Міністерство освіти і науки України
Київський національний університет імені Тараса Шевченка

Факультет інформаційних технологій
Кафедра кібербезпеки та захисту інформації

ДОПУСТИТИ ДО ЗАХИСТУ:
В.о. завідувача кафедри
кібербезпеки
та захисту інформації
_____ Іван ПАРХОМЕНКО
«17» травня 2024 р.

ПОЯСНЮВАЛЬНА ЗАПИСКА

кваліфікаційної роботи

галузь знань *12 Інформаційні технології*
(шифр і назва галузі знань)

спеціальність *125 Кібербезпека*
(код і назва спеціальності)

освітній ступень *магістр*

освітньо-наукова програма *Кібербезпека*
(назва освітньої програми)

на тему: «Метод виявлення кібератак, які використовують штучний інтелект та соціальну інженерію»

Виконавець: студент ІІ курсу, групи КБм-21

_____ **Роман КІНДЕРИСЬ**
(підпис) (Ім'я, ПРІЗВИЩЕ)

	Ім'я, ПРІЗВИЩЕ	**Підпис**
Науковий керівник	Олександр ЛАПТЄВ	
Нормоконтроль	Юрій БАБЕНКО	

Київ 2024

Міністерство освіти і науки України
Київський національний університет імені Тараса Шевченка

Факультет інформаційних технологій
Кафедра кібербезпеки та захисту інформації

ЗАТВЕРДЖЕНО:
В.о. завідувача кафедри
кібербезпеки
та захисту інформації
_____ Іван ПАРХОМЕНКО
«17» листопада 2023 р.

ЗАВДАННЯ
на виконання кваліфікаційної роботи

спеціальності *125 Кібербезпека*
(код і назва спеціальності)

освітній ступень *магістр*

Здобувача КБм-21 (група) Кіндерися Романа Андрійовича (прізвище ім'я по-батькові)

Тема кваліфікаційної роботи Метод виявлення кібератак, які використовують штучний інтелект та соціальну інженерію

1. ПІДСТАВИ ДЛЯ ПРОВЕДЕННЯ РОБОТИ

Рішення засідання кафедри кібербезпеки та захисту інформації факультету інформаційних технологій протокол № 5 від 15.11.2023 р.

2. МЕТА ТА ВИХІДНІ ДАНІ ДЛЯ ПРОВЕДЕННЯ РОБІТ

Об'єкт досліджень	Процес виявлення кібератак які використовують штучний інтелект та соціальну інженерію
Предмет досліджень	Метод виявлення кібератак які використовують штучний інтелект та соціальну інженерію
Мета	Підвищення ефективності виявлення та блокування кібератак які використовують штучний інтелект та соціальну інженерію

Вихідні дані для проведення роботи	Методи захисту від кібератак, які використовують штучний інтелект в атаках соціальної інженерії

3. ОЧІКУВАНІ НАУКОВІ РЕЗУЛЬТАТИ

Наукова новизна	Удосконалення методу виявлення кібератак, які використовують соціальну інженерію та штучний інтелект та розробка класифікації цих атак
Практична цінність	Реалізація удосконаленого методу дозволить підвищити ефективність виявлення та блокування кібератак які використовують штучний інтелект та соціальну інженерію

4. ВИМОГИ ДО РЕЗУЛЬТАТІВ ВИКОНАННЯ РОБОТИ

Робота виконана у повному обсязі відповідно до теми.

5. ЕТАПИ ВИКОНАННЯ РОБОТИ

Найменування етапів робіт	**Строки виконання робіт (початок-кінець)**
Уточнення постановки задачі	17.11.2023 – 29.01.2024
Аналіз літературних джерел	30.01.2024 – 12.02.2024
Обґрунтування вибору рішення	13.02.2024 – 21.02.2024
Виконати аналіз методів кібератак які використовують штучний інтелект та соціальну інженерію.	22.02.2024 – 26.02.2024
Провести дослідження кібератак які використовують штучний інтелект та соціальну інженерію	27.02.2024 – 04.03.2024
Метод виявлення кібератак які використовують штучний інтелект та соціальну інженерію	05.03.2024 – 10.03.2024
Метод виявлення кібератак які використовують штучний інтелект та соціальну інженерію	11.03.2024 – 17.03.2024
Апробація роботи на науково-методичному семінарі	18.03.2024 – 19.03.2024
Розробка рекомендацій щодо виявлення та блокування кібератак які використовують штучний інтелект та соціальну інженерію	20.03.2024 – 17.04.2024
Оформлення та друк пояснювальної записки	18.04.2024 – 25.04.2024

