Developing AI-Driven Socially Relevant Cybersecurity Curriculum Through Collaboration

Long Cheng†
School of Computing
Clemson University
Clemson, SC 29634, USA
lcheng2@clemson.edu

Nishant Vishwamitra
Department of Information Systems and Cyber Security
The Univ. of Texas at San Antonio
San Antonio, TX, USA
nishant.vishwamitra@utsa.edu

Hongxin Hu
Department of Computer Science and Engineering
University at Buffalo, SUNY
Buffalo, New York, USA
hongxinh@buffalo.edu

Xiaohong Yuan
Department of Computer Science
North Carolina Agricultural and Technical State University
Greensboro, NC, USA
xhyuan@ncat.edu

Jeannette M. Wade†
Department of Sociology
North Carolina Agricultural and Technical State University
Greensboro, NC, USA
jmwade1@ncat.edu

Sajad Khorsandroo
Department of Computer Science
North Carolina Agricultural and Technical State University
Greensboro, NC, USA
skhorsandroo@ncat.edu

ABSTRACT

This paper presents the development of AI-Driven Socially Relevant Cybersecurity Curriculum through a collaborative project among Clemson University, North Carolina Agricultural and Technical State University (NC A&T), and the University at Buffalo. We describe the three hands-on labs on cyberbullying detection that we have developed, and discuss our experience of using one of the labs in computer science classes and social science classes.

CCS CONCEPTS

• Security and privacy

KEYWORDS

Artificial intelligence, socially relevant cybersecurity curriculum, cyberbully detection, adversarial attacks

1 Introduction

With the popularity of social media, cyber-harassment became a critical problem as it can have significant negative effects on targeted groups or individuals. In recent years, artificial intelligence has been applied to cyberbullying detection [1]. AI algorithms have been developed and deployed to detect toxic content on social media [2, 3]. Meanwhile, adversaries may exploit vulnerabilities of AI-based classifiers by generating adversarial content to evade existing cyberharassment detectors [4, 5, 6]. However, few educational materials have been designed to engage students by integrating AI and socially relevant cybersecurity through an interdisciplinary approach. Through a collaborative project funded by the NSF SaTC program, we are developing AI-driven socially relevant cybersecurity curricular modules and hands-on labs to engage students with diverse backgrounds. The hands-on labs have been taught in the computer science and social science departments in two universities: NC A&T and Clemson University.

2 AI/ML based Cyberharassment Detection Labs

The hands-on labs were developed on Google Colab platform. To provide students with pre-requisite knowledge, we developed: (1) a brief introduction of about the IPython-based Jupyter Notebook and Google Colab; (2) a broad presentation of general machine learning knowledge; and (3) an introduction about cyberharrassment and automated cyberharassment detection. For each lab, we develop a lab manual with two versions: one for computer science students and one for social science students. Computer science students are required to program some of the features for the labs, while social science students are not required to do such tasks.

The hands-on labs were developed on Google Colab platform. To provide students with pre-requisite knowledge, we developed: (1) a brief introduction of about the IPython-based Jupyter Notebook and Google Colab; (2) a broad presentation of general machine learning knowledge; and (3) an introduction about cyberharrassment and automated cyberharassment detection. For each lab, we develop a lab manual with two versions: one for computer science students and one for social science students. Computer science students are required to program some of the

features for the labs, while social science students are not required to do such tasks.

The three hands-on labs are described below.

Lab 1 AI for Text-based Cyberbully Detection

This lab teaches students how AI models can be used to distinguish between a cyberbullying and non-cyberbully text-based content. Students will learn data preprocessing, training, and the evaluation metrics of AI-based classifiers.

Lab 2 AI for multimodal (image and text) Cyberbully Detection

Cyberbullying can occur in images, and it can also occur in both image and text. Students will extract visual features from images and combine these features with textual features to detect cyberbullying in this lab.

Lab 3 Adversarial Attacks in Cyberbully Detection

AI models are vulnerable to adversarial attacks. In this lab, students will use different algorithms to generate images that can fool models trained to detect cyberbullying, causing the model to produce incorrect output.

3 Teaching Experiences

Lab1 has been used by computer science and social science students in two institutions (NC A&T and Clemson University) across two semesters: Spring 2022 and Fall 2022. Students were given a detailed lab manual and are to complete a set of tasks. Through this process, students learn AI concepts and the application of AI for cyber-harassment detection. Using pre- and post-surveys, we asked students to rate their knowledge or skills in AI and their understanding of the concepts learned. The results showed that the students moderately understood the AI-based cyber-harassment detection lab. In both semesters, computer science students improved their skills in automated cyber-harassment detection and state-of-the-art cyber-harassment detectors after participating in the lab.

The social science students showed significant improvement in their knowledge of how AI works, State-of-the-art cyber-harassment detectors, and automated cyber-harassment detection after a background lecture on the topic in the Fall semester compared to the Spring semester with no background lecture. We learn that having a background lecture before the labs improve understanding of the lab contents, especially for non-computer science students. These findings confirm that the developed lab is viable for teaching AI-driven socially relevant cybersecurity to computer science and non-computer science students and can be used by other institutions.

We have also taught Lab 1 in the GenCyber Summer Camps at NC A&T in 2021 and 2022. The students in the GenCyber summer camps are rising 8th to 12th grade.

4 Conclusion

In this paper we describe our approach of developing AI-driven socially relevant interdisciplinary Cybersecurity Curriculum through a collaborative project. We describe the three hands-on labs on cyberbullying detection that we have developed and discuss our experience of using one of the labs in computer science classes and social science classes. Our experience shows that providing a background lecture is very important for achieving the learning outcomes of the hand-on labs, especially for social science students. Our findings also demonstrate that it is viable to teach AI-driven socially relevant cybersecurity to computer science and non-computer science students and can be used by other institutions.

Future work includes developing hands-on labs on interpretability of AI models for cyberharassment detection, Disparity of AI-based Classifiers for Cyberharassment Detection, and Debiasing AI-driven Cyberharassment Detection Models.

ACKNOWLEDGMENTS

This work is supported in part by the National Science Foundation (NSF) under Grant No. XXXXXX. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of NSF.

REFERENCES

[1] Haoti Zhong, Hao Li, Anna Squicciarini, Sarah Rajtmajer, Christopher Griffin, David Miller, and Cornelia Caragea. Content-driven detection of cyberbullying on the instagram social network. In Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI'16, page 3952–3958, 2016.

[2] AI advances to better detect hate speech. https://ai.facebook.com/blog/ai-advances-to-better-detect-hatespeech/.

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

[4] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In Yoshua Bengio and Yann LeCun, editors, 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings, 2015.

[5] Jinfeng Li, Shouling Ji, Tianyu Du, Bo Li, and Ting Wang. Textbugger: Generating adversarial text against real-world applications. In 26th Annual Network and Distributed System Security Symposium, NDSS 2019, San Diego, California, USA, February 24-27, 2019. The Internet Society, 2019.

[6] Wei Emma Zhang, Quan Z. Sheng, Ahoud Abdulrahmn F. Alhazmi, and Chenliang Li. Adversarial attacks on deep-learning models in natural language processing: A survey. ACM Trans. Intell. Syst. Technol., 11(3):1–41, 2020.