Applications and Risks of Language Models for National Security

Introduction

This technology review discusses recent findings regarding how statistical language models and large language models can be used to not only assist the U.S. intelligence community (IC) but to also create significant national security challenges that the IC must solve. This paper will review two possible applications that language models have in IC-related work, including potential national security concerns that could arise from increasing deployment of these same models.

Using Large Language Models to Generate Intelligence

The use of NLP techniques and tools, including large language models (LLMs), is in its infancy in the national security domain and "decisions made based on the insights provided could have significant consequences for individuals and wider society" (C & Carter, 2023). Currently, the involvement of LLMs in national security is mainly limited to research about the feasibility of integrating these models into everyday tasks in the domain and determining best practices for doing so (Gallagher et. al., 2023).

One possible usage for LLMs within the IC is summarizing large bodies of intelligence text to create briefs. Like many other professionals, IC experts must comb through vast amounts of text data to generate reports and make informed decisions. LLMs can "in a self-supervised manner, extract knowledge from massive corpora of information" to improve productivity and efficiency (C & Carter, 2023). Furthermore, LLMs can cover greater amounts of data than a human researcher in the IC can, which would enable these intelligence reports to be more complete and potentially more effective for important decision-making purposes. These models can not only extract summaries of knowledge from text data, but they can also discover and interpret patterns unknowable to human researchers from directly and indirectly related non-text data using contextual text mining techniques.

However, some important caveats exist to this intelligence summarization and generation use case. Authentication and truthfulness are two vital qualities of national security intelligence reports since decisions of utmost importance, including matters of life and death, are made using these outputs. A common critique of recent LLMs is that they are "prone to factual errors, hallucinations (i.e., fabrication of new information), overconfidence, and susceptibility to adversarial effects" (Gallagher et. al., 2023). A LLM's ability to generate truthful information generally is tested by its ability to answer factual questions with singular answers (e.g., a model's ability to pass a standardized exam). However, intelligence questions do not usually have a single, correct answer that is pre-known (Gallagher et. al., 2023).

To combat this occasionally weakness of tending towards falsifications, LLMs could be forced to list sources used to generate outputs, enabling a human fact-checker to validate the information (C & Carter, 2023). Depending on the number of sources utilized, this manual process may be quick or cumbersome. Additionally, training these models on limited text data instead of large amounts of open-source information could limit the models' access to falsehoods; however, this training technique might invalidate the original purpose of the model, which would be to create intelligence from a vast variety and origin of sources. The open-source category of text data comprises an important piece of IC resources.

Disinformation and Phishing Detection

While the truthfulness of LLM outputs is a cause for hesitation when introducing these NLP tools to use by the IC, on the other hand, language models can be used to help detect falsehoods, providing a second possible application for these models within the national security domain.

Online deception is a concerning cause of public unrest, and it can even interfere with the processes vital to national security. For example, in 2016, "Russian operators executed a campaign to influence the U.S. presidential election [including] efforts to discredit democratic institutions, sow discord, and undermine public trust. […] The Kremlin has waged increasingly brazen disinformation campaigns – operations meant to intentionally spread false or misleading information" (Sedova et. al., 2021). Humans have a poor ability to detect deception and sift through overwhelming amounts of text data, whereas many NLP techniques exist that excel at these tasks, so such a tool could possibly be derived from common statistical language models (SLMs) or LLMs (Zhou, Shi, & Zhang, 2008).

As early as 2008, researchers Zhou, Shi, and Zhang discovered that a "SLM approach to deception detection outperformed a state-of-the-art text categorization method, as well as traditional feature-based methods." The tokens in the model that these researchers developed represent punctuation characters in addition to the traditional words and phrases because punctuation marks can indicate deceivers' rhetoric strategies and help extract deception cues. To decrease complexity and effectively implement Kneser-Ney smoothing (to avoid assigning zero probability to unseen n-grams), the researchers also used stemming and removed words with single occurrences. When evaluating their SLM, the researchers used accuracy, precision, and recall with tenfold cross validation over six datasets consisting of instant messages and emails. On this data, their SLM outperformed basic maximum likelihood estimator models and a support vector machine. In particular, the SLM was "found to achieve a good balance of deception precision and deception recall, mitigating potential biases toward either truth (for example, like humans) or deception (for example, like some feature-based techniques)" (Zhou, Shi, & Zhang, 2008).

