BRIDGING THE CYBERSECURITY WORKFORCE GAP:

EVALUATING AI-DRIVEN MODELS FOR CYBERSECURITY CURRICULUM DEVELOPMENT

Aysun Karamustafaoglu

A Capstone Project Submitted to the

University of North Carolina Wilmington in Partial Fulfillment

Of the Requirements for the Degree of

Master of Science

Department of Computer Science

Congdon School of Supply Chain, Business Analytics, and Information Systems

University of North Carolina Wilmington

2023

Advisory Committee

Ulku Clark , Ph.D

Chair

Geoffrey M. Stoker , Ph.D

Committee Member 1

Ron Vetter , Ph.D

Committee Member 2

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

This research presents a comprehensive comparative analysis of two advanced Natural Language Processing (NLP) models, RoBERTa (Robustly Optimized BERT Pretraining Approach) and SecureBERT, focusing on their application in mapping course descriptions from the University of North Carolina Wilmington's (UNCW) BS Cybersecurity Program to in-demand cybersecurity skills. Selected for their distinct characteristics—RoBERTa's general linguistic capabilities and SecureBERT's specialization in cybersecurity text analysis—these models were evaluated for their effectiveness in aligning academic content with industry needs. The study involved a detailed technical evaluation of each model's ability to interpret and correlate course descriptions with relevant skill sets. This was complemented by a survey among instructors to assess the models' practical applicability in an educational context. The research aimed to determine which model more accurately and comprehensively aligns skills with course content, thus addressing the gap between academic curricula and industry requirements in cybersecurity education. Additionally, the findings incorporate instructor perspectives on the adoption of AI-driven tools, underscoring the importance of further research and development, as well as the need for a continuous dialogue with educators. These insights significantly contribute to the discourse on employing AI-driven tools in curriculum development within the field of cybersecurity education, highlighting their potential to revolutionize the alignment of academic offerings with the evolving demands of the cybersecurity job market. The study's outcomes also entail a pragmatic mapping of the UNCW BS Cybersecurity Program's curriculum to the identified skills, proposing a direct correlation between educational pathways and professional competencies required in the cybersecurity field.

1. INTRODUCTION

In the era of rapid technological advancement, cybersecurity has become a pivotal concern for organizations and individuals alike. This evolution necessitates that cybersecurity education not only keeps pace with industry demands but also aligns effectively with them. A key aspect of achieving this alignment involves the application of Natural Language Processing (NLP) models, particularly in analyzing and interpreting cybersecurity educational content.

In this context, AI-driven models such as RoBERTa and SecureBERT are employed for their advanced text analysis capabilities. These models are utilized to conduct a detailed comparison by mapping the text of course descriptions to the text of in-demand cybersecurity skills. This comparative analysis aims to assess how effectively current cybersecurity courses are preparing students with the skills required in the job market. The process involves a rigorous text analysis where the models examine and correlate the terminologies and concepts from academic course descriptions with those used in industry skill definitions. This assessment helps in determining the relevance and alignment of academic content with industry needs, providing insights into the effectiveness of current cybersecurity education in addressing market demands.

According to the International Information System Security Certification Consortium $(ISC)^2$ Annual Cybersecurity Workforce Study 2023, the global cybersecurity workforce has grown by 8.7% in the past year, with the United States witnessing an 11% increase. Despite this growth, the gap between the number of workers needed and those available has also expanded, underscoring persistent staffing shortages and skills gaps. Notably, 67% of organizations report a shortage of cybersecurity staff, and 92% acknowledge skills gaps, particularly in areas such as cloud computing security, AI/ML, and Zero Trust implementation. Furthermore, for the first time, AI/ML skills have surged to the top five in-demand skills, signifying a pivotal shift in the skills landscape.

Given these emerging trends and the expanding skills gap, this research aims to directly address the pressing need for aligning cybersecurity education with industry demands. By evaluating the effectiveness of RoBERTa and SecureBERT in mapping course content to in-demand skills, this study seeks to provide insights on optimizing cybersecurity curricula to better prepare graduates for the challenges of the current cybersecurity landscape.

This study addresses the challenge of mapping course descriptions from the University of North Carolina Wilmington's BS Cybersecurity Program to these in-demand cybersecurity skills. Given the critical nature of this task, it is essential to evaluate and compare the effectiveness of different AI-driven models in accurately performing this mapping. The primary aim of this research is to conduct a comparative analysis of RoBERTa and SecureBERT to determine which model provides a more accurate and comprehensive alignment of skills with course content.

The methodology involves a detailed technical evaluation of each model, supplemented by an expert survey to gauge practical applicability in an educational setting. Through this approach, we aim to address the gap in understanding the capabilities of these AI-Driven models in the context of cybersecurity education, particularly in light of the identified skills gaps and the increasing criticality of AI/ML competencies.

The findings of this study are expected to contribute significantly to the fields of AI-driven models in education and cybersecurity curriculum development. This paper is structured to first detail the methodologies employed, followed by a presentation of the comparative analysis, the survey results, and concluding with the implications of our findings for future curriculum development in cybersecurity education. By aligning academic offerings with the rapidly evolving demands of the cybersecurity job market, this research seeks to bridge the workforce gap and enhance the readiness of graduates to meet the challenges of this dynamic field. Moreover, this investigation will culminate in a practical application, offering a detailed mapping of UNCW's BS Cybersecurity Program courses to the in-demand skills, thus providing actionable insights for curriculum enhancement and student guidance.

2. BACKGROUND

Understanding the broader context of cybersecurity standards and educational requirements is crucial for this study. This background section delves into key organizations and standards that shape the cybersecurity landscape. By exploring the roles of NIST, CompTIA, INCITS, CyberSeek, and ISC^2, we establish the foundational definitions and guidelines essential for cybersecurity education and workforce development. Additionally, we introduce the concept of Large Language Models (LLMs), specifically RoBERTa and SecureBERT, which are central to our comparative analysis. Their unique capabilities in text analysis make them ideal for assessing the alignment of academic content with industry-defined skills, a critical aspect of this research.

2.1. Standards and Certification Authorities Referenced

2.1.1. CyberSeek

This collaborative initiative offers vital insights into the U.S. cybersecurity job market, aiding in the alignment of educational programs with industry needs. CyberSeek's contributions are integral to this study, particularly for understanding current industry demands for cybersecurity skills and workforce development.

2.1.2. International Information System Security Certification Consortium (ISC)²

The ISC^2 Cybersecurity Workforce Study provides essential insights into global workforce trends and challenges in cybersecurity. This study leverages these insights to inform the understanding of the skills gap and workforce needs, thereby providing a relevant context for the model's evaluation of course and skill alignments.

2.1.3. National Institute of Standards and Technology (NIST)

NIST's role in developing cybersecurity standards and definitions is pivotal. Its guidelines provide the foundational basis for the cybersecurity concepts and skills assessed in this text analysis, ensuring alignment with industry standards and practices.

2.1.4. Computing Technology Industry Association (CompTIA)

CompTIA's contribution to setting industry standards and skill requirements in cybersecurity is noteworthy. Their resources are utilized to inform the understanding of the current skills landscape, crucial for assessing the efficacy of the AI-driven models used in this study for mapping course descriptions to industry-relevant skills.

2.1.5. InterNational Committee for Information Technology Standards (INCITS)

INCITS' contributions to cybersecurity standards significantly inform this study. The committee's work in defining skills and knowledge in the field is central to outlining the competencies assessed through RoBERTa and SecureBERT in mapping cybersecurity course content to industry skills.

2.2. Large Language Models (LLMs)

Large Language Models represent a significant advancement in artificial intelligence, particularly in natural language processing. These models, based on sophisticated neural network architectures such as transformers, are trained on extensive datasets, enabling them to capture complex language patterns. In the context of this research, LLMs are crucial for analyzing and interpreting the nuanced language of cybersecurity education and industry skill requirements.

2.2.1. RoBERTa (Robustly Optimized BERT Pretraining Approach)

RoBERTa builds on the foundations set by BERT (Bidirectional Encoder Representations from Transformers), a transformative model introduced by Google in 2018 that revolutionized natural language processing (NLP). RoBERTa, developed by Facebook AI in 2019, to enhance natural language understanding. It employs a broad corpus for training, including diverse data sets such as CC-News, facilitating effective processing of complex language structures. In this study, RoBERTa's role is to analyze cybersecurity course descriptions, comparing them with industry skill definitions for effective alignment.

2.2.2. SecureBERT - A Domain-Specific Language Model for Cybersecurity

SecureBERT, emerging as a specialized language model, is tailored for the nuanced realm of cybersecurity text analysis. It addresses the intricacies of cybersecurity language, including specific terminology and concepts. Developed on the foundation of RoBERTa, SecureBERT was trained on an extensive corpus of about 1.1 billion words from various cybersecurity sources. This comprehensive training enables SecureBERT to proficiently handle both general and cybersecurity-specific texts. This specialization equips SecureBERT to adeptly handle texts pertinent to cybersecurity courses and industry requirements, making it a key tool in this study for precise analysis of cybersecurity concepts.

3. DATA COLLECTION

The data collection process for this project involved systematic browsing of institutional websites and relevant professional organizations. Two datasets are utilized: skills with their definitions, and courses with their descriptions.

- **Course Descriptions**
  - *UNCW BS Cybersecurity Program*
- **Skill Definitions**
  - *Skills*
    - *Sourced from CyberSeek*
    - *Sourced from ISC²*
  - *Definitions sourced from NIST, CompTIA, INCITS*

3.1. Course Descriptions

Course descriptions were gathered from the Bachelor of Science in Cybersecurity program at the University of North Carolina Wilmington (UNCW), which integrates the expertise of the Congdon School of Supply Chain, Business Analytics and Information Systems, and the Department of Computer Science. This program targets computer, network, and information security, offering tracks in Applied Cybersecurity and Cyber Operations. Its practical learning emphasis, including internships and alignment with industry certifications like Security+ and Certified Ethical Hacker (CEH), prepares graduates for diverse roles in cybersecurity.

Data collection entailed an in-depth review of the UNCW website, focusing on current course offerings in the cybersecurity program. This review included collecting detailed course descriptions for each required and elective course, covering both core and specialized concentrations. A careful compilation and verification of these descriptions was conducted to ensure their accuracy and completeness. The gathered data provides a comprehensive understanding of the curriculum, which is crucial for aligning course content with industry-relevant cybersecurity skills and forms a foundational element of this study's analysis.

3.2. Skill Definitions

In this study, the term 'skills' encompasses both practical abilities and foundational knowledge areas to maintain consistency. This includes a wide range of competencies—examples being firewalls, Linux, Splunk, and ISO 27001—which are all listed under the 'skills' category on the CyberSeek website. This comprehensive approach ensures that the skill set identified is representative of the full spectrum of proficiencies valued in the cybersecurity job market.

3.2.1. Deriving Skills

3.2.1.1. Skills sourced from CyberSeek

The data collection process included a thorough analysis of the CyberSeek website, specifically focusing on entry-level cybersecurity career pathways. This focus was chosen to ensure the academic curriculum's alignment with the skills and competencies required for entry-level positions in the cybersecurity field. From CyberSeek, a detailed list of job titles for entry-level roles was compiled, along with the top skills currently requested by employers and the top future skills anticipated to be in demand over the next five years. This projection underscores the skills anticipated to become increasingly significant in the near future, offering a glimpse into the evolving priorities within the field.

The ordering of these skills reflects their frequency of mention and implied demand within the job postings, with those listed at the top representing the most frequently required competencies for each role.

Key findings include a strong demand for foundational skills like 'Computer Science', 'Cybersecurity', and 'Incident Response', prevalent across various entry-level positions. Emerging skills such as 'Threat Hunting', 'Anomaly Detection', and 'Security Insider Threat Management' are noted for their anticipated future importance. Some skills, like 'SIEM' and 'Firewall/Network Firewalls', are recognized for both their current necessity and expected future relevance.

The following Table 1 highlights these trends, providing a snapshot of the current and future landscape of skills necessary for entry-level cybersecurity roles. This data from CyberSeek is crucial in ensuring

that educational programs in cybersecurity stay aligned with the evolving needs of the job market, particularly for those beginning their careers in this field.

Entry-Level	Top Skills Requested	Top Future Skills Requested with 5-Year Projected	
Cybersecurity Specialist	Cyber Security	Threat Hunting	105%
	Vulnerability	Risk Management Framework	54%
	Computer Science	Threat Intelligence & Response	53%
	Auditing	Network Firewalls	46%
	Information Systems	Phishing	36%
	Risk Analysis		
	Incident Response		
	Security Controls		
	Firewall		
Cyber Crime Analyst	Cyber Threat Intelligence	Threat Hunting	105%
	Cyber Security	Security Information And Event Management (SIEM)	65%
	Incident Response	Anomaly Detection	58%
	Computer Science	Network Firewalls	46%
	Intelligence Analysis	Counter Intelligence	27%
	Digital Forensics		
	Malware Analysis		
	Vulnerability		
	Security Information And Event Management (SIEM)		
Incident & Intrusion Analyst	Incident Response	Threat Hunting	105%
	Cyber Security	Security Information And Event Management (SIEM)	65%
	Incident Management	Anomaly Detection	58%
	Cyber Threat Intelligence	Security Insider Threat Management	57%
	Security Information And Event Management (SIEM)	Counter Intelligence	27%
	Computer Science		
	Linux		
	Splunk		
	Triage		
IT Auditor	Auditing	Blockchain	98%
	Accounting	Security Roles & User Privileges	78%
	Internal Auditing	User Security Management	75%
	Internal Controls	Security Insider Threat Management	57%
	Risk Analysis	ISO 27001	33%
	Finance		
	Information Systems		
	Project Management		
	Public Accounting		

Table 1 - List of In-Demand Cybersecurity Skills: Current and Near Future by CyberSeek

3.2.1.2. Skills sourced from ISC^2 Cybersecurity Workforce Study 2023

The ISC^2 Cybersecurity Workforce Study 2023, in conjunction with insights from CyberSeek, underscores the critical demand and notable shortage of cloud computing security skills in the cybersecurity workforce. Hiring managers consistently identify cloud computing security, along with risk assessment, security analysis, and security engineering, as highly sought-after competencies in potential cybersecurity employees.

