INTERNATIONAL JOURNAL OF ADVANCED ARTIFICIAL INTELLIGENCE RESEARCH (IJAAIR)

eISSN: 3087-4270

Volume. 02, Issue. 02, pp. 09-15, February 2025

ENHANCING AI-CYBERSECURITY EDUCATION: DEVELOPMENT OF AN AI-BASED CYBERHARASSMENT DETECTION LABORATORY EXERCISE

Prof. Michael T. Edwards
School of Cybersecurity and Privacy, Georgia Institute of Technology, USA

Article received: 18/12/2024, Article Accepted: 18/01/2025, Article Published: 28/02/2025
DOI: https://doi.org/10.55640/ijaair-v02i02-02

ABSTRACT

The escalating prevalence of cyberharassment and online abuse poses significant challenges to digital safety and mental well-being, necessitating advanced detection and mitigation strategies. Artificial intelligence (AI), particularly machine learning and natural language processing (NLP), offers powerful tools for identifying such malicious content. However, effectively integrating AI concepts into cybersecurity education, especially concerning social-cybersecurity threats, remains an evolving field. This article details the design and pedagogical rationale behind an AI-based cyberharassment detection laboratory exercise aimed at enhancing AI-cybersecurity education. The lab emphasizes hands-on, experiential learning, guiding students through data preprocessing, model training (e.g., using BERT-based models), evaluation, and crucial analyses of model bias and vulnerability to adversarial attacks. The proposed laboratory serves to equip future cybersecurity professionals with practical skills in developing and critically evaluating AI systems for online safety, while simultaneously fostering an understanding of ethical implications, such as racial bias in detection algorithms. This approach addresses the growing demand for cybersecurity experts adept at leveraging AI, bridging the gap between theoretical knowledge and real-world application in combating complex online threats.

INTRODUCTION

The digital landscape, while offering unprecedented connectivity and opportunities, is increasingly marred by pervasive issues such as cyberharassment, hate speech, and online abuse [1]. These forms of online aggression, which can range from bullying [2] to targeted harassment and hate speech [3], have profound negative impacts on individuals, often leading to psychological distress, social isolation, and even physical harm. The pervasive nature of these threats is underscored by statistics; for instance, cyberbullying rates among adolescents remain significant, with data consistently showing a substantial percentage of young people experiencing online harassment [6]. Addressing these challenges effectively requires a robust understanding of the underlying dynamics of online abuse, often categorized under the umbrella of social-cybersecurity [4, 5], and the development of sophisticated detection mechanisms.

Artificial Intelligence (AI), particularly advancements in machine learning (ML) and natural language processing (NLP), has emerged as a promising frontier in the automated detection of cyberharassment and hate speech [7]. Companies and researchers are actively deploying AI models to sift through vast amounts of online content to identify and flag abusive language [9, 10]. Datasets and challenges like the Hateful Memes Challenge [8] further illustrate the community's effort to advance AI in this domain. However, the application of AI in this sensitive area is not without its complexities and ethical considerations. A significant challenge lies in the potential for AI models to exhibit biases, often reflecting and amplifying existing societal prejudices embedded within training data. For example, AI-driven hate speech detection tools have been documented to display racial bias, mislabeling content from marginalized groups as offensive more frequently than similar content from dominant groups [11, 15]. This inherent bias necessitates a critical and nuanced approach to AI model development and deployment [16, 17].

The rapid evolution of AI in cybersecurity demands a corresponding evolution in cybersecurity education. Traditional cybersecurity curricula, while strong in

network security, cryptography, and digital forensics, often lack sufficient emphasis on the practical application and ethical implications of AI in combating modern online threats. There is a growing need to equip the next generation of cybersecurity professionals with the skills to not only design and implement AI solutions but also to critically evaluate their fairness, robustness, and potential vulnerabilities, such as susceptibility to adversarial attacks [12, 13, 14]. Adversarial examples, subtle perturbations to input data that can cause AI models to misclassify, pose a significant threat to the reliability of AI systems, including those used for cyberharassment detection [14].

