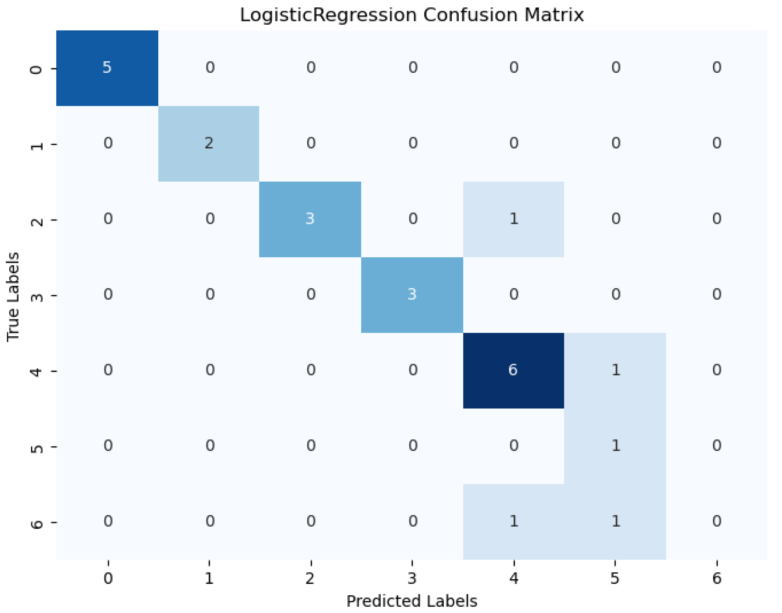
This formula helps in calculating the probability of the tags in the text.

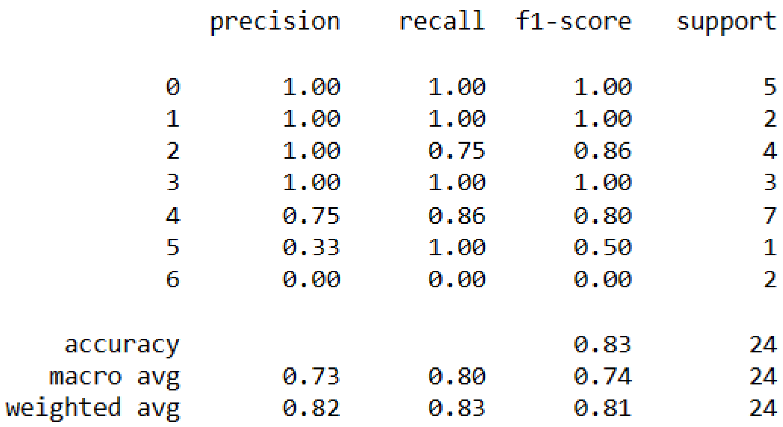
3.4 Advantages of proposed system

The Naive Bayes algorithm has the following advantages:

- It is easy to implement as you only must calculate probability.
- You can use this algorithm on both continuous and discrete data.
- It is simple and can be used for predicting real-time applications.
- It is highly scalable and can easily handle large datasets.