Additionally, the study sheds light on the rising importance of AI and machine learning skills in cybersecurity. Although these skills are not currently at the top of hiring managers' lists, they are increasingly recognized by cybersecurity professionals as vital for future career advancement. This trend signals a transformative shift in the cybersecurity arena, anticipating a growing demand for AI/ML

expertise as these technologies continue to evolve and play a pivotal role in combating cybersecurity threats and enhancing defense mechanisms.

Table 2 presents a table from the ISC^2 study that effectively encapsulates these findings. It contrasts the skills presently sought by hiring managers with those perceived by non-hiring cybersecurity professionals as most valuable for career progression, including opportunities for new jobs and promotions. This comparison provides a holistic perspective of the skills landscape, reflecting diverse viewpoints within the cybersecurity community.

What skills are you most looking for right now when hiring?		**Which of these skills do you think are most in demand for security professionals looking to advance their careers (via new jobs and promotions)?**	
(Asked to Hiring Managers)		***(Asked to Non-Hiring Managers)***	
Cloud Computing Security	32%	Cloud Computing Security	47%
Communication Skills	31%	Governancem Risk Management, and Compliance (GRC)	35%
Risk Assessment, Analysis, and Management	31%	Security Engineering	28%
Security Analysis	28%	Risk Assessment, Analysis, and Management	30%
Security Engineering	28%	Artificial Intelligence / Machine Learning	28%
Governancem Risk Management, and Compliance (GRC)	26%	Zero Trust Implementation	27%
Application Security	24%	Communication Skills	24%
Security Administration	22%	SecOps	24%
SecOps	21%	Digital Forensics and Incident Response	24%
Identity and Access Management	20%	Application Security	23%
NOTE: Base: 7,143-7,184 global cybersecurity professionals *Note: "Don't know/does not apply" responses were removed from the sample base.*			

Table 2 - List of In-Demand Cybersecurity Skills: Current and Near Future by ISC2 Cybersecurity Workforce Study 2023

3.2.1.3. Definitions sourced from NIST, CompTIA, INCITS

The skill definitions for this study were derived from NIST and CompTIA. NIST's glossary served as the primary source, but for skills not fully covered, CompTIA's extensive IT-related resources were used. Where NIST definitions were brief, CompTIA provided the necessary detail.

In situations requiring the synthesis of compound terms, components defined by NIST were integrated into comprehensive definitions. For certain skills outside the scope of both NIST and CompTIA, INCITS provided authoritative definitions.

This streamlined approach ensured a robust and relevant skill list, essential for the textual analysis conducted in the study. Each skill was defined with precision, drawing from these authoritative sources to provide accurate and contextually appropriate definitions.

3.2.2. Refinement of Skill List

The development of the skill list for this study involved a meticulous integration and refinement of data from CyberSeek and ISC^2, with the aim of tailoring it to the specific demands of the cybersecurity profession. This process focused on ensuring relevance and coherence with the study's objectives and the current needs of the industry.

Non-specific or broad knowledge areas, such as 'Information Systems,' were excluded, with the list honed to emphasize technical competencies unique to cybersecurity. General skills applicable across various fields, like 'Communication Skills,' and unrelated categories such as 'Accounting' and 'Finance,' were also omitted for their lack of direct relevance to core cybersecurity roles.

The refinement involved consolidating overlapping skills to reduce redundancy and enhance clarity. For instance, 'Internal Auditing' was merged with 'Auditing' to form a comprehensive 'Auditing / Internal Auditing' category. Similarly, 'Internal Controls' was combined with 'Security Controls' to create 'Security Controls / Internal Controls,' encompassing broader governance mechanisms in cybersecurity.

Skills that naturally fit within broader categories were subsumed under them. 'Risk Analysis' and 'Risk Assessment, Analysis, and Management' were included under 'Governance, Risk Management, and Compliance (GRC),' and 'Security Roles and User Privileges' were expanded to 'User Security Management / Access Management.'

In cases where skills encompassed a wide range of competencies, such as 'Digital Forensics and Incident Response,' they were categorized as individual skills to align with distinct professional roles and enable a detailed analysis.

The final skill list was validated through consultation with a cybersecurity expert, ensuring it accurately reflects the necessary competencies for cybersecurity professionals. This process was crucial in aligning the list with industry standards and emerging trends in the cybersecurity workforce.

Table 3 presents refined cybersecurity skills, organized alphabetically within 'Current' and 'Near Future' categories. Each skill is sourced from authoritative entities such as NIST, CompTIA, or INCITS, ensuring comprehensive and accurate definitions. This table comprises the skill set employed in this study.

Research	Current / Near Future	Skill	Resource
CyberSeek	Current	Anomaly Detection	NIST
ISC2	Current	Application security	NIST
ISC2	Current	Artificial Intelligence/Machine Learning	CompTIA
CyberSeek	Current	Auditing	NIST
ISC2	Current	Cloud Computing Security	NIST, CompTIA
CyberSeek	Current	Computer Science	Incits
CyberSeek	Current	Counterintelligence	NIST
CyberSeek	Current	Cyber Security	NIST, CompTIA
CyberSeek	Current	Cyber Threat Intelligence	NIST, CompTIA
CyberSeek & ISC2	Current	Digital Forensics	NIST
CyberSeek	Current	Firewall	NIST
CyberSeek & ISC2	Current	Governance, Risk Management and compliance (GRC)	NIST, CompTIA
CyberSeek & ISC2	Current	Incident Response / Incident Management	NIST, CompTIA
CyberSeek	Current	Intelligence Analysis	NIST
CyberSeek	Current	Linux	CompTIA
CyberSeek	Current	Malware Analysis	NIST, CompTIA
CyberSeek	Current	Project Management	CompTIA
ISC2	Current	SecOps	CompTIA
ISC2	Current	Security Analysis	NIST
CyberSeek	Current	Security Controls / Internal Controls	NIST
ISC2	Current	Security Engineering	NIST
CyberSeek	Current	Security Information And Event Management (SIEM)	NIST, CompTIA
CyberSeek	Current	Splunk	CompTIA
CyberSeek	Current	Threat Hunting	CompTIA
CyberSeek	Current	Triage	
CyberSeek	Current	Vulnerability	NIST, CompTIA
ISC2	Current	Zero Trust implementation	NIST
CyberSeek	Near Future	Blockchain	NIST
CyberSeek	Near Future	ISO 27001	CompTIA
CyberSeek	Near Future	Phishing	NIST
CyberSeek	Near Future	Risk Management Framework	NIST, CompTIA
CyberSeek	Near Future	Security Insider Threat Management	NIST
CyberSeek	Near Future	Threat Intelligence and Response	NIST, CompTIA
CyberSeek & ISC2	Near Future	User Security Management / Access Management	NIST

Table 3 - Refined List of In-Demand Cybersecurity Skills: Current and Near Future

4. TOOLS UTILIZED

The research employed a range of tools and platforms for efficient data collection, analysis, and model evaluation. These tools were chosen for their robustness, integration capabilities, and ease of use which are crucial for conducting a complex analysis.

The tools and libraries mentioned above establish the foundational framework for the computational tasks integral to this study. While this section introduces these resources and outlines their general roles, detailed discussions regarding their specific applications and contributions throughout various phases of the research are reserved for subsequent sections. This approach allows for a comprehensive exploration of how these tools are instrumental in facilitating in-depth model analysis, data visualization, and interactive data engagement, thus underscoring their significance in achieving the objectives of this research.

4.1. Microsoft 365

Excel was utilized for data collection and the organization of the skill sets and course information. Microsoft Forms was the platform of choice for conducting the evaluation survey to the experts due to its user-friendly interface and compatibility with Excel for data analysis.

4.2. Google Colaboratory (Colab)

This cloud-based Python environment was instrumental for coding, model analysis, and visualization tasks. It facilitated embedding generation, cosine similarity calculations, PCA visualization, and the mapping of skills to courses. Colab's integration with Google Drive allowed for streamlined data management and access.

For the execution of this research within the Google Colaboratory environment, the Python programming language served as the foundation, with several computational tools and libraries employed for data loading, natural language processing (NLP), and visualization tasks. Below is an overview of the Python-based tools used and their specific roles in the research process.

4.2.1. Data Preparation and Preliminary Analysis

- *Pandas:* This foundational library in Python was employed for data manipulation and preliminary analysis. It facilitated the importation and structuring of data from Excel files into dataframe structures, essential for the subsequent stages of analysis.

4.2.2. Library Installation and Computational Environment Configuration

- *NumPy:* Essential for scientific computing within the Python ecosystem, NumPy was utilized for efficient numerical data processing, particularly in array manipulations and operations, critical in handling embeddings.
- *Transformers:* This library, pivotal for natural language processing tasks, was installed to enable the utilization of advanced models such as RoBERTa and SecureBERT. It played a significant role in text tokenization and embeddings generation.
- *Openpyxl:* Incorporated to manage Excel files within Python, this library was integral in processing data stored in Excel format, ensuring seamless data integration into the analytical workflow.
- *Ipywidgets:* Facilitated the creation of interactive elements within Jupyter notebooks. Its integration was key in enhancing user engagement and providing dynamic evaluation and presentation of model outputs.

4.3. Natural Language Processing and Tensor Computation Tools

- *RobertaTokenizer and RobertaModel from Transformers*: These components of the Transformers library were pivotal in the text processing phase with the RoBERTa and SecureBERT models. RobertaTokenizer was used for tokenizing the text data into a format suitable for model input. It converted course descriptions and skill definitions into tokens, ensuring the data was in the appropriate structure for analysis. The RobertaModel, on the other hand, was utilized for generating embeddings from the tokenized text. These embeddings represent high-dimensional numerical vectors that capture the semantic essence of the text, enabling the subsequent stages of cosine similarity calculations, PCA visualization, and clustering. By employing these tools from the Transformers library, the project was able to

leverage the advanced capabilities of the RoBERTa and SecureBERT models effectively, ensuring accurate and meaningful extraction of semantic information from the text data, which was crucial for aligning cybersecurity skills with academic course content.

- *AutoTokenizer, AutoModel, and Pipeline from Transformers:* Provided high-level interfaces for various NLP tasks, complementing the RoBERTa and SecureBERT toolkit in the processing and analysis of text data.
- *Torch:* This deep learning library was crucial for tensor operations, supporting the computational demands of natural language processing models.

4.4. Data Analysis and Visualization

- *Sklearn:* Its PCA module and K-Means algorithm were pivotal in reducing the dimensionality of data and identifying patterns within the embeddings. The '*cosine_similarity*' function from '*sklearn.metrics.pairwise*' was particularly significant in assessing textual similarities.
- *Matplotlib.pyplot and Seaborn:* Employed for data visualization, these libraries were instrumental in creating insightful graphical representations, including scatter plots and histograms, thereby facilitating a clearer understanding of the analyzed data.

4.5. Integration of Interactive Tools

- *IPython.display:* This module was instrumental in managing and displaying interactive widgets within the Jupyter notebook environment. It was utilized in conjunction with ipywidgets to create dynamic user interfaces, such as dropdown menus for selecting course codes and names. Specifically, IPython.display enabled the clear and effective presentation of widgets and handled the updating and clearing of outputs upon user interactions. This functionality was crucial in enhancing the interactivity of the research process, allowing for real-time updates and display of results based on user inputs. Through IPython.display, complex data could be navigated and explored interactively, significantly enhancing user engagement and the overall effectiveness of data presentation.

5. EXPLORATORY DATA ANALYSIS (EDA)

Exploratory Data Analysis (EDA) is a fundamental step in data science, involving comprehensive examination and visual interpretation of data. This crucial process reveals underlying patterns, identifies anomalies, and informs hypotheses, thereby guiding subsequent data pre-processing and modeling strategies.

5.1. Load Data

The EDA began with loading two essential datasets into the Python environment using the Pandas library:

- A dataset of refined skills and their definitions
- A dataset of course names and descriptions

These datasets, loaded from Excel files, reflect the project's goal to align academic curricula with industry-relevant cybersecurity skills.

5.2 Data Cleaning

Data cleaning involved inspecting dataframes for the structure and completeness of data. The UNCW BS Cybersecurity program's dataframe showed 74 entries, and the skills dataframe comprised 34 entries. Key columns, vital for text analysis, contained no missing values.

During the data cleaning phase, the primary focus was on ensuring the uniqueness and accuracy of the course dataset from the UNCW BS Cybersecurity program. Duplicate entries, particularly those associated with the program's two tracks—"Applied Cybersecurity" and "Cyber Operations"—were removed to compile a distinct list of courses. This process resulted in a streamlined dataset comprising 46 unique courses, significantly enhancing the efficiency of text analysis when employing the models.

Original text attributes, such as case and special characters, were meticulously preserved. This decision was influenced by the capabilities of the models in interpreting such textual nuances. Retaining the original format is particularly crucial in the cybersecurity domain, where specific terminologies and acronyms (e.g., 'OS' for 'Operating System') carry distinct meanings. This careful treatment of the text ensures that the analysis accurately reflects the nuances of the cybersecurity field.

