CyberReco: Cybersecurity Workforce Readiness Recommender System

Ramoni O. Lasisi
Department of Computer and Information Sciences
Virginia Military Institute, USA
LasisiRO@vmi.edu

Abstract

Using the United States National Initiative for Cybersecurity Education framework as a guide, we propose *CyberReco - Cybersecurity Workforce Readiness Recommender System*, an AI/ML model that attempts to address readiness gaps needed to prepare users lacking in some cybersecurity knowledge, skills, and activities to be workforce-ready. *CyberReco* is built using natural language processing. We present a hierarchical-based framework that is composed of four components including, *text preparation and normalization, keywords extraction and processing, similarity scores and skills computation, and a recommender component*.

Introduction

Our day to day quality of life is affected by various threats and attacks in the cyber space. This includes, among others, network disruption, loss of vital information, and inability to guarantee privacy control (Dipankar et al. 2010; LI 2018; Lieskovan and Hajný 2021). The emergence of new cyber threats and the continuous rise in cyber attacks on key cyber-physical infrastructure suggest the need to increase the number of experts within the realm of cybersecurity (Lieskovan and Hajný 2021; Lasisi et al. 2022). (Dawson and Thomson 2018) have identify the existence of readiness gap for future cyber workforce development, especially in the non-technical KSAs and social attributes such as values, team organization, and civic duty. Furthermore, (Justice et al. 2022) stated that "currently, there are 464,420 unfilled cybersecurity jobs and the nature of the growth of these jobs suggests additional skills that are not being addressed in degrees cybersecurity programs."

In an attempt to address the workforce readiness gap problem identified above, this paper uses the National Initiative for Cybersecurity Education (NICE) framework, and propose a recommender system, tagged, *CyberReco - Cybersecurity Workforce Readiness Recommender System*. *CyberReco* is built on a core of artificial intelligence and machine learning module using natural language processing. *CyberReco* is able to recommend relevant contents to users based on cybersecurity job description of interest. Datasets for the system were scraped from the National Initiative for Cybersecurity Careers and Studies web pages. The scraped data were collected into two datasets named, nice_framework_work_roles.csv and nice.json.

The *CyberReco* System Design

CyberReco is composed of four components: *text preparation and normalization, keywords extraction and processing, similarity scores and skills computation,* and the *recommender component*. Each component consists of at least a sub component (or module). Figure 1 shows the block diagram for the *CyberReco* recommender system.

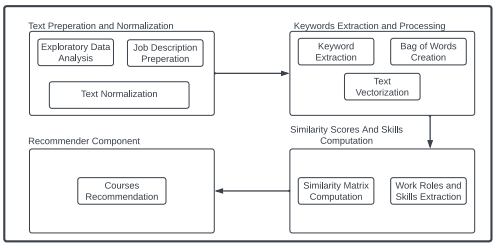

Figure 1: *CyberReco* recommender system block diagram

Text Preparation and Normalization: Textual data is unstructured and contains unwanted items such as non-ascii characters, punctuation, html tags, and links. This component prepares, preprocesses, and transforms raw textual data into clean form that machine learning algorithms can accept as input. There are three modules in this component as described below.

- *Exploratory Data Analysis* - this module loads the two datasets (nice_framework_work_roles.csv and nice.json) into dataframes, drops irrelevant columns from the nice_framework_work_roles.csv dataset, and then separates the KSA dictionary in the nice.json dataset into their respective ids and descriptions.
- *Text preprocessing* - this module takes a text such as the work role or job description and completes preprocessing of the text by removing any html tag, expands contractions, for example the word "didn't" will be changed to did not, "doesn't" will be changed to does not, etc, removes unnecessary numbers, and then breaks down the text into tokens.

- *Text normalization* - this module prepares the relevant columns in the nice_framework_work_roles.csv dataset, including work_role_description, work_role_abilities, work_role_knowledge, work_role_skills, work_role_tasks for modeling. This is done through removing non ascii characters, conversion of the text to lowercase, removal of punctuation, removal of stop words, and completing lemmatization of the texts from each of the columns to their roots.

Below provides an example of a raw textual description of the Security Control Assessor work role and the transformed text after being processed by the text preparation and normalization component.

Original text: *Conducts independent comprehensive assessments of the management, operational, and technical security controls and control enhancements employed within or inherited by an information technology (IT) system to determine the overall effectiveness of the controls (as defined in NIST SP 800-37).*

Transformed text: *conduct independent comprehensive assessments management operational technical security control control enhancements employ within inherit information technology system determine overall effectiveness control define nist sp*

Keywords Extraction and Processing: The transformed text from the previous component remains unstructured as seen above. This component conducts features and keywords extraction and also converts unstructured textual data to structured format (such as table-like) that can be fed into ML algorithms. There are three modules in this component as described below.