To address this educational gap, this article proposes and details the design of an innovative AI-based cyberharassment detection laboratory exercise. This lab is conceived as an integral component of a comprehensive AI-cybersecurity curriculum, leveraging experiential learning principles [24, 25, 26, 27, 30] and project-based learning [20] to provide students with hands-on experience. The primary objective is to empower students to build, test, and analyze AI models for cyberharassment detection, while critically examining issues of bias and adversarial robustness. By engaging with real-world challenges, students will develop not only technical proficiency but also a crucial understanding of the ethical responsibilities inherent in developing AI for social good.

METHODS

Pedagogical Framework and Learning Objectives

The design of the AI-based cyberharassment detection lab is firmly rooted in constructivist and experiential learning theories [24, 25]. It moves beyond traditional lecture-based instruction by adopting a "flipped classroom" approach, where foundational knowledge is acquired independently (e.g., through readings or video lectures), and class time is dedicated to hands-on problem-solving and deeper analytical engagement [18, 19]. This approach is complemented by principles of project-based learning [20] and challenge-based learning [31], allowing students to grapple with authentic, complex problems related to social-cybersecurity [4, 5, 36].

The key learning objectives for students completing this lab include:

1. Technical Proficiency: Gaining practical experience in applying Natural Language Processing (NLP) techniques and machine learning models (specifically deep learning models like BERT [38]) for text classification.

2. Threat Understanding: Deepening their understanding of the nature of cyberharassment and other forms of online abuse.

3. Data Ethics and Bias Awareness: Developing the ability to identify, analyze, and attempt to mitigate bias in AI models, particularly racial bias [11, 15, 16, 17].

4. Adversarial Robustness: Understanding the concept of adversarial attacks on AI systems and practical methods for generating and detecting adversarial examples in text [12, 13, 14].

5. Critical Thinking & Problem Solving: Enhancing analytical skills to evaluate model performance, identify limitations, and propose improvements.

6. Interdisciplinary Perspective: Fostering an appreciation for the intersection of computer science, cybersecurity, and social sciences [21, 23, 32].

Laboratory Exercise Design

The lab is structured into distinct modules, designed to be completed iteratively over several weeks, typically in a dedicated lab session or as part of a larger course project. This modularity allows for flexibility in implementation across various curricula (e.g., cybersecurity, data science, AI ethics courses) [28, 29, 32].

Module 1: Introduction to Cyberharassment and Data Preprocessing

- Concept Introduction: Review of cyberharassment definitions, types, and impact [1, 2, 3, 6]. Discussion on the societal and ethical implications of automated detection.

- Data Acquisition: Students are provided with a pre-curated dataset of online text messages labeled for cyberharassment (e.g., a subset of the Hateful Memes Challenge dataset [8] or similar public datasets, ensuring ethical considerations for data use).

- Text Preprocessing: Hands-on exercises involving tokenization, lowercasing, stop-word removal, and other NLP preprocessing steps using Python libraries. Emphasis on understanding how preprocessing choices can impact model performance.

Module 2: Building and Evaluating a Baseline AI Detection Model

- Feature Engineering/Representation: Introduction to word embeddings (e.g., Word2Vec, GloVe) and the concept of transformer models like BERT for rich text representation [38].

- Model Training: Students train a baseline text classification model (e.g., a simple recurrent neural network, or a pre-trained BERT model fine-tuned for the task) to detect cyberharassment.

• Model Evaluation: Focus on key classification metrics: precision, recall, F1-score, and accuracy. Students learn to interpret these metrics in the context of cyberharassment detection, understanding the trade-offs between false positives and false negatives.

Module 3: Bias Analysis in AI Detection Models

• Concept of Algorithmic Bias: Detailed discussion on how training data can introduce and perpetuate biases, particularly racial bias, into AI models [11, 15]. Review of real-world examples and their societal consequences.

• Bias Detection Techniques: Students are guided to analyze the model's performance across different demographic subgroups (e.g., by analyzing detection rates for language associated with specific racial or ethnic groups, if such annotations are available in the dataset or simulated).