A comparative analysis of text lengths in skill definitions and course descriptions provided insights into the content's verbosity and complexity. A line chart (Figure 1) depicted this comparison, highlighting the character count distribution within the datasets. Skill definitions demonstrated a varied distribution, with a notable concentration around 200-300 characters, suggesting alignment with the models' tokenization limits. The course descriptions showed more uniformity in length, which may necessitate preprocessing steps like truncation. Understanding these aspects is vital for optimizing natural language processing tasks with BERT-based models like RoBERTa and SecureBERT.

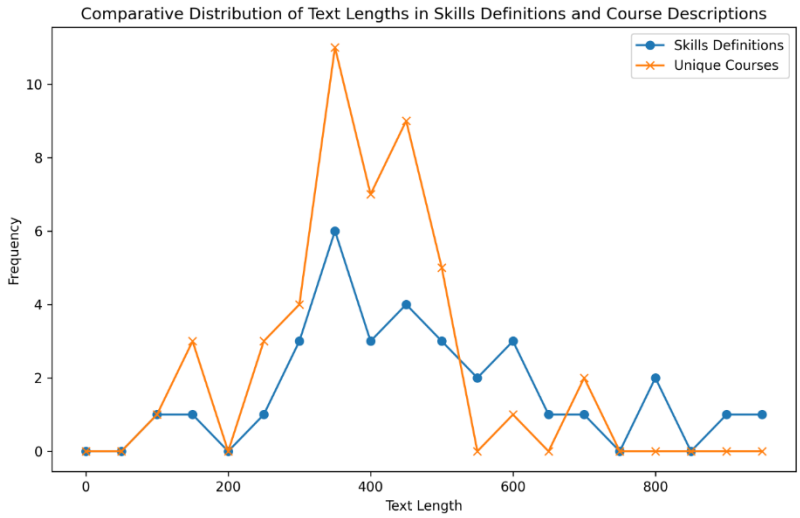

Figure 1 - Comparative Distribution of Text Lengths in Skills Definitions and Course Descriptions

5.3 Data Visualization

Within the exploratory data analysis, the adoption of word clouds reveals predominant themes and concepts in course descriptions and skill definitions, as exemplified in Figure 2. The word cloud for course descriptions prominently features 'security,' 'network,' 'system,' and 'cyber.' These terms suggest

a curriculum with a strong technical focus on the protective measures and architectures of network and system security. Additionally, the presence of 'design' and 'project' indicates that the courses are likely to encompass the planning and execution phases of cybersecurity initiatives, providing a practical, hands-on educational experience.

The word cloud from skill definitions echoes 'security' and 'system,' while distinctively accentuating 'data' and 'information.' This indicates a dual focus on systemic security, data governance, and information management, reflecting the industry's demand for skills that are not only technical but also strategic and analytical. The terms 'analysis,' 'service,' 'control,' and 'risk' point to the need for a comprehensive view of cybersecurity, embracing the assessment of security measures, service provision, risk management, and the implementation of organizational controls.

A comparative analysis reveal:

Shared Core Areas: The commonality of 'security,' 'network,' and 'data' across both word clouds underscores a concerted emphasis on these fundamental cybersecurity pillars.
Educational Emphasis: The course descriptions seem to emphasize more on the practical aspects such as 'system,' 'design,' and 'project,' indicating a curriculum that values the application of knowledge in real-world scenarios.

Industry Skills: Skill definitions appear to prioritize a broader range of competencies, including 'management' and 'analysis,' suggesting that the industry expects professionals to possess both hands-on technical skills and higher-level strategic abilities.

Foundational versus Advanced Skills: The focus on 'network' within course descriptions hints at a foundational approach to cybersecurity education, stressing the importance of building a solid base of knowledge on network structures, protocols, and security strategies. This foundational knowledge serves as a springboard for more advanced topics, aligning with the practical and operational skills represented by terms like 'system' and 'design.'

This analysis underscores the alignment of academic content with industry-required competencies, with educational programs providing the foundational and practical skills necessary for cybersecurity, while

industry demands extend into strategic management and analysis. The word clouds provide a visual representation of the areas of focus within academic and professional realms, highlighting the comprehensive education that spans from foundational knowledge to the application and management of cybersecurity in a professional context.

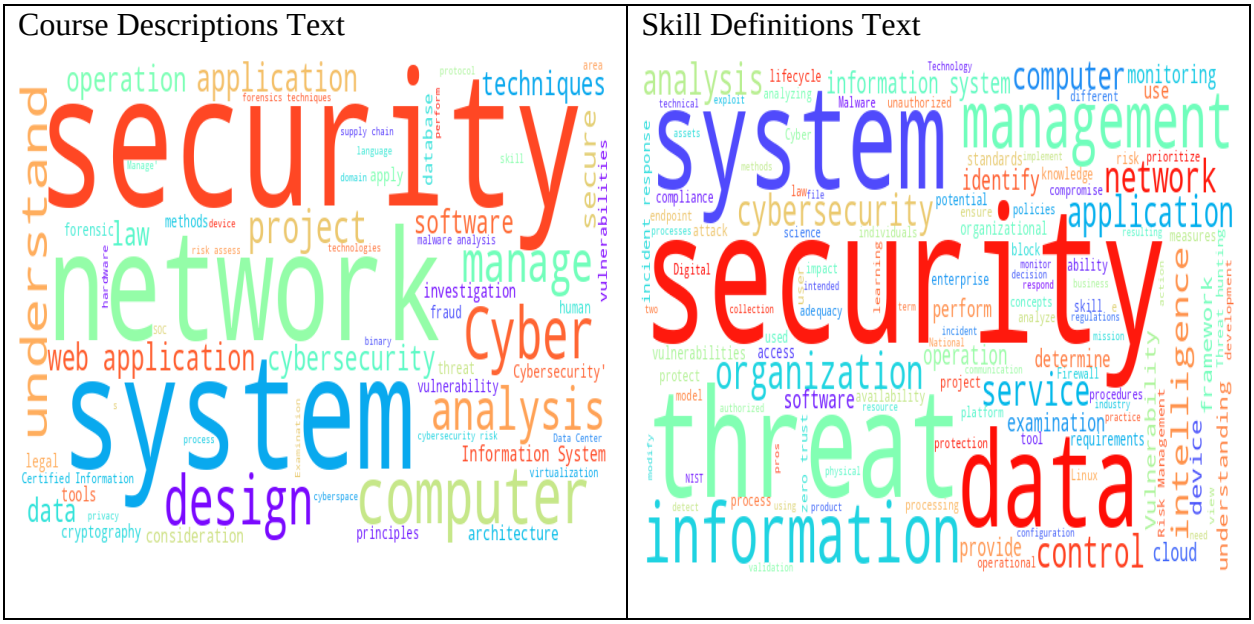

Figure 2 - Comparative Word Cloud Analysis of Cybersecurity Course Descriptions and Industry Skill Definitions

6. COMPREHENSIVE COMPARISON OF AI-DRIVEN MODELS: TECHNICAL ANALYSIS

The forthcoming analysis employs a structured methodological approach to comparatively evaluate the RoBERTa and SecureBERT models, with a focus on their performance and adaptability in analyzing a specialized cybersecurity education corpus.

The methodology for this comparative technical analysis is systematically partitioned into the following sequential components:

- ✓ *Initialization of Models and Tokenizers:* Setting up the foundational elements of both models to ensure a consistent baseline for analysis.
- ✓ *Embedding Generation:* Creating embeddings from textual data to facilitate semantic analysis.
    - o *Token Length:* Assessing token lengths to understand their impact on embedding quality and model constraints.
    - o *Sample Embedding:* Examining a subset of embeddings for initial insights into the semantic space of each model.
- ✓ *Comprehensive Analysis of Embedding Generation:* Evaluating the entire process of embedding generation for accuracy and robustness.
    - o *Visualization using Principal Component Analysis (PCA):* Employing PCA to transform high-dimensional data into a visual, interpretable format.
    - o *Pattern Recognition Through Clustering:* Applying clustering algorithms to identify data groupings.
        - ▪ *Clustering on High-Dimensional Data:* Applying clustering directly to high-dimensional embeddings to explore abstract structures.
        - ▪ *Clustering after Dimensionality Reduction:* Analyzing the effects of PCA on clustering patterns.
    - o *Cosine Similarity Analysis:* Assessing the alignment between course descriptions and skill definitions using cosine similarity.
- ✓ *Threshold-Based Skill Mapping:*
    - o *Determining Optimal Threshold***:** Establishing a similarity score threshold for effective skill-to-course mapping.

- *Mapping Skills to Courses***:** Applying the threshold to derive meaningful course-skill associations.

✓ *Data Structuring and Exportation:* Organizing and preparing the analyzed data for exportation and further examination.

The subsequent sections will detail this comparative analysis, aiming to highlight the distinct characteristics and operational nuances of RoBERTa and SecureBERT in processing the study's data.

6.1. Initialization of Models and Tokenizers

Both projects employed their respective models, RoBERTa and SecureBERT, with each initialized from a pre-trained checkpoint. The choice of model variant was pivotal in each case—'roberta-base' for RoBERTa and 'ehsanaghaei/SecureBERT' for SecureBERT sourced from hugging face.

6.2. Embedding Generation

This section provides insight into the process of converting textual data into numerical embeddings as utilized in language models, demonstrated through an example sentence. This example demonstrates the sequential steps involved in this transformation.

Before delving into the specifics of the embedding generation process, it is pertinent to define several technical terms that underpin the process:

- *Tokenization*: The division of text into tokens, which can be full words or sub-word units. This step is essential for preparing text data for processing by a language model.
- *Hidden States*: Intermediate representations of text within the model, which encapsulate contextual information around each token. These states are generated as the text data passes through the neural network's layers.
- *Vector Space*: A high-dimensional mathematical construct where each token, and the text as a whole, is represented as vectors. Each vector consists of numerous dimensions that collectively capture various attributes and nuances of the text.

- *High-Dimensional Space*: The space in which embeddings exist, often consisting of hundreds or thousands of dimensions. Each dimension can encode different aspects of the text's semantic and syntactic properties, allowing the model to discern and learn from complex patterns within the data.

  Understanding these terms is crucial for comprehending the embedding generation and the nuanced capabilities of the language models employed in this study.

The procedure begins with tokenization, segmenting the text into tokens, which could be entire words or their subcomponents. These tokens are then translated into embedding vectors situated in a high-dimensional space, conventionally with 768 dimensions. Such embeddings enable the models to perceive subtle textual variations and interrelationships, providing a composite semantic interpretation.

The embedding vectors emerge from the model's hidden states, representing intermediate textual features. The averaging of these vectors into a single embedding encapsulates the text's overall semantic and syntactic properties. Although only the initial ten dimensions are typically examined for simplicity, they are part of a broader, abstract feature set that captures the text's overarching meaning within a complex vector space.

6.2.1. Token Length

In the process of embedding generation, an essential step was verifying that the token lengths of our texts did not exceed the 512-token limit of both models. This analysis ensured that the entirety of each text was processed without truncation, preserving the full semantic content for accurate representation in the high-dimensional space.

This phase is foundational for subsequent quantitative text analysis, allowing for semantic-level comparisons. It is pivotal for aligning cybersecurity skill definitions with course descriptions, enhancing the project's ability to undertake a detailed, data-driven exploration of their interplay.

6.2.2 Sample Embedding

To exemplify the embedding process, a comparative analysis was conducted on identical excerpts from the skills definitions and course descriptions datasets, employing both RoBERTa and SecureBERT models. The following Table 4 displays the first ten dimensions of the embeddings, providing an illustrative insight into the models' capability to encode semantic information. This comparison elucidates the nuanced differences and similarities in how each model represents the same textual data.

Text	RoBERTa Embedding (first 10 dimensions)	SecureBERT Embedding (first 10 dimensions)
Text 1 (from skill definitions): Computer Science is the branch of science and technology that is concerned with the processing of data, information, and knowledge by means of computers.	0.017230192199349403, 0.11629314720630646, 0.06312257796525955, 0.11054687201976776, 0.2602975368499756, 0.012217069044709206, -0.03875125199556351, 0.18014103174209595, 0.04846182465553284, -0.11705897748470306	0.008710913360118866, -0.029571400955319405, -0.022374190390110016, 0.05121944844722748, 0.06825514137744904, 0.023958908393979073, -0.004141054581850767, 0.046699829399585724, -0.013553441502153873, -0.04359157010912895
Text 2 (from course descriptions): The course name is 'Cyber Policy, Legal, Ethics, Compliance' and the course description is 'Exploration of the intersection between cybersecurity and policy, law, ethics, privacy, compliance, and law enforcement. Examination of applicable laws and policies related to cyber defense. Topics include privacy and the internet, the effectiveness of cybersecurity applications in preventing crime and abuse, and the application of ethical concepts in cybersecurity.'	0.05628405138850212, -0.0018783085979521275, 0.03107105940580368, 0.1243169978260994, -0.03676924109458923, 0.0675598680973053, -0.019529923796653748, 0.024785608053207397, 0.043169498443603516, -0.03915724903345108	2.7830985800392227e-06, -0.020860152319073677, -0.022103041410446167, 0.02739141322672367, -0.004768958315253258, 0.03067748062312603, 0.00284026050940156, -0.00875475350767374, -0.012785546481609344, -0.027178790420293808

Table 4 - Comparative Embedding Analysis: First Ten Dimensions

In comparing the embedding outputs for two distinct texts, RoBERTa's embeddings exhibit a considerable range and variance, with values extending from approximately -0.03875 to 0.26029 for the first text, and -0.03915 to 0.12431 for the second. Such dispersion suggests RoBERTa's comprehensive feature capture within the text. Conversely, SecureBERT's embeddings display a more constrained range and lower magnitude, hovering around -0.04359 to 0.06825 and -0.02717 to 0.03067 for the respective texts, potentially indicating a more nuanced or conservative interpretation.

