MINISTERUL EDUCAȚIEI ȘI CERCETĂRII AL REPUBLICII MOLDOVA
Universitatea Tehnică a Moldovei
Facultatea Calculatoare, Informatică și Microelectronică
Department of Software and Automation Engineering

Admis la susținere
Șef departament:
Fiodorov Ion, dr., conf. univ.

_____
„ ___ ” _____ 2024

Investigarea atacurilor cibernetice bazate pe exploatarea modelelor lingvistice în contextul inteligenţei artificiale

Investigating cyber attacks based on the exploitation of linguistic models in the context of artificial intelligence

Proiect de master

Student: _____ Todos Alexandru, SI-221M
Coordonator: _____ Bulai Rodica, univ. lect.
Consultant: _____ Bulai Rodica, univ. lect.

Chișinău, 2024

Rezumat

Todos Alexandru, "Investigarea atacurilor cibernetice bazate pe exploatarea modelelor lingvistice în contextul inteligenţei artificiale", teza de masterat, Chișinău, 2024

Structura tezei. Teza cuprinde 53 de pagini și include introducere, 4 capitole, concluzii, o bibliografie de 19 de surse și 24 de figuri.

Cuvinte cheie: inteligență artificială, crimă cibernetică, securitatea datelor, limbaj natural, rețele neurale

Domeniul de studiu: Aspectele teoretice și practice ale Inteligenței Artificiale și a modelelor lingvistice, în vederea elaborării unui studiu de caz privind folosirea IA în scopuri abuzive de către răufăcătorii cibernetici.

Obiectivele studiului: 1. Studierea fenomenului Inteligenței Artificiale și a etapelor sale de evoluție. 2. Afișarea stării prezente de utilizare în scopuri abuzive și malicioase a IA pentru profitul infractorilor în domeniul cibernetic. 3. Prezentarea potențialelor scenarii de viitor a utilizării instrumentelor de Inteligență Artificială pentru efectuarea atacurilor cibernetice. 4. Analiza impactului atacurilor cibernetice cu folosirea tehnologiilor IA asupra busines-ulurilor, consumatorilor. Prezentarea dificultăților ce vor apărea în viitor în domeniul securității informaționale. Specularea asupra intențiilor răufăcătorilor și prezentarea unei liste de recomandări contra atacurilor cu folosirea de tehnologii IA.

Ipoteza de cercetare. Analiza acestei teme și prezentarea unor cazuri reale a folosirii IA în scopuri malițioase ar putea scoate în evidență problema discutate, prezență potențialele riscuri pentru organizații și promovarea IA ca instrument pentru a ajuta dezvoltatorii software în crearea lucrilor bune.

Valoarea aplicativă a lucrării. Studiul efectuat poate ajuta organizațiile ce stau în spatele tehnologiilor IA în modificarea algoritmelor acestora, astfel în cît ele să fie cu mult mai greu de utilizat pentru infractorii cibernetici. La fel, acesta poate ajuta diverse organizații în implimentarea politicilor și tacticelor de apărare contra atacuri cibernetice cu folosirea IA.

Noutatea și originalitatea. Deși lucrarea se bazează pe unele studii efectuate anteriror de alți experți în domeniul de securitatea, teza prezintă atacurile ce sunt deja posibile și atacuri cibernetice ce ar putea fi posibile în viitor. Aceste lucruri au fost deduse doar în baza cunoștințelor și posibilităților prezente astăzi.

Abstract

Todos Alexandru, "Investigating cyber attacks based on the exploitation of linguistic models in the context of artificial intelligence", master thesis, Chișinău, 2024

Thesis structure. The thesis contains 53 pages and includes an introduction, 4 chapters, conclusions, a bibliography of 19 sources and 24 figures.

Keywords: artificial intelligence, cybercrime, data security, natural language, neural networks

Field of study: Theoretical and practical aspects of Artificial Intelligence and linguistic models to develop a case study on the misuse of AI by cyber criminals.

Objectives of the study: 1. To study the phenomenon of Artificial Intelligence and its stages of evolution. 2. To show the present state of misuse and malicious use of AI for the profit of cyber criminals. 3. Presenting potential future scenarios of the use of AI tools to carry out cyber attacks. 4. Analysis of the impact of cyber attacks using AI technologies on businesses, consumers. Presentation of future challenges in the field of information security. Speculating on the intentions of the perpetrators and presenting a list of recommendations against attacks using AI technologies.

Research hypothesis. Analysis of this topic and presentation of real cases of using AI for malicious purposes could highlight the discussed issue, the presence of potential risks for organizations and promotion of AI as a tool to help software developers in creating good things.