- *Keywords Extraction* - this module extracts keywords from a transformed textual data, such as the work role or job description. We employ the well-known Rapid Keyword Extraction (RAKE) to complete this function. Below gives a list of the keywords extracted by this module when given the above transformed text as input.

  [conduct, independent, comprehensive, assessments, management, operational, technical, security, control, enhancements, employ, within, inherit, information, technology, system, determine, overall, effectiveness, define, nist, sp]
- *Bag of Words Creation* - this module creates a new feature referred to as a 'bag of words.' The 'bag of words' is a combination of the transformed texts from the columns of the nice_framework_work_roles.csv dataset and the keywords extracted from the corresponding work role title for each row of the dataset.
- *Text Vectorization* - the intermediate result of the 'bag of words' consists of only text. To fit a machine learning model which involves computing similarity scores between texts, we need inputs that are numeric. So we convert the 'bag of words' feature to numbers by creating set of vectors containing the frequency of occurrence of each word in the 'bag of words.'

Similarity Scores and Skills Computation

- *Similarity scores computation* - this module compares two work role descriptions and computes their similarity score using the pairwise cosine similarity score. The higher the similarity score between the two work role descriptions, the more similar the two work roles are. In computing the cosine similarity scores we used the CountVectorizer and TfidVectorizer to convert texts of each of the descriptions to numerical values before fitting transforms.
- *Work roles and skills extraction* - using an identified work role from the NICE framework, this module computes other work roles that are similar to the identified work role based on top similarity scores. For example, given a work role such as 'Technical Support Specialist,' it returns the top five work roles that are similar to 'Technical Support Specialist.' This module also identifies skills that are needed for a given work role. The common skills across the work roles are extracted using either set intersection i.e, skills common to similar work roles, or frequency count i.e., skills occurring most across similar work roles.

Recommender Component The only module in this component uses the similarity skills extracted from either the set intersection or frequency count to provide suggested courses that potentially satisfy identified work roles skills. These skills correspond to deficiencies that users need to develop or improve on. The process begins by a user supplying description of a cybersecurity job role of interest to *CyberReco*. The texts of the job description and those of the work roles in the nice_framework_work_roles.csv dataset undergo text preprocessing and normalization. The outcomes of this stage are normalized texts that are ready for similarity computations. The NICE framework work roles that are similar to the specified normalized job description are extracted in the subsequent stage. Using similarity scores computed for textual data from normalized job description and work roles, the work roles with top similarity scores are extracted.

The next step involves extraction of skills. Top skills that are representative of the required skills needed for the work roles are extracted. These skills are then used to provide suggested recommendation of contents for the user. Suggested contents are based on three levels: novice or entry level, emerging or intermediate level, and expert or advanced level.

Conclusions

Using the U.S. National Institute of Standards and Technology's National Initiative for Cybersecurity Education framework, we propose *CyberReco - Cybersecurity Workforce Readiness Recommender System*, an AI/ML model that attempts to address the readiness gaps needed to prepare users lacking in some cybersecurity skills and competencies to be workforce-ready. *CyberReco* is founded on three functionalities of web scraping, NLP, and a recommender system.

Acknowledgments

Supported in part by the Commonwealth Cyber Initiative and the National Security Agency under Grant/Cooperative Agreement Number H98230-21-l-0167.

References

[Dawson and Thomson 2018] Dawson, J., and Thomson, R. 2018. The future cybersecurity workforce: going beyond technical skills for successful cyber performance. *Frontiers in Psychology*.

[Dipankar et al. 2010] Dipankar, S. S.; Vivek, D.; Roy, S. S.; Ellis, C.; and Wu, Q. 2010. A survey of game theory as applied to network security. In *Proceedings of the 43rd Hawaii International Conference on Systems Sciences*.

[Justice et al. 2022] Justice, C.; Sample, C.; Loo, S. M.; Ball, A.; and Hampton1, C. 2022. Future needs of the cybersecurity workforce. In *Proceedings of the 17th International Conference on Information Warfare and Security*.

[Lasisi et al. 2022] Lasisi, R. O.; Menia, M.; Farr, Z.; and Jones, C. 2022. Exploration of AI-enabled contents for undergraduate cyber security programs. In *The 35th International FLAIRS Conference*.

[LI 2018] LI, J. 2018. Cyber security meets artificial intelligence: a survey. *Frontiers of Information Technology & Electronic Engineering*.

[Lieskovan and Hajný 2021] Lieskovan, T., and Hajný, J. 2021. Building open source cyber range to teach cyber security. In *ARES 2021: The 16th International Conference on Availability, Reliability and Security*.