This contrast is particularly pronounced in specific dimensions where RoBERTa shows a marked positive value, signifying an attribute that is strongly represented, while SecureBERT's corresponding value approaches zero, suggesting a more subdued feature presence. These observations imply that the

models may prioritize and weigh textual features differently, which could significantly influence their performance in downstream NLP tasks such as clustering or classification. The broader variance in RoBERTa may render it more attuned to textual nuances, whereas SecureBERT's focused embeddings might drive more targeted interpretations of text similarities and groupings.

Although the analysis presently focuses on the initial ten dimensions for illustrative purposes, a holistic understanding necessitates examining the entire spectrum of dimensions to capture the full breadth of the models' text processing capabilities.

In the comparative analysis of the RoBERTa and SecureBERT model outputs, distinct patterns emerge in the embeddings' variance, range, and magnitude. RoBERTa's embeddings demonstrate a greater variance and range, suggesting a broader capture of textual features, whereas SecureBERT presents a more conservative magnitude of embeddings, potentially indicating a more precise or focused representation of features. Notably, specific dimensions reveal significant differences; for instance, RoBERTa's fourth dimension in the first text sample registers a pronounced positive value, in contrast to SecureBERT's closer-to-zero value, implying divergent interpretations of certain textual features by the two models.

These disparities are not merely numerical but could substantially affect the outcomes of downstream NLP tasks such as similarity analysis, clustering, or classification. RoBERTa's sensitivity to text variances might result in a different understanding of text similarity or groupings than SecureBERT's concentrated embeddings. While this analysis, grounded in the first ten dimensions, offers valuable insights, it is merely a fragment of the entire embedding landscape. A comprehensive analysis would require consideration of all dimensions to fully grasp the models' processing and encoding capabilities.

6.3. Embedding Generation

6.3.1. Visualization using Principal Component Analysis (PCA)

The analytical journey through the linguistic landscape of RoBERTa and SecureBERT models engages Principal Component Analysis (PCA) for dimensionality reduction, a method chosen for its adeptness in translating the complexity of high-dimensional data into a comprehensible two-dimensional plane. This technique is adeptly applied to both models, yielding scatter plots that visually narrate the

similarities and divergences between the embeddings of skill definitions and course descriptions, thereby illuminating their potential congruence.

PCA, a staple in the realm of data science, adeptly reduces the intricacies of multidimensional data while retaining its core attributes. The application of PCA to the embeddings derived from both the RoBERTa and SecureBERT models distills these rich, multidimensional representations into a simplified, two-dimensional space. This reduction not only eases the visualization process but also enhances the visibility of the data's intrinsic structure. The resultant two-dimensional scatter plots offer a visual exposition of each model's method for organizing and differentiating the embeddings associated with skills and course descriptions. These visual insights are pivotal for unearthing the spatial relationships and clustering proclivities within the embeddings, providing clarity on how the models interpret and categorize the semantic content of the text.

Figure 3 encapsulates the PCA-transformed embeddings from both models, depicting the interplay and distribution of skills and course embeddings within an accessible bidimensional framework. These visualizations are the precursors to deeper analyses, such as clustering and pattern recognition, enabling the identification of subterranean patterns and groupings that remain concealed within the high-dimensional landscape.

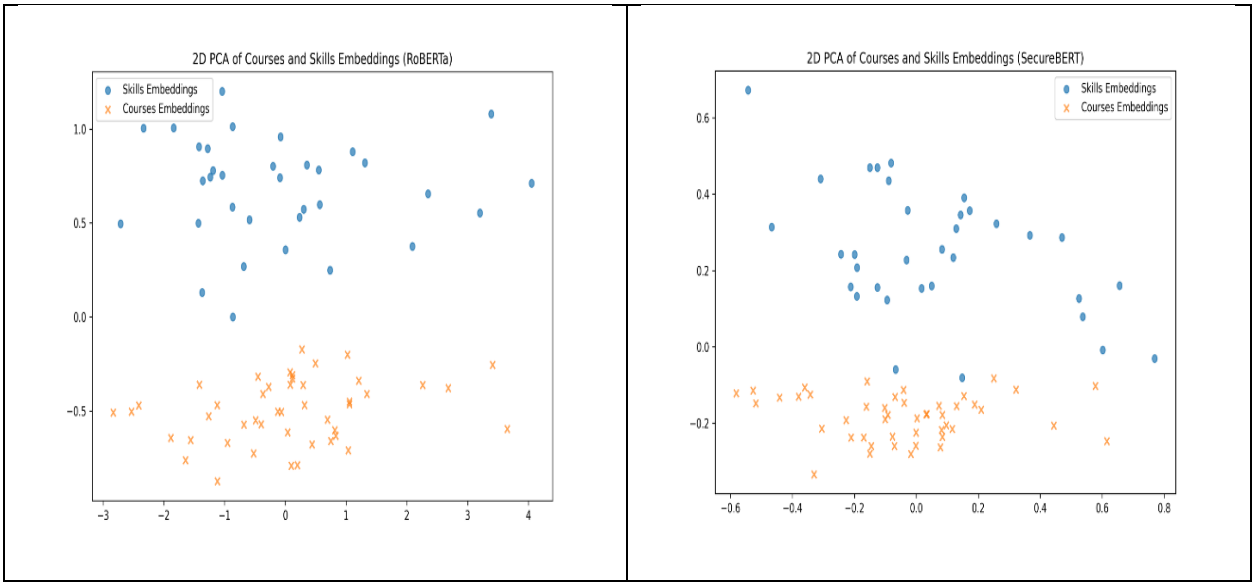

Figure 3 - Comparative Two-Dimensional PCA Scatter Plots of Embeddings

The PCA scatter plot for RoBERTa reveals a broad scatter across the axes, indicative of a substantial variability in the embeddings. The skills embeddings, depicted in blue, are widely scattered, demonstrating RoBERTa's capacity to distinguish between the nuances of various skills. The courses embeddings, colored in orange, although somewhat clustered, exhibit notable dispersion, suggesting that RoBERTa discerns a multifaceted array of features within the course descriptions, showcasing its intricate understanding of educational content.

Conversely, the PCA scatter plot for SecureBERT illustrates a pronounced clustering, particularly within the skills embeddings, delineated in blue. This compact grouping indicates SecureBERT's inclination towards a more concentrated and consistent depiction of skills. The courses embeddings, also clustered but with a relatively narrow spread, point to SecureBERT's streamlined representation of course data. A discernible separation between the skills and courses embeddings in SecureBERT's plot implies the model's acute differentiation between these categories of text.

A synthesis of the analytical insights from both models posits that RoBERTa affords a detailed and widespread representation of embeddings, potentially advantageous for tasks necessitating fine-grained textual discrimination. On the other hand, SecureBERT's depiction seems to coalesce around general themes, possibly offering benefits for the elucidation of broader categorical distinctions within the data.

Upon comparing the two models, RoBERTa appears to offer a granular representation conducive to distinguishing subtle textual differences. SecureBERT, conversely, seems to provide a holistic view, beneficial for identifying overarching themes. The choice between RoBERTa and SecureBERT may thus depend on the specific demands of subsequent NLP applications, where detailed differentiation or broad categorization is required.

6.3.2. Pattern Recognition Through Clustering

Clustering algorithms are employed to unveil patterns within the high-dimensional embeddings from the RoBERTa and SecureBERT models, aiming to illuminate obscured thematic structures.

The Elbow Method was instrumental in determining the optimal number of clusters, balancing the model's complexity with within-cluster variance. This method plots the sum of squared distances within clusters against varying cluster counts, seeking the inflection point where additional clusters do not significantly enhance within-cluster cohesion, as displayed in Figure 4.

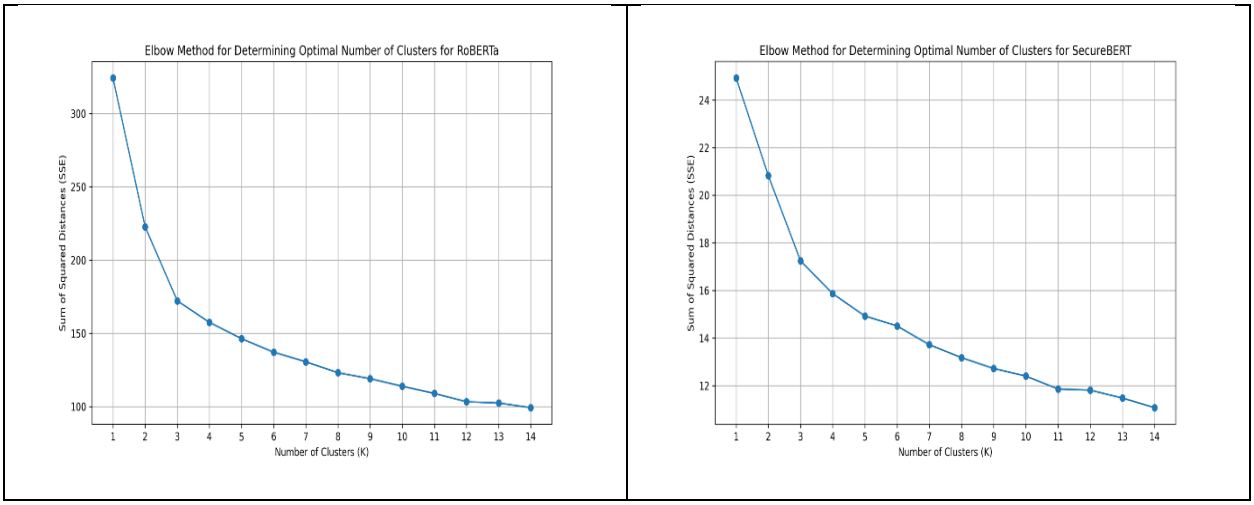

Figure 4 - Comparative Elbow Method Analysis for Optimal Cluster Determination of Embeddings

While the Elbow plots for both models lack a pronounced 'elbow,' indicating a definitive number of clusters, manual inspections and empirical observations suggest three clusters as the most coherent grouping for both RoBERTa and SecureBERT. This number avoids over-segmentation, where clusters may become sparsely filled or empty, ensuring meaningful categorization of the data.

Adopting three clusters for both models standardize the comparison, emphasizing differences arising from the models' inherent data representations rather than from clustering methodology. This consistent approach aligns with the research objective to compare the models' interpretive abilities.

The study advances to clustering within the original high-dimensional space as well as the PCA-reduced two-dimensional space, subsequent to establishing the optimal cluster number. This bifurcated approach allows for an analytical comparison between the intricate patterns inherent in the high-dimensional embeddings and their counterparts in the simplified space afforded by PCA. A meticulous manual inspection follows, where the clusters are scrutinized against domain expertise to confirm their pertinence and integrity.

6.3.2.1. Clustering on High-Dimensional Data

In this phase, K-means clustering was utilized to categorize the high-dimensional embeddings into coherent segments. This algorithm, known for its iterative assignment of data points to the nearest cluster centroid and recalibration of centroids to minimize variance within clusters, is particularly effective for grouping data based on feature similarity, thus serving as an apt tool for this study's comparative analysis.

The clustering application to the embeddings from the RoBERTa and SecureBERT models reveals intricate patterns within the cybersecurity domain. Figure 5, displaying RoBERTa's clustering, and Figure 5, illustrating SecureBERT's, are placed adjacent for a direct visual comparison of the models' data structuring.

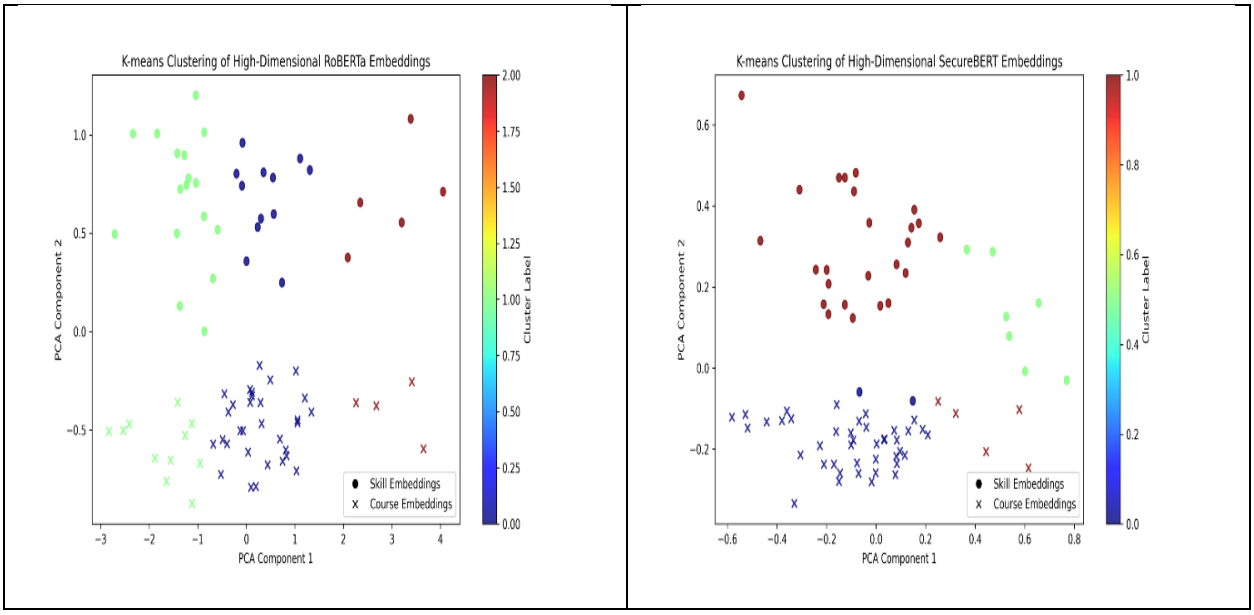

Figure 5 - Comparative K-Means Clustering of Embeddings

RoBERTa's clustering demonstrates considerable variability, suggesting a multifaceted representation of cybersecurity concepts. Conversely, SecureBERT's clustering is more concentrated, implying a targeted interpretation of the thematic content. Notably, the PCA Component 1 range is broader in RoBERTa's plot, indicating greater variance within the embeddings.