Application value of the paper. The study conducted can help organizations behind AI technologies in modifying their algorithms so that they are much harder to use for cyber criminals. Likewise, it can help various organizations in implementing policies and tactics to defend against cyber attacks using AI.

Novelty and originality. Although the paper is based on some previous studies by other security experts, the thesis presents attacks that are already possible and cyber attacks that could be possible in the future. These have only been deduced based on the knowledge and possibilities present today.

Contents

LIST OF ACRONYMS:

AGI – Artificial General Intelligence

AI – Artificial Intelligence

API – Application Programming Interface

BERT – Bidirectional Encoder Representations from Transformers

Captcha – Completely Automated Public Turing test to tell Computers and Humans Apart

DALL-E – Text-to-image generation AI model

ELMo – Embeddings from Language Model

ENISA – European Network and Information Security Agency

ETSI – European Telecommunications Standards Institute

FairSeq – Facebook AI Research Sequence-to-Sequence Toolkit

GAN – Generative Adversarial Network

GitHub – Version control and collaboration platform

GPT – Generative Pre-trained Transformer

HFT – High-Frequency Trading

HHM – Hidden Markov Models

IoT – Internet of Things

Java – Programming language and development environment

KYC – Know Your Customer

LLM – Large Language Model

LSTM – Long Short-Term Memory

MFA – Multi-factor Authentication

ML – Machine Learning

NER – Named-Entity Recognition

NLP – Natural Language Processing

NLG – Natural Language Generation

OSINT – Open-Source Intelligence

Python – Programming language and development environment

RNN – Recurrent Neural Networks

SQL – Structured Query Language

URL – Uniform Resource Locator

INTRODUCTION

Large language models have emerged as critical tools with unmatched capabilities in the constantly evolving field of artificial intelligence, changing how we interact with, analyze, and generate massive amounts of textual data. These models, that are based on deep learning architectures, have an incredible ability to understand, explain, and generate human-like language that mimics the rich complexities of natural language.

While AI and ML algorithms could offer significant benefits to society, they may also create a variety of digital, physical, and political risks. Despite the fact that these innovations have transformed several sectors, like natural language processing, content production, and personal assistance, they also raise serious concerns about the possible exploitation and abuse of this technology. Large language models, if placed in the wrong hands or poorly managed, have the potential to become powerful cyber threats. This also applies to such a popular tool as ChatGPT, that has been valuable to developers everywhere, even criminals.

This research project looks into the complexities of huge language models and investigates the potential security threats they present. The subject examines the vulnerabilities and techniques by which malicious individuals might use these models to plan and carry out cyber attacks, as well as manipulate information to manipulate persons and systems. It aims to bring information about the delicate balance required for using the potential of these models for good, while protecting against exploitation and encouraging responsible use within ethical constraints. Please keep in mind that the information supplied is only for general knowledge and educational purposes.

This research will use a variety of approaches, combining qualitative and quantitative methodologies. An in-depth investigation of existing facts will serve as the basis, diving into research publications, case studies, and industry reports to understand the AI-driven cyber-attacks. Additionally, real-life tests and simulations will be performed to verify hypotheses and develop potential attack scenarios.

This investigation also looks at current approaches to safer AI, such as preventing AI from being utilized for organizing cyberattacks or preventing attacks against AI-based mechanisms and tools. AI systems may be exposed to threats as a result of their own vulnerabilities.

The mission of the master's research is to examine the developing environment of large language models and raise awareness about their potential security vulnerabilities by following the following objectives:

- **analyze** security risks posed by large language models and **identify** vulnerabilities that could be exploited by hackers;
- **highlight** the consequences of big language models being abused for cyber assaults, misinformation campaigns, and other criminal acts, highlighting the potential harm to individuals, companies, and society at large;
- **promote** ethical and responsible use of large language models;
- **develop** strategies and security measures aimed at minimizing the identified risks and increasing the resilience of large language models against potential abuse;
- **guide** and serve as a foundational reference for future research and development in the field of AI.

In addition to studying the present state of AI technologies, this study tries to predict the potential future use of these technology by criminals. This is a challenging task, but one that the cybersecurity industry and law enforcement must take on in their never-ending effort to remain one step ahead of the criminal world.

The results of this research will provide an exhaustive understanding of the difficulties of cyber-attacks using AI language models. It aims to contribute to the intellectual discussion on cybersecurity by providing practical information and strategic recommendations for improving AI-driven linguistic models against exploitation, in order to protect digital infrastructures.