• Mitigation Strategies: Introduction to techniques for debiasing models, such as data augmentation, re-sampling, or adversarial debiasing [16, 17]. Students attempt to apply a simple debiasing technique and evaluate its impact on fairness and performance. This practical engagement fosters critical thinking about AI ethics [31].

Module 4: Adversarial Attacks and Model Robustness

• Introduction to Adversarial Examples: Explain the concept of adversarial examples in NLP and their implications for cybersecurity [12, 13, 14]. Discuss how subtle changes to text can trick models (e.g., by changing a few characters or words [13]).

• Generating Adversarial Text: Students use a pre-built tool or implement a simplified adversarial attack algorithm (e.g., character-level perturbations or synonym replacement) to generate adversarial versions of non-harassing text that the model misclassifies as harassment, or vice-versa.

• Evaluating Robustness: Students assess their model's susceptibility to these generated adversarial examples. Discussion on the challenges of making AI models robust against such attacks and potential defensive strategies.

Tools and Environment

The lab utilizes a common programming environment for data science and AI:

• Programming Language: Python, due to its extensive libraries and readability.

• Libraries: pandas for data manipulation, numpy for numerical operations, scikit-learn for traditional ML models and evaluation metrics, tensorflow or pytorch for deep learning models (especially transformer models like BERT [38]), and specialized NLP libraries (e.g., transformers, nltk, spacy).

• Development Environment: Jupyter Notebooks or Google Colab, providing an interactive and collaborative platform.

• Data Collection/Annotation (Optional for future iterations): For more advanced labs, students could engage in data collection or annotation using tools like Qualtrics [37] for survey-based data or custom annotation interfaces.

Assessment Strategy

Student learning outcomes are assessed through a combination of deliverables:

• Lab Reports: Comprehensive reports for each module detailing methodology, code implementation, results, analysis of findings (including bias and robustness), and critical reflections.

• Code Submission: Clean, well-commented code demonstrating their implementation.

• Presentations: (Optional) Students present their findings, particularly the bias analysis and adversarial attack results, fostering communication skills [23].

• Quizzes/Exams: Short assessments to test conceptual understanding of AI models, bias, and adversarial attacks.

This structured, hands-on approach aims to cultivate deep learning, problem-solving skills, and ethical awareness, which are crucial for students from diverse backgrounds entering the fields of data science and cybersecurity [21, 32, 33].

Results (Expected Educational Outcomes)

The implementation of this AI-based cyberharassment detection laboratory exercise is anticipated to yield several significant educational outcomes, directly addressing the stated learning objectives and enhancing the preparedness of students for real-world AI-cybersecurity challenges.

Enhanced AI and Cybersecurity Technical Competency

Students completing the lab are expected to demonstrate a practical understanding of the AI development pipeline for text classification. This includes:

• Proficiency in NLP Techniques: Students will gain hands-on experience with fundamental NLP tasks such as text preprocessing (tokenization, normalization),

vectorization, and leveraging pre-trained language models like BERT [38] for feature extraction and fine-tuning.

• Model Development and Evaluation: Participants will be able to train, validate, and test machine learning models for classification tasks, interpreting standard performance metrics (precision, recall, F1-score, accuracy) within the context of cyberharassment detection.

• Implementation Skills: Through coding assignments, students will develop practical programming skills in Python using relevant AI/ML libraries, applying theoretical knowledge to concrete problems.

Deepened Understanding of Online Abuse and Social-Cybersecurity

The lab will provide a tangible connection between abstract AI concepts and the very real problem of online abuse [1]. By working with actual or simulated cyberharassment data, students will:

• Contextualize Cyber Threats: Develop a more nuanced understanding of the characteristics and challenges associated with detecting cyberharassment, distinguishing it from general online discourse.

• Appreciate Social-Cybersecurity: Recognize the critical role of AI in addressing social cybersecurity issues [4, 5] and the importance of interdisciplinary approaches to solving such complex problems.

Heightened Awareness and Analytical Skills for AI Ethics and Bias

One of the most crucial outcomes is an improved ethical understanding and analytical capability regarding AI systems:

• Bias Identification: Students will gain practical experience in investigating and identifying potential biases (e.g., racial bias [11, 15]) within AI models. They will learn to critically analyze why certain groups of text or individuals might be disproportionately affected by a model's predictions.