A detailed examination of RoBERTa's first cluster encompasses a diverse spectrum of cybersecurity topics, indicative of the model's broader categorical comprehension. SecureBERT's equivalent cluster offers a more selective categorization of skills, potentially enhancing specificity in cybersecurity applications.

Each model encapsulates a blend of technical and strategic aspects of cybersecurity, albeit with distinctive grouping methodologies. RoBERTa's clusters are inclusive, combining varied cybersecurity knowledge, whereas SecureBERT's clusters are discerning, segregating practical from strategic content more sharply.

Ultimately, the selection between RoBERTa and SecureBERT for practical application would depend on the particular nuances required in the cybersecurity context. For nuanced distinctions, RoBERTa's expansive clustering may be beneficial, while for tasks necessitating clear-cut thematic categories, SecureBERT's focused clusters might be preferred. The manual inspection corroborates the quantitative analysis, confirming the relevance and coherence of the selected clusters against established domain knowledge.

6.3.2.2 Clustering after Dimensionality Reduction

The process commenced with an initial analysis of the high-dimensional data, followed by the application of Principal Component Analysis (PCA) to reduce the data dimensions. This initial stage involved engaging with the full complexity of the high-dimensional data, utilizing PCA as a more advanced technique to simplify the data, reduce noise, and potentially enhance the performance of clustering algorithms. The aim of dimensionality reduction, preceding the clustering of embeddings, was to mitigate the 'curse of dimensionality' that often obscures meaningful patterns within high-dimensional data. By reducing the dimensions, it was possible to retain the most significant features of the data, thus simplifying its complexity without sacrificing the essence of its underlying structure.

Post-dimensionality reduction, the K-means clustering algorithm was employed, leading to the formation of discernible clusters that encapsulated aspects of both datasets. This step proved crucial in refining the embeddings, thereby highlighting salient groupings for efficient processing by the K-means algorithm.

Figures 6 illustrates the PCA-reduced embeddings from the RoBERTa and SecureBERT models, respectively. These visualizations demonstrate that both models effectively capture the intricate relationships between skill definitions and course descriptions in the field of cybersecurity.

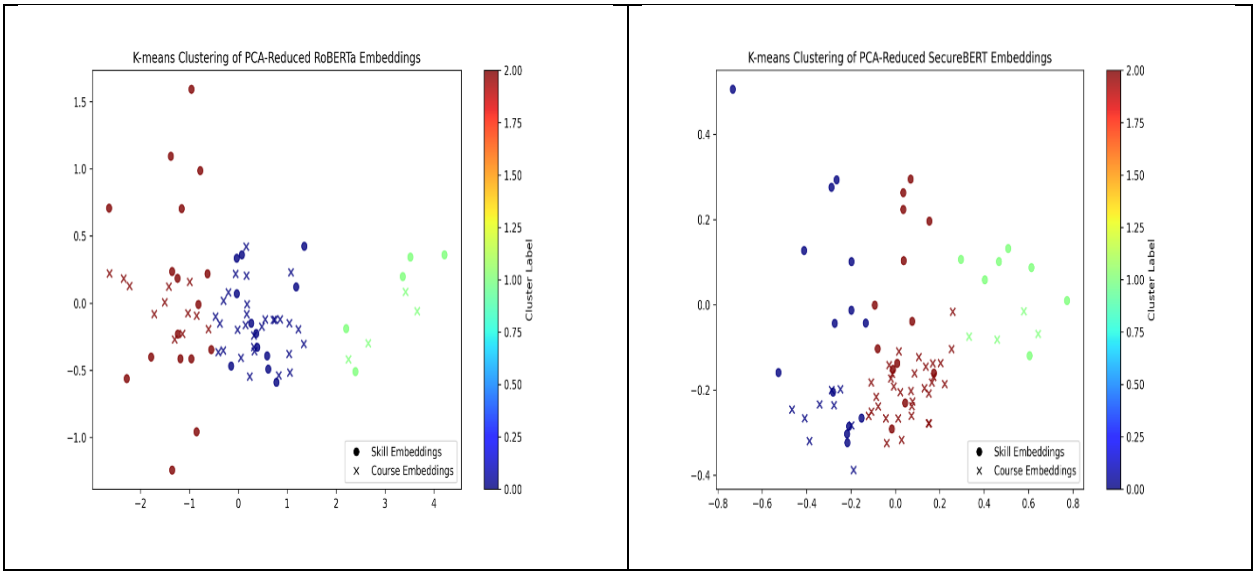

Figure 6 - Comparative PCA-Reduced Embeddings

The first cluster from the RoBERTa model emerged as a comprehensive repository of cybersecurity expertise, encompassing a range from auditing to advanced threat detection and response. This cluster illustrated both the practical aspects of cybersecurity, such as incident management and Security Information and Event Management (SIEM), and strategic elements like policy and legal considerations.

Conversely, SecureBERT's comparable cluster displayed notable variances in point density and distribution, indicating different emphases on certain skills and courses. This variation may reflect divergences in the underlying training data or algorithmic differences between the two models.

The PCA Component scales in the SecureBERT visualizations were more constrained, suggesting less variance among the embeddings compared to the broader distribution observed in the RoBERTa figure. The clusters in SecureBERT appeared to have a tighter grouping with less overlap, potentially indicating more distinct separations between the categories of embeddings in the model. Interestingly, in both figures, the 'Skill Embeddings' and 'Course Embeddings' were not entirely segregated into separate clusters, suggesting some semantic similarity between the embeddings for skills and courses.

From these visualizations, it was inferred that while dimensionality reduction and clustering captured some data patterns, the true semantic relationships might be more intricate. These plots offer a simplified view conducive to guiding further analysis but do not fully encapsulate the nuances of high-dimensional data.

Both RoBERTa and SecureBERT clusters converged on pivotal cybersecurity skills such as threat intelligence, risk management frameworks, and cloud security, underscoring the fundamental competencies universally recognized in the cybersecurity field. However, there was a notable distinction in the clustering patterns between the models. RoBERTa's clusters displayed a broad spectrum, incorporating a diverse range of skills and courses, suggesting a model with a wide-ranging understanding of cybersecurity fields. In contrast, SecureBERT's clusters showed a more segmented approach, potentially offering a more focused identification and categorization of cybersecurity aspects.

In conclusion, the application of K-means clustering following dimensionality reduction on RoBERTa and SecureBERT embeddings yielded insightful delineations of cybersecurity competencies. The resultant clusters affirmed the models' capability to discern and categorize complex cybersecurity knowledge into coherent groupings. Therefore, the choice between employing RoBERTa or SecureBERT for specific cybersecurity applications should be informed by the granularity and specificity of the clustering outcomes, aligning them with the objectives of the respective task.

6.3.3 Cosine Similarity Analysis

The ensuing phase of the investigation employed cosine similarity for the analytical assessment of the correspondence between courses and skills. Cosine similarity, a robust method for textual analysis, quantifies the congruence between text-based embeddings. This similarity analysis involved computing the cosine similarity between the embeddings of course descriptions and skill definitions.

Cosine similarity, with values ranging from -1 to 1, measures the cosine of the angle between two vectors within a multidimensional space. A value closer to 1 denotes higher similarity. In text embeddings, this metric elucidates the degree to which a course's content correlates with specific skills, offering an objective measure to gauge the alignment between educational content and industry skill requisites.

This methodological approach yielded a systematic set of cosine similarity scores for each course description, which is pivotal for the next phase of mapping skills to course content. An exemplar of this analytical process is depicted in Table 5, showcasing the course “Fundamentals of Cybersecurity” alongside its similarity scores as determined by different models, sorted from highest to lowest.

Upon analyzing and comparing the outcomes from the RoBERTa and SecureBERT models for the course "Fundamentals of Cybersecurity," subtle yet significant variations in the models' embedding spaces were revealed. The initial similarity scores provided by RoBERTa, commencing at 0.98, indicate a closely matched embedding space, signifying a high similarity between the vectors of the course and the skills. In contrast, SecureBERT's initiation at a slightly reduced score of 0.96 suggests a strong, albeit marginally less similar, relationship.

For RoBERTa, the most aligned skills—Governance, Risk Management and Compliance (GRC), Risk Management Framework, Cyber Security, Vulnerability, and Cloud Computing Security—register similarity scores exceeding 0.98, denoting a tight correlation within the model's semantic interpretation of these concepts in relation to the course.

SecureBERT concurs with RoBERTa on the high relevance of the same skills to the course content, albeit with marginally lower scores around 0.94, hinting at a nuanced disparity in representation. Notably, SecureBERT positions Cyber Threat Intelligence immediately after these skills, while RoBERTa lists Security Controls / Internal Controls, indicating a modest reordering and difference in emphasis.

Mid-range scores exhibit divergence; RoBERTa assigns higher similarity to 'Security Information And Event Management (SIEM)' and 'Threat Hunting', while SecureBERT assigns lower scores, possibly reflecting distinct training data or model-specific interpretations.

The lower end of the similarity spectrum accentuates more pronounced disparities. For instance, RoBERTa ascribes a higher similarity to 'Computer Science', contrasting with SecureBERT’s lower valuation, suggesting RoBERTa's propensity to link general computer science knowledge more integrally with cybersecurity.

In conclusion, while RoBERTa and SecureBERT largely concur on the skills pertinent to the "Fundamentals of Cybersecurity" course, their quantification of similarity exhibits distinct patterns. Such differences could be attributed to variations in the models' pre-training datasets, internal architectures, or fine-tuning processes. These disparities provide insights into the potential applications of each model, contingent on the specific focus desired within a cybersecurity curriculum or the alignment of course content with professional skillsets. The detailed nature of these similarities highlights the importance of a discerning approach when utilizing these models for educational and industry alignment purposes.

Course Name	RoBERTa		SecureBERT	
	All Skills	Similarity Score	All Skills	Similarity Score
Fundamentals of Cybersecurity	Governance, Risk Management and compliance (GRC)	0.986426234	Governance, Risk Management and compliance (GRC)	0.966022134
	Risk Management Framework	0.985317349	Risk Management Framework	0.964459836
	Cyber Security	0.984902561	Cyber Security	0.960484147
	Vulnerability	0.98459357	Vulnerability	0.952383637
	Cloud Computing Security	0.981398165	Cloud Computing Security	0.947519302
	Security Controls / Internal Controls	0.978641331	Cyber Threat Intelligence	0.934803247
	Cyber Threat Intelligence	0.974425435	Zero Trust implementation	0.929624259
	Zero Trust implementation	0.966344774	Security Controls / Internal Controls	0.928451896
	Security Information And Event Management (SIEM)	0.965769768	Linux	0.914993167
	Linux	0.963323593	Incident Response / Incident Management	0.909022808
	Threat Hunting	0.962159395	Application security	0.908066928
	Application security	0.962089479	Auditing	0.90415132
	Incident Response / Incident Management	0.961221695	Security Information And Event Management (SIEM)	0.904099166
	Digital Forensics	0.960287213	Digital Forensics	0.900013089
	Malware Analysis	0.953531146	Threat Hunting	0.898291111
	Auditing	0.951073349	Threat Intelligence and Response	0.895757318
	Threat Intelligence and Response	0.949518561	Malware Analysis	0.878554583
	ISO 27001	0.94116509	Blockchain	0.875453055
	Blockchain	0.94113028	ISO 27001	0.872630715
	Security Insider Threat Management	0.933940172	Security Insider Threat Management	0.871974528
	Firewall	0.933699846	Project Management	0.867681384
	Project Management	0.932896256	Firewall	0.858315051
	SecOps	0.928739905	SecOps	0.856490731
	Security Engineering	0.924318492	Artificial Intelligence/Machine Learning	0.85525012
	Counterintelligence	0.924311817	Anomaly Detection	0.851487637
	Phishing	0.922961056	Security Engineering	0.850573659
	Anomaly Detection	0.921794295	Security Analysis	0.849391818
	Security Analysis	0.920971155	Splunk	0.841044068
	Intelligence Analysis	0.918134987	Counterintelligence	0.837941349
	Artificial Intelligence/Machine Learning	0.912192702	Phishing	0.837591529
	Splunk	0.910400331	Intelligence Analysis	0.834623456
	User Security Management / Access Management	0.901119471	User Security Management / Access Management	0.827574313
	Triage	0.881614983	Computer Science	0.781591654
	Computer Science	0.876796961	Triage	0.729303002

Table 5 - Comparative Cosine Similarity Analysis for 'Fundamentals of Cybersecurity' Course

6.4. Threshold-Based Skill Mapping

6.4.1 Determining Optimal Threshold

The determination of an optimal threshold for skill-to-course mapping was conducted through an analysis of histograms, specifically titled "Course & Skill - Distribution of Cosine Similarities to Decide Threshold" for the RoBERTa and SecureBERT models, respectively. These histograms are foundational for several reasons:

- *Threshold Guidance*: The visualizations underpin the selection process for cosine similarity thresholds. Scrutiny of the distribution patterns facilitates the identification of appropriate cutoffs that distinguish relevant from irrelevant skill-to-course alignments.
- *Dataset Characterization*: The histograms elucidate the degree of alignment between course content and skill definitions. A higher mean value suggests robust alignment, whereas a broader distribution indicates variability in the similarities.
- *Strategic Analysis Framework*: Insights from the similarity score distribution inform the refinement of methodologies for skill-to-course mapping, based on these similarity metrics.

