INL/EXP-21-63785-Revision-0

Cybersecurity Training Curriculum Analysis through the application of Machine Learning Text Classification and Natural Language Processing

August 2021

Changing the World's Energy Future

Kevin Tian, Bengisu Cuneyit, Gary Martin Deckard

Idaho National Laboratory

INL is a U.S. Department of Energy National Laboratory operated by Battelle Energy Alliance, LLC

INL/EXP-21-63785-Revision-0

Cybersecurity Training Curriculum Analysis through the application of Machine Learning Text Classification and Natural Language Processing

Kevin Tian, Bengisu Cuneyit, Gary Martin Deckard

August 2021

Idaho National Laboratory
Idaho Falls, Idaho 83415

http://www.inl.gov

Prepared for the
U.S. Department of Energy
Under DOE Idaho Operations Office
Contract DE-AC07-05ID14517

Cybersecurity Training Curriculum Analysis through the application of Machine Learning Text Classification and Natural Language Processing

Tian, Kevin[1], Cuneyit, Bengisu[1], Deckard, Gary M.[2]

Purdue University[1], Idaho National Laboratory (Mentor)[2]

Introduction: In the domain of cybersecurity, there are a wide variety of education and training courses offered by a variety of training providers (commercial vendors, governmental, and academic institutions). Unfortunately, a comprehensive cross-provider mapping of the body of course offerings does not currently exist. The inability to compare training courses by topic or cybersecurity work-role and the level of difficulty (competency level) affects all organizations in identifying and selecting potential training opportunities for their personnel performing cybersecurity duties.

Our Project: We utilized Machine Learning (ML) Text Classification methodologies as an effort to organize an accurate and thorough catalog of course offerings that establishes competency level equivalencies across providers. We theorized that it was possible to align the course offerings to an accepted framework of industry work-roles such as the National Institute for Standards & Technology (NIST) National Initiative for Cybersecurity Education (NICE) [2] by work-role and competency level.

Methodology: Training offerings across the cybersecurity field have associated course descriptions with specific text-based attributes for each course. Correspondingly, the NICE Framework and ISU-INL research driven ICS work-roles contain text-based descriptions and verbiage for their associated Tasks, Knowledge, and Skills (TKSs).

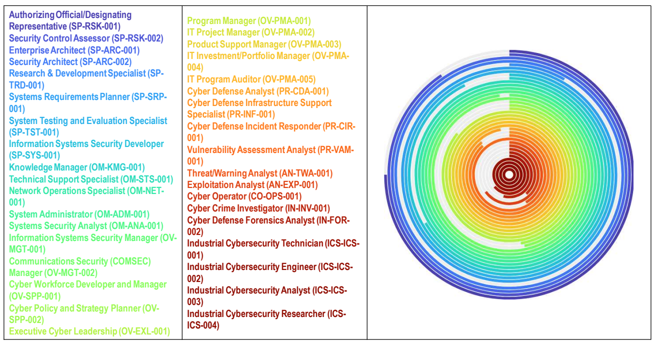

Table 1: Roles performed by respondents (graph showing percentage within role) [1]

Category	Phase I	Phase II
Data Source:	Jung Repository & CYBER-CHAMP catalog	NICE roles reference spreadsheet
Classification Algorithm	• Multinomial Naïve Bayes (NB) • Support Vector Machines (SVM)	• Cosine similarity • Euclidian Distance
Categories	Novice, Fundamentals, Intermediate, Advanced, Expert	53 NICE Work-Roles, ICS Work-Roles

Table 2: Breakdown of Phase I and Phase II characteristics

Data processing

1. (Phase I): Raw course data was aggregated into one cleaned text paragraph for each row in the repository (minus the 'Course Competency' column) split into a python tuple. The first entry being the raw text and the second entry the course competency.
   (Phase II): NIST role descriptions aggregated into one raw text paragraph then cleaned.
2. Cleaned text was vectorized into a Term Frequency – Inverse Document Frequency vector (TF-IDF).
3. Classification algorithm applied to vectorized output

Our Data: The data used in this project primarily comes from **Randall Jung's repository of Cybersecurity courses** that were included with his master's thesis. These courses were labeled by the shown categories and more.

Vendor	Course Number	Course Title	Course Description	Course Duration	Competency Level	Bloom's Level
ISA	FG02	Mathematics for Instrumentation Technicians	This course is specifically designed for the instrument technician who may be struggling with mathematical computations or those who need a basic refresher….	4 days	Fundamentals	Remember

Table 3: Abbreviated table of training data from Jung's repository [1]

Results: Using optimized hyperparameters for SVM within our model yielded a high preliminary accuracy of 88%, while using Naïve Bayes saw a marginal improvement from 47.05% to 66.83% accuracy. Accuracy for evaluation of our model was calculated using True Positive (TP), True Negative (NP), False Positive (FP), and False Negative (FN) using the formula below and accuracy after optimization is shown in table 4.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

(Optimized) SVM Accuracy	(Optimized) Naïve Bayes Accuracy
88.08 %	66.83 %

Table 4: Accuracy for Naive Bayes and Support Vector Machine classifiers after optimization

Figure 1 describes the accuracy of our models' predictions using SVM. The intersection of a row and column in a confusion matrix represents the percentage correctly classified between labels . High values along the diagonal indicate correct classification, and high values on the non-diagonal entries signal misclassification.

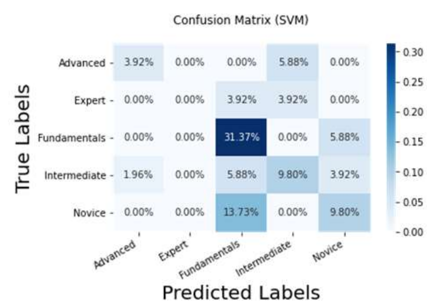

Figure 1: Confusion matrix for SVM

Upon further analysis of SVM's confusion matrix in figure 1, we can see that the model is having a difficult time distinguishing between the Fundamentals & Novice, Fundamentals & Intermediate, and isn't able to categorize Expert at all.

Conclusion: Through the combination of Phase I and Phase II approaches, our model pipeline can receive general characteristics of a Cybersecurity course and successfully categorize the course into the proper competency level and return the associated NICE work roles with cosine similarity serving as a proxy for percent match for relevancy.

Input

"Is a high-level introductory course designed to expose participants to the challenges and frameworks used in implementing and sustaining a cyber security program at a nuclear and/or radiological facility."

Output

```
In [3]: b_script.analyze(course)
Out[3]: {'compLevel': 'Novice',
         'roles': ['ovmgt001 : 60.22%',
         'prcda001 : 51.55%',
         'sprsk001 : 50.87%',
```

While the results are preliminary due to a limited dataset, our approach suggests an automated solution to a manual problem that currently exists within the Cybersecurity space can exist and should be researched further.

Acknowledgements: We would like to thank Idaho National Labs for providing us with this opportunity, our mentor Gary M. Deckard, and all those who have supported this research through feedback and peer review.

References:

[1] Jung, R. (2020, April). Challenges for the General Schedule 2210 Series. *Unpublished*. USAF Air War College.

[2] Petersen, R., Santos, D., Smith, M., & Witte, G. (2020). *Workforce Framework for Cybersecurity (NICE Framework).* National Institute of Standards and Technology.