Найменування етапів робіт	Строки виконання робіт (початок-кінець)
Оформлення презентацій та отримання рецензій	26.04.2024 – 12.05.2024
Подача пакету документів на розгляд ЕК	13.05.2024 – 18.05.2024

6. РЕАЛІЗАЦІЯ РЕЗУЛЬТАТІВ ТА ЕФЕКТИВНІСТЬ

Економічний ефект Зменшення витрат на методи захисту від кібератак, які використовують штучний інтелект та соціальну інженерію

Соціальний ефект Підвищення ефективності виявлення та блокування кібератак які використовують штучний інтелект та соціальну інженерію.

7. ДОДАТКОВІ ВИМОГИ

Завдання видав _____ Олександр ЛАПТЄВ
(підпис) (Ім'я, ПРІЗВИЩЕ)

Завдання прийняв до виконання _____ Роман КІНДЕРИСЬ
(підпис) (Ім'я, ПРІЗВИЩЕ)

Дата видачі завдання: 17.11.2023 р.
Термін подання кваліфікаційної роботи до ЕК 17.05.2024 р.

РЕФЕРАТ

Обсяг роботи 82 сторінки., 49 рисунків і 51 джерело літератури.

Об’єкт дослідження – процес виявлення кібератак які використовують штучний інтелект та соціальну інженерію.

Метою дипломної роботи є підвищення ефективності виявлення та блокування кібератак які використовують штучний інтелект та соціальну інженерію.

У процесі вирішення поставлених завдань у дипломній роботі були використані: методи аналізу, спостереження, індукції та моделювання.

У роботі досліджено процес виявлення кібератак які використовують штучний інтелект та соціальну інженерію; типи та основні сценарії таких атак, реалізовано симуляцію атаки та метод її виявлення; розроблено класифікацію атак та рекомендації їх виявлення для користувачів.

Практичне значення роботи полягає у розробці класифікації атак соціальної інженерії із використанням штучного інтелекту та у розробці методу їх виявлення.

Результати досліджень можуть бути використані для підвищення стану захищеності користувачів від атак соціальної інженерії.

Наукова новизна дослідження полягає у:

- Розробці методу виявлення атак соціальної інженерії із використанням штучного;
- розробці класифікації атак соціальної інженерії із використанням штучного інтелекту

Одним із напрямків подальших досліджень є розширення створеної класифікації, а також вдосконалення методу виявлення шляхом аналізу вкладень, а також підключення ще однієї моделі штучного інтелекту, яка аналізуватиме текст повідомлень на наявність в них соціальної інженерії.

Ключові слова: чутлива інформація, соціальна інженерія, атака, штучний інтелект, загрози, ризики, вразливість, програмне забезпечення із відкритим кодом.

ЗМІСТ

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ ТА СКОРОЧЕНЬ

БД – база даних

ОС – операційна система

ПЗ – програмне забезпечення

ШПЗ - шкідливе програмне забезпечення

OSINT – Open Source Intelligence

SAT – Security Awareness Training

SMS – Short Message Service

URL – Uniform Resource Locator

ШІ – штучний інтелект

СІ – соціальна інженерія

API – application programming interface

YAML - Yet Another Markup Language

YARA – Yet Another Hilarious Acronym

TN – True negative

TP – True positive

IP – Internet Protocol

DMARC - Domain-based Message Authentication, Reportingand Conformance

SPF- Sender Policy Framework

DKIM - Domain Keys Identified Mail

GSM - Global System for Mobile Communications

SPAM - sales promotional advertising mail

DNS - Domain Name System

RSA - Rivest, Shamir and Adleman

ВСТУП

Інформація завжди була неймовірно цінним ресурсом. З давніх часів люди ставились до знань як до скарбу, який потрібно ретельно оберігати, обмежувати до нього доступ, захищати шляхом впровадження різного роду механізмів безпеки. Із процесом еволюції людства інформація як актив, також змінювалась, трансформувались її форми та обсяги, формати зберігання і передавання. Завдяки науково-технічному прогресу, в наші дні інформація найчастіше зберігається в оцифрованому вигляді і на те є низка своїх причин. Очевидно, це набагато зручніше, коли доступ є у всіх необхідних суб'єктів одночасно, коли не потрібно витрачати великі обсяги часу на передачу, копіювання та поширення, тощо. Проте, ця зручність не позбавила людей від притаманних цьому активу ризиків. Як наприклад, несанкціонований доступ, втрата доступу, втрата власне кажучи самого активу. Тому, завдання забезпечення стану захищеності є актуальним і сьогодні, а враховуючи темпи росту обсягів інформації воно залишатиметься таким завжди.