Histogram Analysis:
Figure 7 illustrates the frequency distributions of cosine similarity scores for the RoBERTa and SecureBERT models. These distributions provide a quantitative overview of the similarity ranges and the general alignment of course content with corresponding skills.

A dashed line within each histogram represents the mean cosine similarity score, serving as a central metric of the dataset's similarities. Kernel Density Estimate (KDE) plots overlaid on the histograms offer a continuous probability density function, enhancing interpretability. Histograms aid in establishing thresholds that signify significant similarity. The thresholds are derived based on the distribution patterns and mean values, seeking to balance the inclusion of relevant skills with the exclusion of non-essential associations.

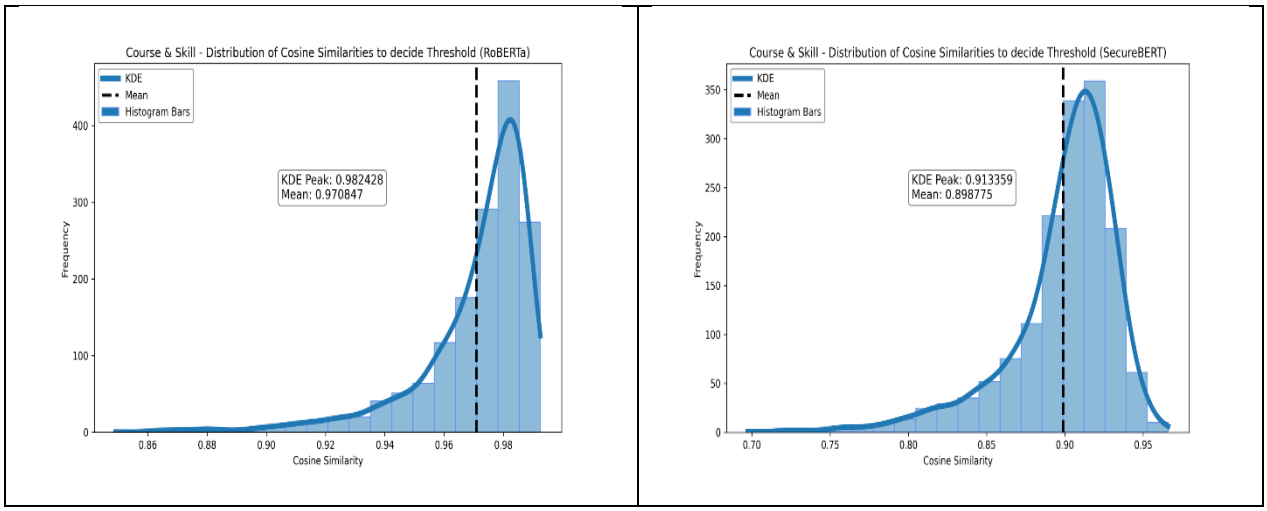

Figure 7 - Comparative Histogram Analysis of Cosine Similarity Scores for Skill-to-Course Mapping

For RoBERTa, the histogram indicates a distribution skewed towards higher similarity scores, with a mean of 0.970847 and a KDE peak at 0.982428. In contrast, the SecureBERT histogram, with a mean of 0.898775 and a KDE peak at 0.913359, suggests a wider range of relationships between course content and skill definitions.

6.4.2 Mapping Skills to Courses

Tailored threshold values are necessary due to the distinct distributions between the models. For RoBERTa, a threshold of 0.9835 is deemed suitable to encapsulate relevant skills effectively. Conversely, for SecureBERT, a lower threshold of 0.924 is chosen to include necessary skills while accounting for the model's broader variance in scores.

6.5. Data Structuring and Exportation

Upon establishing threshold-based mapping, the outputs were collated and presented to course instructors for evaluation via a survey. This step not only validates the theoretical approach but also engages cybersecurity experts to assess the practical applicability of the model outputs in correlating educational content with requisite skills.

7. EXPERT EVALUATION OF AI-DRIVEN MODEL RESULTS WITH A SURVEY

Following up the computational technical analysis, this section outlines the preparatory steps leading to the quantitative analysis of RoBERTa and SecureBERT models. The examination, conducted by experts, assesses the models' accuracy and efficiency in educational applications.

The rapid advancement in educational technologies, especially in cybersecurity, necessitates a rigorous evaluation of AI-driven models by subject matter experts. Considering the sophistication of AI in textual technical analysis, it is imperative to examine these models' practicality and effectiveness in real-world educational settings. Therefore, a survey was conducted to collect expert feedback from instructors, which is critical for assessing the alignment of course content with industry-relevant skills as interpreted by the AI-driven models.

The survey was designed to evaluate three principal dimensions of the AI-driven models' output:

- The degree of accuracy, and efficacy with which each model maps course descriptions to relevant skill sets.
- The pertinence of the skills identified by the models in correlation to the actual course content and learning outcomes.
- The feasibility of applying these AI-generated insights to the development of cybersecurity curricula.

Out of the 37 unique courses in the UNCW BS Cybersecurity Program, 27 were evaluated by experts through this survey. The remaining courses were not included due to the unavailability of instructors for participation in the survey. Survey participants were presented with data from RoBERTa and SecureBERT, along with corresponding course descriptions and skill definitions. The survey incorporated both quantitative and qualitative questions to form a holistic perspective of the models' performance, specifically from an educational standpoint.

7.1 Survey Questions and Their Purposes

Accuracy of Course-Skill Mappings: The questions aimed to determine the precision with which the models' outputs align with actual course content, providing a direct measure of the models' accuracy.

Relevance of Identified Skills: The survey sought to establish the significance and relevance of the skills identified by the models, ensuring that the models pinpoint skills that are genuinely applicable to the courses.

This preparatory stage sets the stage for the subsequent analysis, where the results of the survey are subjected to a detailed statistical evaluation in the section 'Evaluating AI-Driven Model Efficacy.' Following this, we discuss 'Instructors' Perspectives on AI Integration in Cybersecurity Curriculum Development,' which represents the second part of the survey's outcomes.

7.2 Performance and Statistical Analysis of Survey Result

This section presents the methodological framework and findings from the statistical analysis of the survey data, assessing the text analysis capabilities of RoBERTa and SecureBERT models. Two key metrics were used: Accuracy Score and Relevance Score, reflecting the models' proficiency in skill-to-course alignment.

The Accuracy Score, defined as the quotient of skills instructors deem directly relevant over the total identified by the models, quantifies the precision of the models in aligning skills with course content. Mathematically, the Accuracy Score is calculated as follows:

$$\text{Accuracy Score} = \frac{\text{Directly Relevant Count}}{\text{Model Skill Count}}$$

The Relevance Score, which includes both directly and briefly relevant skills, offers a broader assessment of the models' discernment abilities. The Relevance Score is calculated using the equation:

$$\text{Relevance Score} = \frac{\text{Directly Relevant Count} + \text{Briefly Relevant Count}}{\text{Model Skill Count}}$$

For instance, RoBERTa's identification of five skills in a course, deemed two directly relevant and one briefly relevant by instructors, would culminate in an accuracy score of 0.4 and a relevance score of 0.6.

Such quantitative assessments are visually rendered in bar charts, facilitating a lucid comparison across different courses, as illustrated in Figure 8 and 9.

The analysis of the accuracy score reveals that RoBERTa exhibits exemplary performance in courses like CIT 301 (Cloud Computing & Virtualization), which covers a range of topics cloud computing technologies, including deployment of services across various models such as SaaS, PaaS, DaaS, and IaaS, as well as issues related to VPNs, firewalls, and two-factor authentication, and MIS 322 (Information Assurance), focusing on standards, rules, and regulations. This breadth aligns with RoBERTa's extensive training corpus, enabling it to accurately match a wide range of skills to the course descriptions.

Conversely, SecureBERT's domain-specific training gives it an edge in specialized courses such as CYBR 340 (Cybersecurity Program Management), which delves into security awareness, industry practices, and the managerial aspects of security planning, and CYBR 431 (Cloud Security), a senior-level course, focuses on cloud security architecture and the principles, patterns, standards, and technologies that govern secure cloud-based services. SecureBERT's domain-specific focus on cybersecurity language enables it to excel in these courses, where nuanced understanding and application of cybersecurity concepts are essential.

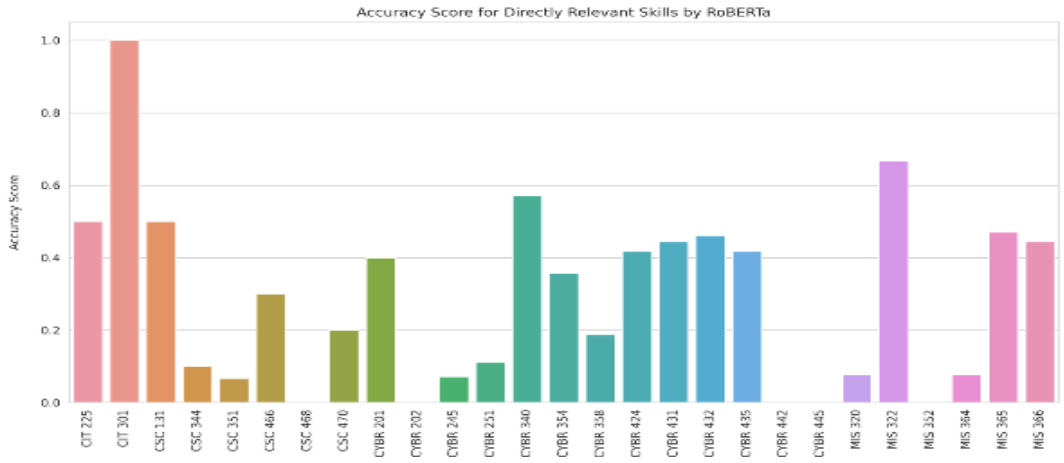

Figure 8 - Bar Charts of Accuracy Score - RoBERTa

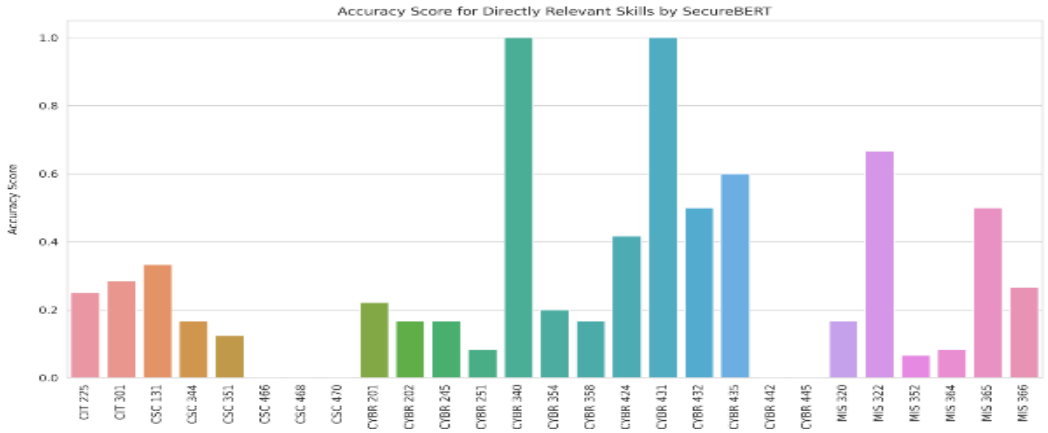

Figure 9 - Bar Charts of Accuracy Score - SecureBERT

Figures 10 and 11 elucidate the relevance score comparisons, revealing the unique strengths of the RoBERTa and SecureBERT models. RoBERTa outperforms in CIT 225 (Linux Administration), addressing the configuration and administration of Linux operating systems. In CIT 301 (Cloud Computing & Virtualization), which surveys and configures a wide array of cloud computing technologies and services, RoBERTa's ability to process diverse technical concepts proves advantageous. Additionally, RoBERTa excels in MIS 365 (Ethical Hacking), focusing on vulnerability exploitation to protect information systems, where RoBERTa's expansive training corpus aptly matches the course's focus on security architectures and penetration testing.

On the other hand, SecureBERT's tailored capabilities performes better in CSC 470 (Hardware Security), where the intricate knowledge of hardware and physical security is paramount; CYBR 340 (Cybersecurity Program Management), which requires a nuanced understanding of security policies and practices; and CYBR 431 (Cloud Security), dealing with advanced cloud security architectures. SecureBERT's specialized training enables it to adeptly handle the detailed, domain-specific content of these courses. Moreover, in CYBR 435 (Cyber Fraud and Investigation), SecureBERT's focused training on cybersecurity specifics provides an edge in mapping skills related to fraud prevention and forensic investigation. Notably, both models perform equally well in CYBR 445 (Certified Information Systems Security Professional), suggesting a convergence in their ability to deal with the advanced topics that make up the CISSP domains, which demand both breadth and depth of understanding in cybersecurity. The relevance score analysis underscores the importance of model training in relation to course

specificity. RoBERTa's generalist approach makes it versatile for a wide range of IT-related topics, while SecureBERT's focused training aligns it with the specialized and advanced areas of cybersecurity, indicating the significant role of domain-specific knowledge in the performance of AI-driven models.

