Using Large Language Models for Social Cybersecurity Analyses*

Iain J. Cruickshank[1][0000−0002−4205−5806], Robert Thomson[1][0000−0001−9298−2870], and Nathaniel D. Bastian[1][0000−0001−9957−2778]

[1] Army Cyber Institute, West Point, NY
[2] {iain.cruickshank, robert.thomson, nathaniel.bastian}@westpoint.edu
https://cyber.army.mil/Research/Research-Labs/Cognitive-Security

Keywords: Large Language Models · Social Cybersecurity · Zero-shot Machine Learning

1 Background

Large language models (LLMs) have emerged as powerful tools in natural language processing, revolutionizing various, diverse fields. LLMs, such as OpenAI's GPT-series of models, have demonstrated exceptional capabilities in understanding and generating human-like text, driven by their ability to learn patterns and structures from vast amounts of training data. These advancements in AI have opened up new avenues for research and application development. In particular, these models can be used with minimal input data, and careful instructing (or prompting), to perform analyses that often required humans or substantial amounts of labeled and curated data.

At the same time, the field of Social Cybersecurity has emerged as an interdisciplinary field that explores the human aspects of cybersecurity, focusing on the interaction between individuals, technology, and the socio-technical systems in which they operate. Of particular note, the field of social cybersecurity has generated several insights and techniques for analyzing online data, particularly for countering mis- and disinformation. In this workshop, we will present how to harness the power of LLMs in order to augment social cybersecurity workflows. The use of these new models — and paradigms of working with AI — allow for new types of analyses as well as the ability to quickly run more traditional types of social cybersecurity analyses.

* This work was conducted within the Cognitive Security Research Lab at the Army Cyber Institute at West Point and supported in part by the Office of Naval Research (ONR) under Support Agreement No. USMA 20057. The views expressed in this paper are those of the authors and do not reflect the official policy or position of the United States Military Academy, the United States Army, the Department of Defense, or the United States Government.

2 Demonstration

For this tech demonstration, we will demonstrate how to use open-source, freely available LLMs and Python software packages to work through a social cybersecurity workflow. Specifically, we will demonstrate how open-source LLMs can be used to do a series of tasks with social media data to understand the nature of comments in the social media space and classify those comments relative to social cybersecurity concepts. Specifically, we will demonstrate the following:

Analyze a space of social media posts: Given a collection of social media posts, we will show how you can use packages like BERTopic, combined with an LLM, to break a corpus of text into semantically similar components (i.e. topic modeling) and then describe those topics in short, human-meaningful descriptions (as opposed to the most probable words per topic). We will further show how you can incorporate additional information into the prompt for the LLM to produce more meaningful summaries of the comments (e.g., using a prompt of "produce a short summary of the following comments *as they relate to [ITEM]*").

Zero-shot labeling social media comments: Given the comments and information about them from the previous step, we will then demonstrate how one can use the HuggingFace software package along with Accelerate in order to label the comments based on a custom set of labels, without any supervision of the machine learning model. These labels can also be expanded to short phrases instead of the traditional one-word labels for additional capability in the analysis.

Prompt engineering for classifying the stance social media comments: Then we will show how an LLM can be used for more complex text labeling and identification tasks. In particular, we will show how an analyst can use prompting techniques, like few-shot prompting and reasoning prompting in order to label the stance of the comments towards a particular topic or entity. The use of prompting combined with an LLM enables what is otherwise a very difficult or labor-intensive type of social cybersecurity analysis (e.g., stance labeling) to be done at scale and with speed.

Using a Language-Vision Model to analyze user profiles: Finally, we will show how language vision models (LVMs) can be combined with software like BERTopic to cluster together user profile images and describe the user profile text at the same time. For example in a recent analysis of Parler data, we have found a cluster of images of lions that frequently use terms like "Patriot" and "Party" to describe their accounts. This type of analysis allows for cross-model analyses of social media data, which previously required humans to perform.

Overall, we will demonstrate how new breakthroughs in LLM models can be easily and cheaply applied to social cybersecurity workflows. We hope this tech demo inspires future work of applying AI to social cybersecurity and better social cybersecurity tools to defeat mis- and disinformation.