• Ethical Implications: The lab fosters critical discussions and hands-on engagement with the ethical responsibilities of AI developers, particularly in sensitive domains like content moderation and online safety.

• Bias Mitigation Awareness: Students will be exposed to basic strategies for debiasing models [16, 17] and understand the trade-offs involved in balancing fairness and performance.

Practical Experience with Adversarial Attacks and Robustness

The module on adversarial attacks is designed to prepare students for real-world vulnerabilities of AI systems:

• Adversarial Thinking: Students will understand how adversaries might exploit AI models, specifically through the generation of adversarial text examples [12, 13, 14].

• Model Vulnerability Analysis: They will gain hands-on experience in generating simple adversarial attacks and assessing the robustness of their detection models, leading to a deeper appreciation of the challenges in building truly secure AI systems.

• Foundation for Defense: The exposure to attacks provides a foundational understanding necessary for later learning about defensive mechanisms and robust AI design.

Development of 21st-Century Skills

Beyond technical knowledge, the experiential nature of the lab contributes to broader skill development:

• Problem-Solving: Students engage in iterative problem-solving, debugging, and optimizing their AI models.

• Critical Thinking: They are challenged to critically evaluate data quality, model performance, and ethical implications, moving beyond simply running code [23].

• Data Literacy: Enhanced ability to work with and interpret complex textual data.

• Self-Efficacy: The hands-on, practical experience in building functional AI models for a relevant cybersecurity problem is expected to boost students' self-efficacy and confidence in tackling complex cyber-social challenges [35].

These outcomes collectively contribute to producing a new generation of cybersecurity professionals who are not only technically proficient in AI but also ethically aware and capable of addressing the complex socio-technical challenges of the digital age.

DISCUSSION

The designed AI-based cyberharassment detection laboratory exercise represents a vital advancement in cybersecurity and AI education. By deeply integrating technical AI concepts with the pressing social issue of online abuse, it bridges a critical gap in traditional curricula. The hands-on, experiential approach, building on established pedagogical frameworks [20, 24, 25, 26, 27, 30], ensures that learning is active, relevant, and

impactful. This contrasts with purely theoretical instruction, which, while foundational, may not fully prepare students for the complexities of real-world AI deployment and its ethical considerations [18, 19].

The direct engagement with actual or simulated cyberharassment data forces students to confront the nuances of online communication and the difficulties in automated content moderation. This experiential learning is crucial for developing not just technical competence but also a critical understanding of the social implications of AI systems. The explicit focus on bias analysis and adversarial attacks moves beyond mere model building, instilling a crucial awareness of AI's limitations and vulnerabilities [11, 12, 13, 14, 15, 16, 17]. This ethical dimension is increasingly recognized as paramount in AI education, particularly in applications that directly impact human rights and social equity. Students learn that an AI model is not merely a black box but a system whose output can have real-world consequences, demanding careful consideration of fairness and robustness. This aligns with broader efforts to integrate ethics into data science and AI curricula [31].

Challenges in Implementation: Despite the significant benefits, implementing such a lab is not without its challenges.

- Computational Resources: Training deep learning models like BERT [38] can be computationally intensive, requiring access to powerful GPUs, which might be a constraint for some educational institutions. Cloud-based platforms (e.g., Google Colab, AWS, Azure) can mitigate this, but still require proper management.

- Data Availability and Sensitivity: While public datasets exist (e.g., [8]), ensuring that the data used for the lab is ethically sourced, appropriately anonymized, and does not expose students to overly graphic content is crucial. The sensitivity of cyberharassment data requires careful curation.

- Instructor Expertise: Instructors need a strong background in both AI/NLP and cybersecurity, as well as an understanding of the social and ethical dimensions of online abuse. This interdisciplinary expertise might be a limiting factor in some departments [32, 33].

- Student Background Heterogeneity: Students may come from diverse academic backgrounds [21, 32], necessitating adaptable materials and support to ensure all participants can engage effectively with the technical and conceptual challenges.