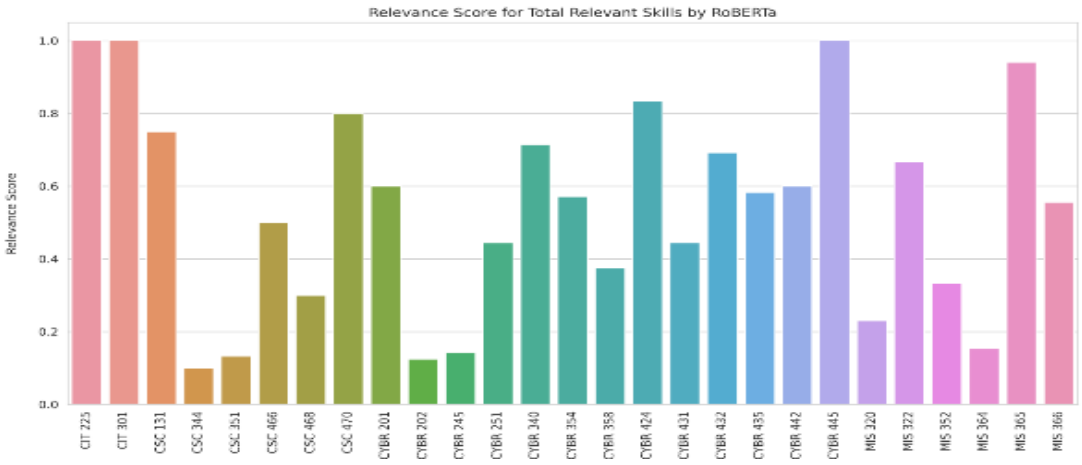

Figure 10 - Bar Charts of Relevance Score - RoBERTa

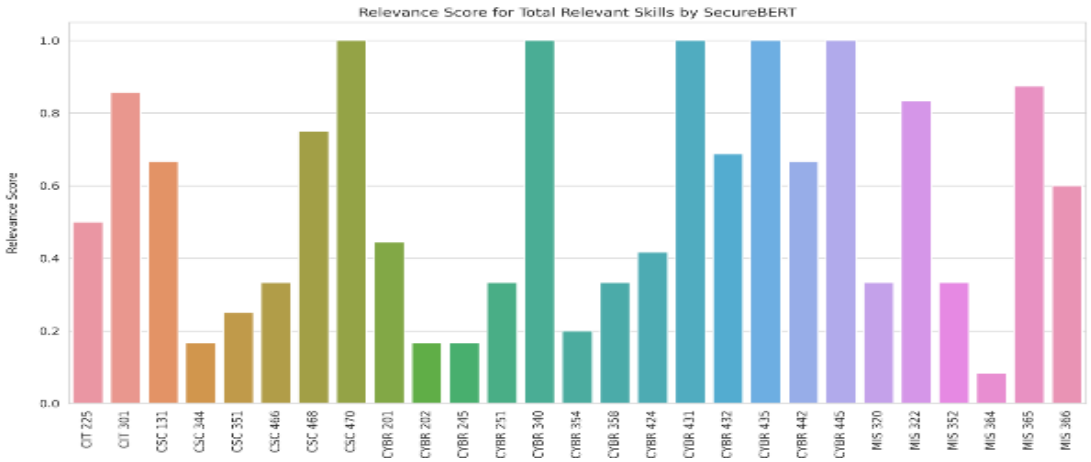

Figure 11 - Bar Charts of Relevance Score - SecureBERT

The mean and standard deviations were calculated to establish an overall performance indicator for the models. RoBERTa's slightly higher mean accuracy score suggests a nuanced edge in identifying directly relevant skills. In contrast, SecureBERT's higher mean relevance score and greater standard deviation

for both accuracy and relevance scores imply a wider scope in skill identification and with greater variability across different course contexts, as presented in Table 6.

	RoBERTA	SecureBERT
Mean Accuracy Score for Directly Relevant Skills	0.290	0.275
Mean Relevance Score for Total Relevant Skills	0.540	0.555
Standard Deviation of Accuracy Score	0.253	0.279
Standard Deviation of Relevance Score	0.288	0.311

Table 6 - Comparative Performance Analysis: Mean and Standard Deviation

The section concludes with an Analysis of Variance (ANOVA), a statistical test used to evaluate the differences in performance between the RoBERTa and SecureBERT models. The F-value measures the extent of variance in performance between the models, while the P-value assesses the probability that any observed differences occurred by chance. An F-value of 0.034 and a P-value of 0.854 (Table 7) indicate there is no significant discrepancy in performance between the models according to the expert evaluations. The observed differences in model performance are not substantial enough to determine a clear preference for either model based on the survey data. The findings imply that external factors, such as the specific nature of course descriptions or the expertise of the respondents, may also influence the perceived efficacy of the models.

	F-Value	P-Value
ANOVA	0.034	0.854

Table 7 - ANOVA Results for Performance Comparison

In conclusion, the detailed investigation deduces that RoBERTa, with its generalized training, holds a marginal advantage in courses with a broad thematic scope. SecureBERT, with its focused cybersecurity-specific training, demonstrates pronounced strengths in specialized subject areas. This comparative analysis underscores the potential of AI-driven models to enhance the alignment of academic curricula with industry-relevant skills, thereby reinforcing the models' applicability in the design and development of educational content.

7.3 Instructors' Perspectives on AI Integration Analysis of Survey Result

This section delves into instructors' perspectives regarding the integration of AI-driven models in curriculum design, particularly within the realm of cybersecurity education. It presents an analysis of survey results that explore not only the perceived benefits of using AI-driven models like RoBERTa, SecureBERT, and ChatGPT for enhancing course-to-skill mapping accuracy but also investigates the alignment of course descriptions with educational outcomes. Furthermore, the survey sought to gauge instructors' readiness to adopt these AI-generated insights into their teaching practices, a critical aspect of modern pedagogical approaches. This analysis is pivotal in understanding the current landscape of AI integration in education and identifying potential areas for improvement and further research.

The survey probed the potential of detailed syllabi, which include topics and learning outcomes, in refining the accuracy of course-to-skill mappings created by the RoBERTa and SecureBERT models. As shown in the first part of Figure 12, titled "Potential of Detailed Syllabi to Enhance Accuracy of the Models," a substantial majority of survey participants, approximately 86%, affirmed this potential. Within this group, 19% expressed definitive certainty, while 67% agreed with a likelihood of improvement. However, a minority of 15% remained neutral or undecided, indicating a trend towards endorsing enriched syllabi for enhanced performance of AI-driven models in educational contexts.

Furthermore, the survey assessed the alignment of current course descriptions with detailed syllabi, encompassing actual course topics and outcomes. This aspect is depicted in the second part of Figure 12, titled "Alignment of Course Descriptions with Detailed Syllabi." Here, responses indicated a recognition of the need for improvements, with 52% of respondents supporting this view. Conversely, 37% of respondents did not commit to a definitive viewpoint on this matter, and 11% perceived no need for modification. This distribution suggests a significant degree of uncertainty regarding the current adequacy of course descriptions in aligning with detailed educational outcomes.

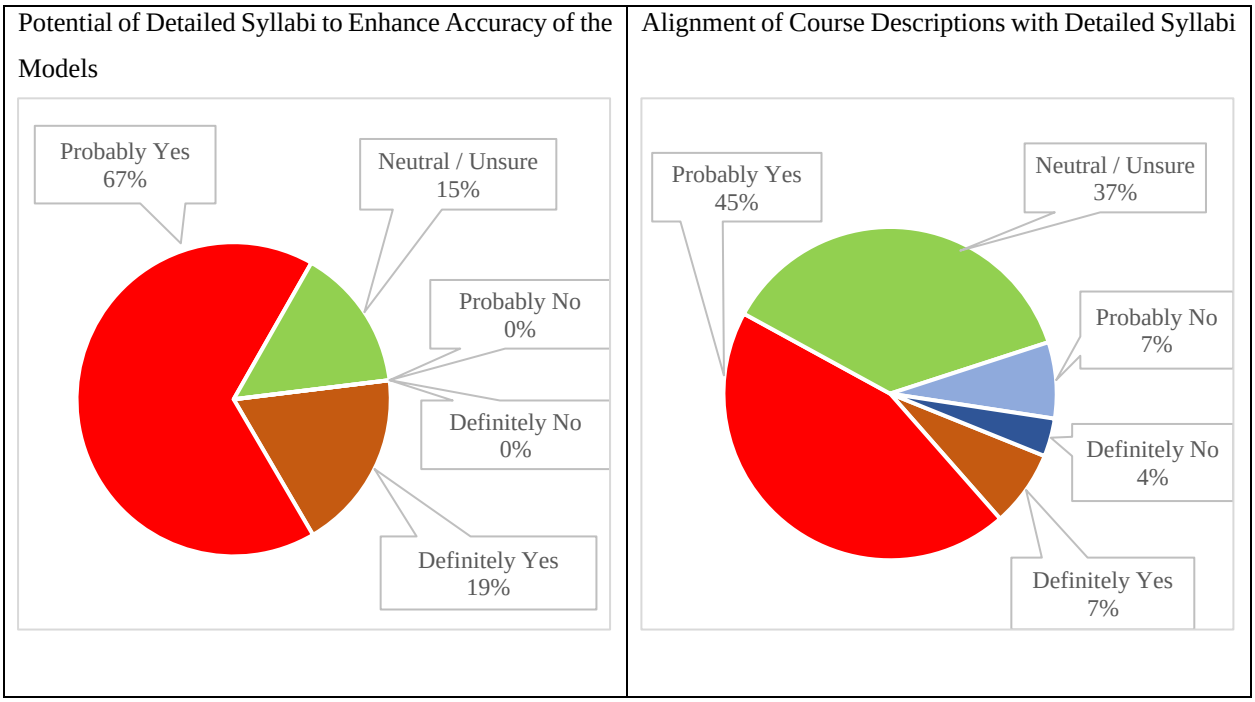

Figure 12 - Survey Insights on Detailed Syllabi's Impact on AI Model Accuracy and Course Description Alignment

The likelihood of employing AI insights for course development was examined, revealing a diverse array of opinions among instructors. As depicted in Figure 13, the first part, titled “Instructors' Likelihood to Utilize AI-Driven Insights from RoBERTa and SecureBERT,” indicates a plurality, 37%, appeared likely to utilize insights from these models. However, a notable 44.4% expressed skepticism, and 19% of the instructors remained neutral. This diversity in responses highlights the need to address concerns and bolster the credibility of RoBERTa and SecureBERT as AI-driven tools in education.

The survey also explored the receptivity towards ChatGPT insights among instructors, particularly those who were previously uncertain or skeptical about other AI models. This aspect, shown in the second part of Figure 13 titled “Instructors' Receptivity to ChatGPT Insights Among Skeptics of RoBERTa and SecureBERT Models,” reveals that 35% of these respondents were inclined to likely use ChatGPT, in contrast to 47% who remained unlikely, with 18% maintaining neutrality. This pattern indicates a discerning yet tentative acceptance of AI tools, underscoring the need for a more detailed elucidation of the unique attributes that make certain AI technologies, such as ChatGPT, more appealing for educational application.

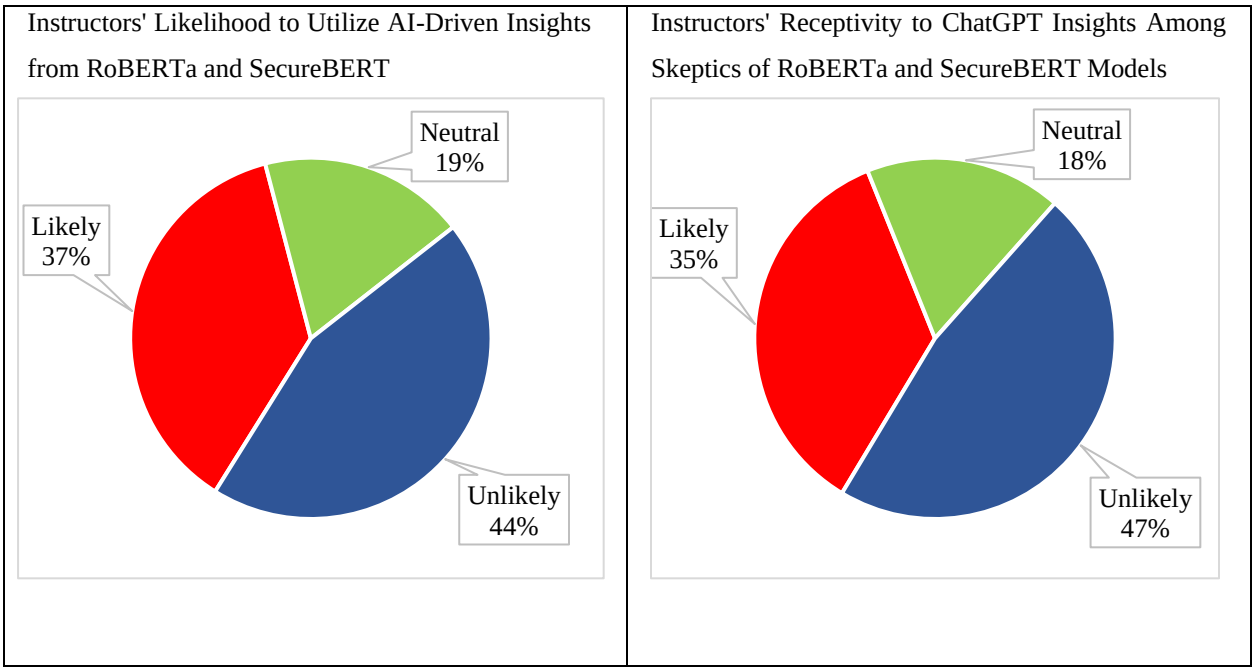

Figure 13 - Distribution of Instructors' Perspectives on Willingness to Employ AI-Driven Insights for Course Development and Revision

8. DISCUSSION OF FINDINGS

This research has illuminated key aspects of the application and efficacy of AI-driven models, specifically RoBERTa and SecureBERT, in cybersecurity education. The technical analysis revealed significant differences in the performance and adaptability of these models, with each model demonstrating unique strengths that cater to different aspects of cybersecurity education.