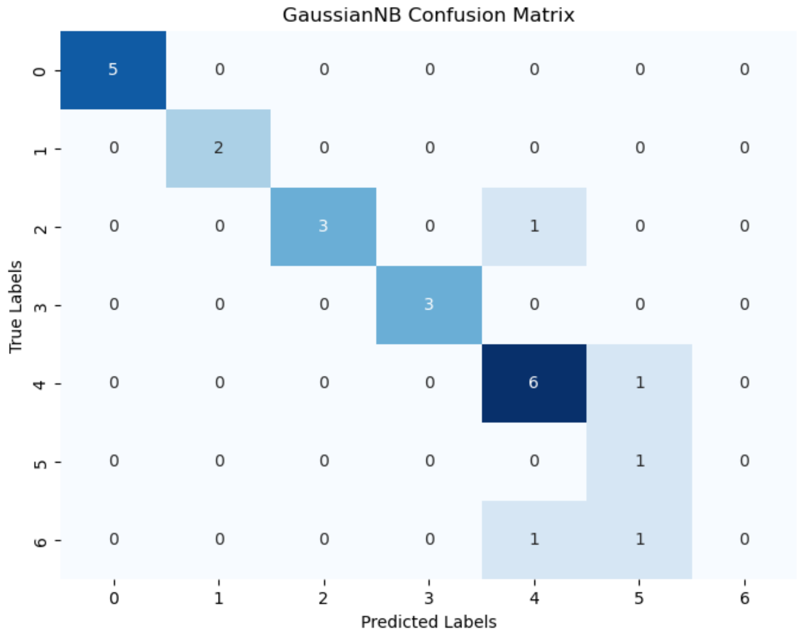
4. Results and Discussion

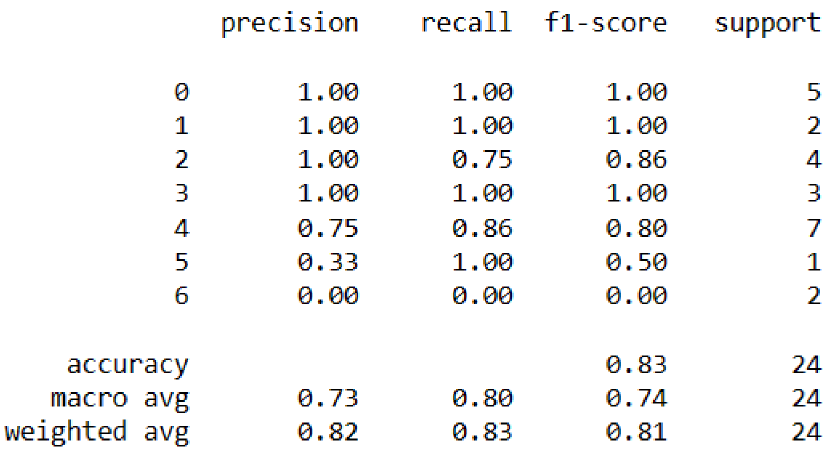



```
              precision    recall  f1-score   support

           0       1.00      1.00      1.00         5
           1       1.00      1.00      1.00         2
           2       1.00      0.75      0.86         4
           3       1.00      1.00      1.00         3
           4       0.75      0.86      0.80         7
           5       0.33      1.00      0.50         1
           6       0.00      0.00      0.00         2

    accuracy                           0.83        24
   macro avg       0.73      0.80      0.74        24
weighted avg       0.82      0.83      0.81        24
```





```
              precision    recall  f1-score   support

           0       1.00      1.00      1.00         5
           1       1.00      1.00      1.00         2
           2       1.00      0.75      0.86         4
           3       1.00      1.00      1.00         3
           4       0.75      0.86      0.80         7
           5       0.33      1.00      0.50         1
           6       0.00      0.00      0.00         2

    accuracy                           0.83        24
   macro avg       0.73      0.80      0.74        24
weighted avg       0.82      0.83      0.81        24
```

5. Conclusion and Future Scope

The results of our study demonstrate the effectiveness of the proposed approach in predicting cyber insurance policy patterns. By using TF-IDF for feature extraction and the multinomial naive Bayes classifier for classification, we achieved high accuracy in predicting the patterns. This indicates that the combination of these techniques can be a valuable tool for insurance companies in understanding and predicting policy trends in the cyber insurance domain.

Future Scope

Although our approach shows promising results, there are several avenues for future research and improvement. Firstly, the incorporation of more advanced machine learning algorithms and techniques could potentially enhance the accuracy and robustness of policy pattern prediction. Additionally, exploring the integration of other textual analysis methods, such as word embeddings or deep learning models, may provide further insights into policy patterns and improve prediction performance. Furthermore, expanding the dataset used for training and testing could lead to more comprehensive and generalizable results. Lastly, investigating the impact of external factors, such as regulatory changes or technological advancements, on policy patterns could contribute to a more holistic understanding of the cyber insurance industry.