Багато людей помилково вважають, що їх інформаційні активи захищати не потрібно, часто відповідаючи на застереження фахівців із інформаційної безпеки репліками типу «Та кому взагалі можуть стати в нагоді мої данні?» На це питання є однозначна відповідь – зловмисникам. Прикладом, яким можна довести людям неправильність такої точки зору є кошти, точніше їх цифровий еквівалент, тому що складно зараз знайти людину, в якої немає відкритого особистого рахунку в банку, який окрім того що передбачає наявність на ньому грошей, також вимагає від суб'єкта якому належить, надати установі низку персональних даних(дату народження, ідентифікаційний код, паспортні дані, номер телефону та багато іншого). Така інформація вважається чутливою (з англ. sensitive data). Під такими даними розуміють інформацію при несанкціонованому доступі до якої, або її втраті суб'єкт, якому вона належить, буде поставлений в становище втрати певних переваг та збільшить ймовірність настання ризиків інформаційної безпеки, які можуть призвести до небажаних наслідків «Чи готові ви розпрощатись із всіма своїми коштами, дати

кіберзлочинцям оперувати ними від вашого імені, розпоряджатись вашою особистістю як своєю?». Ці питання є риторичними.

Соціальні інженери у зловмисних мотивах використовують різні методи для знаходження та отримання доступу до такої інформації. Найчастіше шляхом розвідки та аналізу в мережі Інтернет. На жаль, користувачі у зв'язку з невисоким рівнем освіченості та брак знань у сфері кібергігієни можуть добровільно поширювати таку інформацію, що в свою чергу дозволяє зловмисникам витрачати менше часу на онти підготування та експлуатацію атак. Також, останнім часом все більш популярним стає використання технологій штучного інтелекту для здійснення кібератак. Здебільшого такі атаки спрямовані на користувачів, оскільки загальноприйнято вважати саме людей найбільш вразливою ланкою будь-якої інформаційно комунікаційної системи, їх називають атаками соціальної інженерії. Цей термін означає використання психологічних методів та соціальних навичок для отримання доступу до інформації, яка зазвичай захищена технічними засобами безпеки

У зв'язку з тим, що такий підхід зловмисників є достатньо новим, питання ефективного виявлення та блокування кібератак із використанням штучного інтелекту перебуває в стані активної розробки та досі є невирішеним.

Описані вище положення визначають актуальність даної роботи, яка присвячена аналізу атак соціальної інженерії із використанням штучного інтелекту.

Об'єкт дослідження: процес виявлення кібератак які використовують штучний інтелект та соціальну інженерію

Предмет дослідження: метод виявлення кібератак які використовують штучний інтелект та соціальну інженерію.

Мета дослідження: підвищення ефективності виявлення та блокування кібератак які використовують штучний інтелект та соціальну інженерію

Перелік питань, які мають бути розроблені:

- аналіз методів виявлення кібератак які використовують штучний інтелект та соціальну інженерію.;
- розробка методу виявлення кібератак які використовують штучний інтелект та соціальну інженерію;

• розробка рекомендацій щодо виявлення та блокування кібератак які використовують штучний інтелект та соціальну інженерію

У процесі вирішення поставлених завдань у дипломній роботі були використані: методи аналізу, спостереження, індукції та моделювання; метод пошуку найкоротших шляхів між вершинами графу.

Наукова новизна одержаних результатів:

- Розроблено метод виявлення атак соціальної інженерії із використанням штучного інтелекту, що базують тільки на програмному забезпеченні із відкритим кодом, та можуть бути використані як для корпоративних цілей так і для користувацьких;

- вперше розроблено класифікацію атак соціальної інженерії із використанням штучного інтелекту. За допомогою цієї класифікації можна однозначно описати будь-яку атаку СІ із використанням ШІ.

Практичне значення роботи полягає у розробці класифікації атак соціальної інженерії із використанням штучного інтелекту та у розробці методу їх виявлення

Результати досліджень можуть бути використані для підвищення стану захищеності користувачів від атак соціальної інженерії.

Основні наукові положення і результати роботи тезисно опубліковані в збірці тез та доповідей VII Міжнародної науково-практичної конференції “Проблеми кібербезпеки інформаційно-телекомунікаційних систем”.