- Keeping Pace with AI Advances: The field of AI is rapidly evolving, requiring the lab content and tools to be continuously updated to remain relevant.

Future Directions and Implications: Future iterations of this laboratory could expand in several ways. Incorporating more advanced debiasing techniques and exploring different adversarial attack vectors could deepen student understanding. Integrating defensive mechanisms against adversarial attacks would be a natural next step, allowing students to design more robust AI systems. Furthermore, the lab could be extended to include other forms of online malicious behavior, such as misinformation or online radicalization, broadening its scope within social-cybersecurity. Longitudinal studies could also track the long-term impact of such experiential learning on students' career choices and their approaches to ethical AI development.

This AI-based cyberharassment detection lab offers a compelling model for future cybersecurity and AI curricula. By combining rigorous technical training with critical ethical analysis and practical problem-solving, it prepares students not just to be skilled developers but also responsible architects of a safer digital future. The emphasis on hands-on application and critical thinking is essential for empowering the next generation of professionals to effectively combat the complex and evolving landscape of online abuse.

REFERENCES

[1] K. Thomas, D. Akhawe, M. Bailey, D. Boneh, E. Bursztein, S. Consolvo, N. Dell, Z. Durumeric, P. Kelley, D. Kumar, D. McCoy, S. Meiklejohn, T. Ristenpart, and G. Stringhini, “Sok: Hate, harassment, and the changing landscape of online abuse,” pp. 473–493, 2021.

[2] D. E. S. Swearer, “Research on school bullying and victimization: What have we learned and where do we go from here?” School Psychology Review, p. 365–383, 2013.

[3] S. Wachs, M. F. Wright, and A. T. Vazsonyi, “Understanding the overlap between cyberbullying and cyberhate perpetration: Moderating effects of toxic online disinhibition,” Criminal Behaviour and Mental Health, vol. 29, no. 3, pp. 179–188, 2019.

[4] “Center for Informed Democracy & Social-Cybersecurity,” https://www.cmu.edu/ideas-social-cybersecurity/.

[5] “Social-Cybersecurity,” http://www.casos.cs.cmu.edu/projects/projects/social_cyber_security.php.

[6] J. W. Patchin, “Cyberbullying Statistics,” https://cyberbullying.org/2019-cyberbullying-data, 2019.

[7] D. Ducharme, “Machine learning for the automated identification of cyberbullying and cyberharassment,”

Ph.D. dissertation, University of Rhode Island, 2017.

[8] "Hateful Memes Challenge and Data Set," https://ai.facebook.com/tools/hatefulmemes/.

[9] H. Zhong, H. Li, A. Squicciarini, S. Rajtmajer, C. Griffin, D. Miller, and Caragea, "Content-driven detection of cyberbullying on the instagram social network," p. 3952–3958, 2016.

[10] "AI advances to better detect hate speech," https://ai.facebook.com/blog/ai-advances-to-better-detect-hate-speech/.

[11] "Google's Hate Speech Detection A.I. Has a Racial Bias Problem," https://fortune.com/2019/08/16/google-jigsaw-perspective-racial-bias/.

[12] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," 2015. [Online]. Available: https://arxiv.org/abs/1412.6572

[13] J. Li, S. Ji, T. Du, B. Li, and T. Wang, "Textbugger: Generating adversarial text against real-world applications," 2019. [Online]. Available: https://www.ndss-symposium.org/ndss-paper/textbugger-generating-adversarial-text-against-real-world-applications/

[14] W. E. Zhang, Q. Z. Sheng, A. A. F. Alhazmi, and C. Li, "Adversarial attacks on deep-learning models in natural language processing: A survey," ACM Trans. Intell. Syst. Technol., vol. 11, no. 3, pp. 1–41, 2020.