More recently, a researcher at the Ivan Franko National University of Lviv in Ukraine examined the possibility of using the Llama 2 LLM to detect disinformation and fake news. Using a Parameter-Efficient Fine-Tuning/Low Rank Adaptation approach, Pavlyshenko demonstrated that the Llama 2 model can "perform a deep analysis of texts and reveal complex styles and narratives" for the purposes of "revealing disinformation and propaganda narratives, fact checking, fake news detection, manipulation analytics, [and] extracting named entities with their sentiments" (Pavlyshenko, 2023). As in the case of summarizing intelligence sources and extracting knowledge from related patterns, accuracy is important and simultaneously challenging. Pavlyshenko found that accuracy can be improved by utilizing more trustworthy training data and providing iterative, corrective feedback to the model.

The ability of LLMs and SLMs to generate text quickly and cheaply for spear phishing campaigns is well-known and a great cause for concern, particularly among the national security community due to potentially high-profile targets of these phishing campaigns (Hazell, 2023). At the same time, language models could help detect and prevent these types of cybersecurity attacks. One strategy for deploying LLMs against phishing campaigns is to train on previous, trusted messages and then have the LLM evaluate new messages for “inconsistencies in writing style or flag suspicious email addresses, making it easier for users to notice potential threats” (Hazell, 2023). The lack of attention span and ignorance traits that make humans susceptible to phishing attacks are not vulnerabilities of LLMs; LLMs can methodically scrutinize each incoming message for suspicious signals.

I believe that these models can be used for good in another way: generating text data quickly and cheaply for training programs (for humans) that aim to combat the spread of disinformation and susceptibility to similar threats, such as phishing. LLMs and SLMs could enable well-meaning practitioners to educate others effectively and efficiently by creating realistic examples, thereby decreasing the potency of ill-intentioned users aiming to spread disinformation and implement cyber-attacks.

Conclusion

While the increasing prevalence of NLP language models may cause concern among those in the national security community, I believe that with careful training and modifications, these models might ameliorate many of the problems caused by bad actors with access to the same technology. LLMs and SLMs have proven effective against the spread of disinformation and some forms of cyberattacks that involve social engineering. Additionally, these models can add capabilities to the IC that could improve productivity, coverage, and the effectiveness of intelligence work.

~ Anna Otte

References

C, A. & Carter, R. (2023, Jul.). Large language models and intelligence analysis. *Centre for Emerging Technology and Security Expert Analysis*. https://cetas.turing.ac.uk/sites/default/files/2023-07/cetas_expert_analysis_-_large_language_models_and_intelligence_analysis.pdf

Gallagher, S. et. al. (2023, Sept.). A retrospective in engineering large language models for national security. *Carnegie Mellon University Software Engineering Institute*. https://insights.sei.cmu.edu/documents/5752/6125_Retrospective_in_Engineering_Large_Language_Models_for_National_Security__vzeyF95.pdf

Hazell, J. (2023, May 12). *Large language models can be used to effectively scale spear phishing campaigns*. arXiv.org. https://arxiv.org/abs/2305.06972

Pavlyshenko, B. (2023, Sept. 12). *Analysis of disinformation and fake news detection using fine-tuned large language model.* arXiv.org. https://doi.org/10.48550/arXiv.2309.04704

Sedova, K. et. al. (2021, Dec.). AI and the future of disinformation campaigns. *Center for Security and Emerging Technology*. https://cset.georgetown.edu/publication/ai-and-the-future-of-disinformation-campaigns/

Zhou, L., Shi, Y., & Zhang, D. (2008, Aug.). A Statistical Language Modeling Approach to Online Deception Detection. *IEEE Transactions on Knowledge and Data Engineering*, vol. 20, no. 8, pp. 1077-1081. https://doi.org/10.1109/TKDE.2007.190624