BIBLIOGRAPHY

1. B. H. Choo, «The emergence of Large Language Models (LLMs) - The Low Down - Momentum Works», 23 March 2023. Available at: https://thelowdown.momentum.asia/the-emergence-of-large-language-models-llms/ [Viewed: 10 September 2023]

2. O. Calzone, «An Intuitive Explanation of LSTM», *Medium*, 10 April 2022. Available: https://medium.com/@ottaviocalzone/an-intuitive-explanation-of-lstm-a035eb6ab42c. [Viewed: 10 September 2023]

3. Z. Aston, «10.1. Long Short-Term Memory (LSTM) — Dive into Deep Learning 1.0.3 documentation», 7 december 2023, Available at: https://d2l.ai/chapter_recurrent-modern/lstm.html. [Viewed: 10 September 2023]

4. M. Brundage, K. Mayer, "Lessons Learned on Language Model Safety and Misuse.", 3 March 2023. Available at: https://openai.com/research/language-model-safety-and-misuse [Viewed: 10 September 2023]

5. N. Carlini, «Extracting Training Data from Large Language Models», *arXiv.org*, 14 December 2020 Available at: https://arxiv.org/abs/2012.07805v2. [Viewed: 10 September 2023]

6. S. Palka, D. McCoy, «Fuzzing E-mail Filters with Generative Grammars and {N-Gram} Analysis», 2015. Available at: https://www.usenix.org/conference/woot15/workshop-program/presentation/palka. [Viewed: 10 September 2023]

7. «Weaponizing Machine Learning: Humanity Was Overrated Anyway», *Bishop Fox*. Available at: https://bishopfox.com/resources/weaponizing-machine-learning-humanity-was-overrated-anyway. [Viewed: 10 September 2023]

8. «Internet Organised Crime Threat Assessment (IOCTA) 2020», *Europol*. Available at https://www.europol.europa.eu/publications-events/main-reports/internet-organised-crime-threat-assessment-iocta-2020. [Viewed: 28 October 2023]

9. ThoughtfulDev, EagleEye [software]. 15 February 2018. Available at https://github.com/ThoughtfulDev/EagleEye [Viewed: 28 October 2023]

10. CorentinJ, Real-Time-Voice-Cloning [software]. 3 July 2019. Available at https://github.com/CorentinJ/Real-Time-Voice-Cloning [Viewed: 28 October 2023]

11. S. Hilt “The Sound of a Targeted Attack”, 27 August 2017, Available at: https://documents.trendmicro.com/assets/pdf/The-Sound-of-a-Targeted-Attack.pdf [Viewed: 28 October 2023]

12. «A „silly“ attack made ChatGPT reveal real phone numbers and email addresses», *Engadget*, 29 november 2023 Available at: https://www.engadget.com/a-silly-attack-made-chatgpt-reveal-real-phone-numbers-and-email-addresses-200546649.html. [Viewed: 28 October 2023]

13. T. Simonite “OpenAI’s Text Generator Is Going Commercial.”, Wired, 11 June 2020, Available at: https:// www.wired.com/story/openai-text-generator-going- commercial/ [Viewed: 28 October 2023]

14. «ESET Whitepaper: Can Artificial Intelligence Power Future Malware?» Available at https://www.eset.com/me/whitepapers/can-artificial-intelligence-power-future-malware/. [Viewed: 15 December 2023]

15. H. Zhang, «StackGAN: Text to Photo-realistic Image Synthesis with Stacked Generative Adversarial Networks». arXiv, 4 August 2017 doi: 10.48550/arXiv.1612.03242. Available at: http://arxiv.org/abs/1612.03242. [Viewed: 15 December 2023]

16. «Wannacry Ransomware», *Europol*. Available at https://www.europol.europa.eu/wannacry-ransomware. [Viewed: 15 December 2023]

17. M. Rajpal, W. Blum, R. Singh, «Not all bytes are equal: Neural byte sieve for fuzzing». arXiv, 9 November 2017 doi: 10.48550/arXiv.1711.04596. Available at: http://arxiv.org/abs/1711.04596. [Viewed: 15 December 2023]

18. L. H. Newman, «AI Wrote Better Phishing Emails Than Humans in a Recent Test», *Wired*. Available at: https://www.wired.com/story/ai-phishing-emails/. [Viewed: 15 December 2023]

19. C. T. Thanh, I. Zelinka, «A Survey on Artificial Intelligence in Malware as Next-Generation Threats», Mendel, December 2019, doi: 10.13164/mendel.2019.2.027. Available at: https://mendel-journal.org/index.php/mendel/article/view/105. [Viewed: 15 December 2023]