References

[1] Gallagher Cyber Insurance Market Conditions Report: Guidance as the cyber insurance market continues to harden. https://www.ajg.com/us/news-and-insights/2021/jan/2021-cyber-insurance-market-report/ (2021), [Online; accessed 18-July-2022]

[2] ENISA Threat Landscape 2021. https://www.enisa.europa.eu/publications/enisa-threat-landscape-2021 (2021), [Online; accessed 18-July-2022]Return to ref 4 in article

[3] Report, H.: Don't let cyber be a game of chance. https://www.hiscoxgroup.com/sites/group/files/documents/2021-04/Hiscox%20Cyber%20Readiness%20Report%202021.pdf (2021), [Online; accessed 18-July-2022]

[4] Aziz, B.: Others A systematic literature review of cyber insurance challenges. In: 2020 International Conference on Information Technology Systems and Innovation (ICITSI), pp. 357–363 (2020)

[5] Dambra, S., Bilge, L., Balzarotti, D.: SoK: Cyber insurance? technical challenges and a system security roadmap. In: 2020 IEEE Symposium On Security And Privacy (SP), pp. 1367–1383 (2020)

[6] Kure, H.I., Islam, S. & Mouratidis, H. An integrated cyber security risk management framework and risk predication for the critical infrastructure protection. Neural Comput & Applic 34, 15241–15271 (2022). https://doi.org/10.1007/s00521-022-06959-2

[7] Albasheer H, Md Siraj M, Mubarakali A, Elsier Tayfour O, Salih S, Hamdan M, Khan S, Zainal A, Kamarudeen S. Cyber-Attack Prediction Based on Network Intrusion Detection Systems for Alert Correlation Techniques: A Survey. Sensors. 2022; 22(4):1494. https://doi.org/10.3390/s22041494

[8] Tsohou, A., Diamantopoulou, V., Gritzalis, S. et al. Cyber insurance: state of the art, trends and future directions. Int. J. Inf. Secur. 22, 737–748 (2023). https://doi.org/10.1007/s10207-023-00660-8

[9] J. Zhao, M. Shao, H. Wang, X. Yu, B. Li, X. Liu, Cyber threat prediction using dynamic heterogeneous graph learning, Knowledge-Based Systems, Volume 240, 2022, 108086, ISSN 0950-7051,https://doi.org/10.1016/j.knosys.2021.108086.

[10] M. Husák, J. Komárková, E. Bou-Harb and P. Čeleda, "Survey of Attack Projection, Prediction, and Forecasting in Cyber Security," in IEEE Communications Surveys & Tutorials, vol. 21, no. 1, pp. 640-660, Firstquarter 2019, doi: 10.1109/COMST.2018.2871866.

[11] Singh, A., Akhilesh, K.B. (2020). The Insurance Industry—Cyber Security in the Hyper-Connected Age. In: Akhilesh, K., Möller, D. (eds) Smart Technologies. Springer, Singapore. https://doi.org/10.1007/978-981-13-7139-4_16

[12] Hwang S-Y, Shin D-J, Kim J-J. Systematic Review on Identification and Prediction of Deep Learning-Based Cyber Security Technology and Convergence Fields. Symmetry. 2022; 14(4):683. https://doi.org/10.3390/sym14040683

[13] Sarker IH, Abushark YB, Alsolami F, Khan AI. IntruDTree: A Machine Learning Based Cyber Security Intrusion Detection Model. Symmetry. 2020; 12(5):754. https://doi.org/10.3390/sym12050754

[14] Lu, H., Zhang, G., Shen, Y. (2020). Cyber Security Situation Prediction Model Based on GWO-SVM. In: Barolli, L., Xhafa, F., Hussain, O. (eds) Innovative Mobile and Internet Services in Ubiquitous Computing . IMIS 2019. Advances in Intelligent Systems and Computing, vol 994. Springer, Cham. https://doi.org/10.1007/978-3-030-22263-5_16

[15] Z. Zhang, H. A. Hamadi, E. Damiani, C. Y. Yeun and F. Taher, "Explainable Artificial Intelligence Applications in Cyber Security: State-of-the-Art in Research," in IEEE Access, vol. 10, pp. 93104-93139, 2022, doi: 10.1109/ACCESS.2022.3204051.

[16] G. Mumtaz et al., "Classification and Prediction of Significant Cyber Incidents (SCI) using Data Mining and Machine Learning (DM-ML)," in IEEE Access, doi: 10.1109/ACCESS.2023.3249663.