[15] M. Sap, D. Card, S. Gabriel, Y. Choi, and N. A. Smith, "The risk of racial bias in hate speech detection," pp. 1668–1678, Jul. 2019. [Online]. Available: https://www.aclweb.org/anthology/P19-1163

[16] E. Okpala, L. Cheng, N. Mbwambo, and F. Luo, "Aaebert: Debiasing bert-based hate speech detection models via adversarial learning," in 2022 21st IEEE International Conference on Machine Learning and Applications (ICMLA). IEEE, 2022, pp. 1606–1612.

[17] V. Nanda, S. Dooley, S. Singla, S. Feizi, and J. P. Dickerson, "Fairness through robustness: Investigating robustness disparity in deep learning," CoRR, vol. abs/2006.12621, 2020. [Online]. Available: https://arxiv.org/abs/2006.12621

[18] J. C. Nwokeji, R. Stachel, and T. Holmes, "Effect of instructional methods on student performance in flipped classroom," pp. 1–9, 2019.

[19] S. Eybers and M. Hattingh, "Teaching data science to postgraduate students: A preliminary study using a" flip" classroom approach." International Association for Development of the Information Society, 2016.

[20] J. C. Nwokeji and P. S. T. Frezza, "Cross-course project-based learning in requirements engineering: An eight-year retrospective," pp. 1–9, 2017.

[21] Y. Velaj, D. Dolezal, R. Ambros, C. Plant, and R. Motschnig, "Designing a data science course for non-computer science students: Practical considerations and findings," pp. 1–9, 2022.

[22] R. Matovu, J. C. Nwokeji, T. Holmes, and T. Rahman, "Teaching and learning cybersecurity awareness with gamification in smaller universities and colleges," pp. 1–9, 2022.

[23] L. Samavedham and K. Ragupathi, "Facilitating 21st century skills in engineering students," The Journal of Engineering Education, vol. 26, no. 1, pp. 38–49, 2012.

[24] A. Y. Kolb and D. A. Kolb, "Learning styles and learning spaces: Enhancing experiential learning in higher education," Academy of management learning & education, vol. 4, no. 2, pp. 193–212, 2005.

[25] D. A. Kolb, Experiential learning: Experience as the source of learning and development. FT press, 2014.

[26] A. Barman, S. Chen, A. Chang, and G. Allen, "Experiential learning in data science through a novel client-facing consulting course," pp. 1–9, 2022.

[27] G. I. Allen, "Experiential learning in data science: Developing an interdisciplinary, client-sponsored capstone program," pp. 516–522, 2021.

[28] S. Rosenthal and T. Chung, "A data science major: Building skills and confidence," pp. 178–184, 2020.

[29] P. Anderson, J. Bowring, R. McCauley, G. Pothering, and C. Starr, "An undergraduate degree in data science: curriculum and a decade of implementation experience," pp. 145–150, 2014.

[30] E. Serrano, M. Molina, D. Manrique, and L. Baumela, "Experiential learning in data science: From the dataset repository to the platform of experiences," pp. 122–130, 2017.

[31] D. A. Martin and G. Bombaerts, "Enacting socio-technical responsibility through challenge based learning in an ethics and data analytics course," pp. 1–7, 2022.

[32] A. F. Salazar-Gomez, A. Bagiati, N. Minicucci, K. D. Kennedy, X. Du, and C. Breazeal, "Designing and implementing an ai education program for learners with diverse background at scale," pp. 1–8, 2022.

[33] H. A. Hashim, C. Tatarniuk, and B. Harasymchuk, "First year engineering design: Course design, projects, challenges, and outcomes," pp. 1–9, 2022.

[34] T. Lowe and C. Rackley, “Cybersecurity education employing experiential learning,” 2018.

[35] A. Konak, “Experiential learning builds cybersecurity self-efficacy in k-12 students,” Journal of Cybersecurity Education, Research and Practice, vol. 2018, no. 1, p. 6, 2018.

[36] C. M. B. Turner and C. F. Turner, “Analyzing the impact of experiential pedagogy in teaching socio-cybersecurity: Cybersecurity across the curriculum,” Journal of Computing Sciences in Colleges, vol. 34, no. 5, pp. 12–22, 2019.

[37] “Qualtrics,” https://www.qualtrics.com/

[38] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” arXiv preprint arXiv:1810.04805, 2018.