8.1 Technical Analysis and Model-Specific Applications

Embedding Generation and Token Analysis: RoBERTa and SecureBERT were analyzed for their abilities in semantic interpretation. RoBERTa showed considerable variability in its embeddings, indicating a robust capability to handle a wide range of content. SecureBERT displayed a more constrained range, reflecting its targeted interpretative strength.

PCA and Clustering Techniques: The use of PCA and clustering algorithms provided insights into how each model organizes and interprets cybersecurity data. RoBERTa distinguished diverse features, indicating its suitability for a comprehensive analysis of various topics. SecureBERT exhibited pronounced clustering, ideal for specialized courses.

Cosine Similarity Analysis: This analysis offered an objective measure of the alignment between course content and industry skills. RoBERTa generally yielded higher similarity scores, suggesting a closer alignment with course and skill content. SecureBERT showed strong alignment as well but with slightly lower similarity scores, indicative of its focused approach.

Model-Specific Similarity Thresholds: The establishment of these thresholds is crucial for customizing AI-driven curriculum development. RoBERTa's higher threshold aligns with its broader feature capture, while SecureBERT's lower threshold suits its more conservative similarity range.

8.2 Comparative Effectiveness in Skill-to-Course Mapping

The comparative analysis of RoBERTa and SecureBERT highlighted their distinct advantages in skill-to-course mapping. RoBERTa showed higher accuracy in courses with a broad thematic scope, while

SecureBERT was preferred for courses with specialized cybersecurity knowledge. This direct comparison underscores the importance of selecting the appropriate model based on the specific educational content and objectives.

8.3 Expert Perspectives and Instructors' Openness to AI Integration:

Expert evaluations and the survey of instructors revealed a cautiously optimistic attitude towards incorporating AI insights into curriculum development. The growing acceptance of AI integration in educational settings was evident, especially towards advanced AI tools like ChatGPT. This shift in instructors' perspectives is critical, as it indicates an evolving readiness to embrace AI-driven methodologies in pedagogical practices.

8.4 Instructors’ Perspectives on AI Tools

The survey also highlighted a discerning yet tentative acceptance of AI tools among instructors. Notably, a significant shift was observed among previously skeptical instructors, with a considerable proportion showing a favorable view towards ChatGPT. This pattern points to the need for elucidating the specific attributes of AI tools that instructors find appealing or reliable for educational applications.

8.5 Importance of Customizing AI-Driven Curriculum Development:

The establishment of model-specific similarity thresholds and the nuanced application of each model based on their distinct features highlight the importance of customizing AI-driven curriculum development. This approach ensures that academic programs are effectively aligned with industry needs, enhancing the relevance and efficacy of cybersecurity education.

The insights gleaned from our research not only advance our understanding of AI-driven models but also inform the strategic development of curricula that address the dynamic requirements of the cybersecurity industry.

9. CONCLUSIONS AND IMPLICATIONS FOR CYBERSECURITY EDUCATION

The outcomes of this study demonstrate the transformative potential of AI-driven models in synchronizing cybersecurity education with the requisites of the industry. The effective mapping of course content to skills in demand by RoBERTa and SecureBERT is anticipated to significantly contribute to the preparation of a skilled cybersecurity workforce. The research broadens the existing understanding of AI-driven models within educational frameworks and accentuates the necessity for their tailored application, contingent on specific course requirements.

9.1 Future Directions for Research

Subsequent investigations are encouraged to explore the application of AI-driven models within diverse pedagogical contexts and academic disciplines. The development of advanced models, tailored to the particular educational requirements, is expected to augment the precision of course-to-skill mappings. A comprehensive understanding of the dynamics of AI integration and the evolving needs of the cybersecurity job market remains a priority.

9.2 Practical Implications and Curriculum Alignment

The insights from this study provide a strategic path for cybersecurity educators and curriculum developers to align academic offerings with market demands. Detailed skill mapping within UNCW's BS Cybersecurity program could guide students in selecting courses and ensures that they are not only gaining theoretical knowledge but also acquiring practical skills that are directly relevant to the evolving demands of the cybersecurity job market.

9.3 Practical Application: Course-to-Skill Mapping in UNCW BS Cybersecurity Program

Building upon the study's findings, Table 8 distills the course-to-skill mapping for the UNCW BS Cybersecurity Program. This mapping is a direct embodiment of the research's implications, offering a pragmatic instrument for students to align their academic pursuits with the competencies sought in the cybersecurity industry. It is a strategic guide, ensuring that the theoretical knowledge imparted in the classroom is complemented by practical skills that resonate with the current and future needs of the cybersecurity job market. Table 8 offers a visual representation of these alignments and gaps, providing clear indicators for potential curriculum evolution.

The following analysis of the program's course offerings highlights the strengths and opportunities within the curriculum to meet these industry-aligned competencies:

- *Foundational Skills*: The curriculum's integration of fundamental skills such as "Anomaly Detection," "Cloud Computing Security," and "Firewall" within core courses highlights the program's commitment to providing a solid cybersecurity foundation.

- *Integrated Skills*: A thematic thread of "Application Security," "Auditing," and "Incident Response / Incident Management" woven throughout both core and elective courses exemplifies the curriculum's holistic approach to cybersecurity education.

- *Current and Emerging Trends*: The proactive inclusion of burgeoning fields like "Blockchain" and "Artificial Intelligence/Machine Learning" within the core offerings reflects the program's dedication to future-proofing its students' skillsets.

- *Specialization Opportunities*: By positioning specialized areas such as "Digital Forensics" and "Threat Hunting" within elective courses, the program affords students the flexibility to pursue their individual cybersecurity interests and career aspirations.

- *Depth and Breadth of Learning*: The curriculum's versatility is evident in its treatment of "Linux," "Malware Analysis," and "Security Analysis," ensuring students gain a comprehensive understanding that spans the theoretical to the practical.

- *Opportunities for Curriculum Development*: Notably absent from direct instruction, "ISO 27001" represents a strategic opportunity for curriculum expansion to encompass this critical information security standard.

- *Tool-Specific Instruction*: While hands-on training with "Splunk" is not provided, the curriculum's focus on SIEM tools equips students with adaptable skills applicable to a variety of security platforms.

- *Competency Development*: The development of "Triage" competencies is implied rather than explicitly taught, emerging organically as students engage with coursework that demands prioritization and incident handling.

- *Advanced Topics*: The current absence of "Security Insider Threat Management" and "Zero Trust Implementation" from direct instruction invites future enhancements to the curriculum, aligning with the cybersecurity sector's evolving priorities.

- *Alignment with Industry Requirements*: The comprehensive coverage of "Vulnerability" and "User Security Management / Access Management" in core courses aligns the program closely with the foundational competencies in high demand across the cybersecurity industry.

- *Curriculum Responsiveness*: The program's current structure demonstrates responsiveness to industry trends, yet the incorporation of emerging and critical topics like "Zero Trust Implementation" would further solidify its relevance.

UNCW BS Cybersecurity program's core and elective coursework provide a robust foundation and diverse specialization opportunities, preparing students for the dynamic field of cybersecurity.

Skills In-Demand	Directly Taught	Briefly Taught
Anomaly Detection	Core	Elective
Application security	Core & Elective	Core & Elective
Artificial Intelligence/Machine Learning	Core	Elective
Auditing	Core & Elective	Core & Elective
Blockchain	Core	Elective
Cloud Computing Security	Core	Core & Elective
Computer Science	Core	Elective
Counterintelligence	Elective	Core & Elective
Cyber Security	Core & Elective	
Cyber Threat Intelligence	Core & Elective	Core & Elective
Digital Forensics	Elective	Core & Elective
Firewall	Core	Core & Elective
Governance, Risk Management and compliance (GRC)	Core	
Incident Response / Incident Management	Core & Elective	Core & Elective
Intelligence Analysis	Elective	Elective
ISO 27001	Not Covered	Core & Elective
Linux	Core & Elective	Core & Elective
Malware Analysis	Elective	Core & Elective
Phishing	Core	
Project Management	Core	Core & Elective
Risk Management Framework	Core	Core
SecOps	Elective	Elective
Security Analysis	Elective	Core & Elective
Security Controls / Internal Controls	Core	Core & Elective
Security Engineering	Elective	Elective
Security Information And Event Management (SIEM)	Core & Elective	Core
Security Insider Threat Management	Not Covered	Not Covered
Splunk	Not Covered	Not Covered
Threat Hunting	Elective	Core & Elective
Threat Intelligence and Response	Elective	Core & Elective
Triage	Not Covered	Not Covered
User Security Management / Access Management	Core	
Vulnerability	Core & Elective	Not Covered
Zero Trust implementation	Not Covered	Core & Elective

Table 8 - Course-to-Skill Mapping for UNCW BS Cybersecurity Program

In conclusion, this study serves as a catalyst for the advancement of cybersecurity education, highlighting the critical role of AI-driven models in the development of curricula. Strategic alignment of course content with industry demands is essential for effectively preparing graduates to navigate the challenges of the cybersecurity landscape. The integration of AI models and the alignment of educational content with in-demand skills effectively bridges the gap between academic preparation and industry requirements, equipping graduates with the competencies necessary for success in the dynamic field of cybersecurity.

10. REFERENCES

1. Aghaei, E., Niu, X., Shadid, W., & Al-Shaer, E. (2022). SecureBERT: A Domain-Specific Language Model for Cybersecurity. ArXiv, abs/2204.02685. Retrieved from https://ar5iv.labs.arxiv.org/html/2204.02685
2. Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. Retrieved from https://arxiv.org/abs/1409.0473
3. CyberSeek. (n.d.). Cybersecurity supply and demand heat map. Retrieved from https://www.cyberseek.org/heatmap.html
4. Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. North American Chapter of the Association for Computational Linguistics. ArXiv, abs/1810.04805. Retrieved from https://ar5iv.labs.arxiv.org/html/1810.04805
5. Doe, J. (2021, April 1). What is embedding and what can you do with it? Towards Data Science. Retrieved from https://towardsdatascience.com/what-is-embedding-and-what-can-you-do-with-it-61ba7c05efd8
6. Editor, C. S. R. C. C. (n.d.). Information system - glossary: CSRC. CSRC Content Editor. Retrieved April 12, 2023, from https://csrc.nist.gov/glossary/term/information_system
7. Gatos, L. (2023). Transformer models: an introduction and catalog. Retrieved from https://ar5iv.labs.arxiv.org/html/2302.07730
8. International Information System Security Certification Consortium. (2023). [ISC2 Cybersecurity Workforce Study 2023]. Retrieved from https://media.isc2.org/-/media/Project/ISC2/Main/Media/documents/research/ISC2_Cybersecurity_Workforce_Study_2023.pdf?rev=28b46de71ce24e6ab7705f6e3da8637e
9. International Information System Security Certification Consortium. (2023). Our association and mission. ISC2. Retrieved from https://www.isc2.org/about
10. Kerner, S. M. (2023, September). What are large language models (LLMs)? TechTarget. Retrieved from https://www.techtarget.com/whatis/definition/large-language-model-LLM
11. Lavton, G. TechTarget. (n.d.). Transformer model. Retrieved from https://www.techtarget.com/searchenterpriseai/definition/transformer-model
12. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. ArXiv, abs/1907.11692. Retrieved from https://ar5iv.labs.arxiv.org/html/1907.11692
13. Lutkevich, B. (2023, October 3). 16 of the best large language models. TechTarget. Retrieved from https://www.techtarget.com/whatis/feature/12-of-the-best-large-language-models
14. Maitra, M. (2023, September 12.). Importance and impact of exploratory data analysis. DZone. Retrieved from https://dzone.com/articles/importance-and-impact-of-exploratory-data-analysis
15. National Institute of Standards and Technology. (2023, June 6). About NICE. Retrieved from https://www.nist.gov/itl/applied-cybersecurity/nice/about
16. National Institute of Standards and Technology. (n.d.). Retrieved from https://csrc.nist.gov/glossary
17. NVIDIA. (n.d.). What is a transformer model? Retrieved from https://blogs.nvidia.com/blog/what-is-a-transformer-model/
18. Our association and mission: (ISC)². Our Association and Mission | (ISC)². (n.d.). Retrieved from https://www.isc2.org/About
19. University of North Carolina Wilmington. (n.d.). Cybersecurity, B.S. Retrieved from https://uncw.edu/academics/majors-programs/cse-csb/cybersecurity-bs/

20. University of North Carolina Wilmington. (n.d.). Cybersecurity, B.S. Retrieved from https://uncw.edu/academics/majors-programs/cse-csb/cybersecurity-bs/
21. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Retrieved from https://arxiv.org/abs/1706.03762
22. Wang, Y., & Liu, K. (2018). Deep learning for natural language processing: advantages and challenges. National Science Review, 5(1), 24-26. https://doi.org/10.1093/nsr/nwx106
23. Xuemei L., & Huirong F. (2023, November 19). SecureBERT and LLAMA 2 Empowered Control Area Network Intrusion Detection and Classification. Retrieved from https://arxiv.org/abs/2311.12074