РОЗДІЛ 1
ОГЛЯД АТАК СОЦІАЛЬНОЇ ІНЖЕНЕРІЇ ІЗ ВИКОРИСТАННЯМ ШТУЧНОГО ІНТЕЛЕКТУ ТА МЕТОДІВ ЇХ ВИЯВЛЕННЯ

Перш ніж безпосередньо розглядати методи та методики виявлення атак варто зрозуміти принцип їх роботи, масштаб та значущість, Про що піде мова в наступному підрозділі.

1.1. Огляд атак соціальної інженерії із використанням штучного інтелекту.

Атаки соціальної інженерії – це атаки із використанням маніпулювання людьми з метою отримання конфіденційної інформації, отримання несанкціонованого доступу до систем або здійснення шахрайських дій [1].

Спеціалістами у сфері захисту інформації було представлено значну кількість різних методик, які можуть використовуватись для опису та класифікації атак соціальної інженерії. Здебільшого вони використовують як базис певні фундаментальні атаки, із варіаціями розбиття їх на підкласи [2]. Також, варто відмітити, що новизна класифікації, тобто від моменту її представлення до моменту, коли потрібно описати атаку пройшло небагато часу, напряму впливає і на кількість нових атак, які в ній можна прокласифікувати. Оскільки інформаційні технології розвиваються, зараз із значною швидкістю, то цей розвиток має вплив і на темпи появ нових атак, а не всі з них можна однозначно описати на базі попередньо вивчених. Яскравим прикладом такої, є представлена на веб-сайті Кевіна Мітника [3]. Вона є досить вагомою в світі інформаційної безпеки, велика кількість фахівців послуговуються нею, та довіряють їй, в основному у зв’язку із репутацією спеціаліста (спершу він сам виступав в ролі атакуючого, тому чітко розуміє цілі, засоби та методи при проведенні атаки). Вона є досить короткою, містить в собі усього 6 пунктів, з невеликим розбиттям деяких на підпункти:

1) Фішинг, який поділяється на Цільовий(англ. Spear phishing) та Whaling (дослівно з англ. “китобійна”), який теж свого роду є цільовим, проте метою його є атака на людей, які займають якусь високу посаду, або роль в суспільстві.

2) Вішинг та Смішинг – Вішинг - фішингова атака, в якій зловмисник комунікує із жертвою за допомогою технологій передачі голосу(можливо і відео), тобто телефонним дзвінком через оператора стільникової мережі, месенджери, та інші засоби комунікації із можливістю транслювання звуку. Смішинг - це фішинг із використанням коротких текстових повідомлень(SMS).

3) Претекстинг (анг. Pretexting) - відбувається, коли зловмисник видає себе за іншу особу, часто за якогось високопоставленого чиновника чи директора компанії, та змушує жертву виконувати його накази

4)Спонукання(англ. Baiting) - зловмисник спонукає жертву до завантаження якогось ШПЗ на свій пристрій. До прикладу, надсилає якийсь файл під видом безкоштовної копії ліцензійного продукту, або передає якийсь пристрій переносу інформації при цьому переконуючи жертву в відсутності шкідливості в ньому.

5) Несанкціоноване отримання фізичного доступу(англ. Tailgating та Piggybacking). Tailgating – полягає в непомітному супроводі авторизовонаго коритувача, з метою проникнення після нього в зону із строгим контролем. Piggybacking - є по суті тим же, тільки з однією важливою відмінністю: під час такої атаки авторизований для входу суб’єкт знає про плани зловмисника та дозволяє йому їх здійснити.

6) Послуга за послугу(лат Quid pro quo) - тип атаки, в якій зловмисник обмінює послугу на інформацію, при цьому, часто сам зловмисник своїми діями викликає в жертви бажання скористатись цією послугою, або не бути в змозі відмовитись від неї.

В більшості із цих атак можна також імплементувати елементи використання штучного інтелекту. Зловмисники вдаються до цього аби скоротити час на атаку, зробити її більш цільовою та ефективною [4]. Розглянемо варіанти таких впроваджень на прикладі двох найпопулярніших таких атак фішингових та вішингових. Оскільки інженери захисту інформації протидіють їм кожного дня десятками років.

Темпи та об’єми фішингу постійно зростають, згідно із даними Vade Security [5]. Як можна побачити із рисунку 1.1, кількість фішингових атак за третій квартал 2023-го року варіюється від 110 до 210 мільйонів.

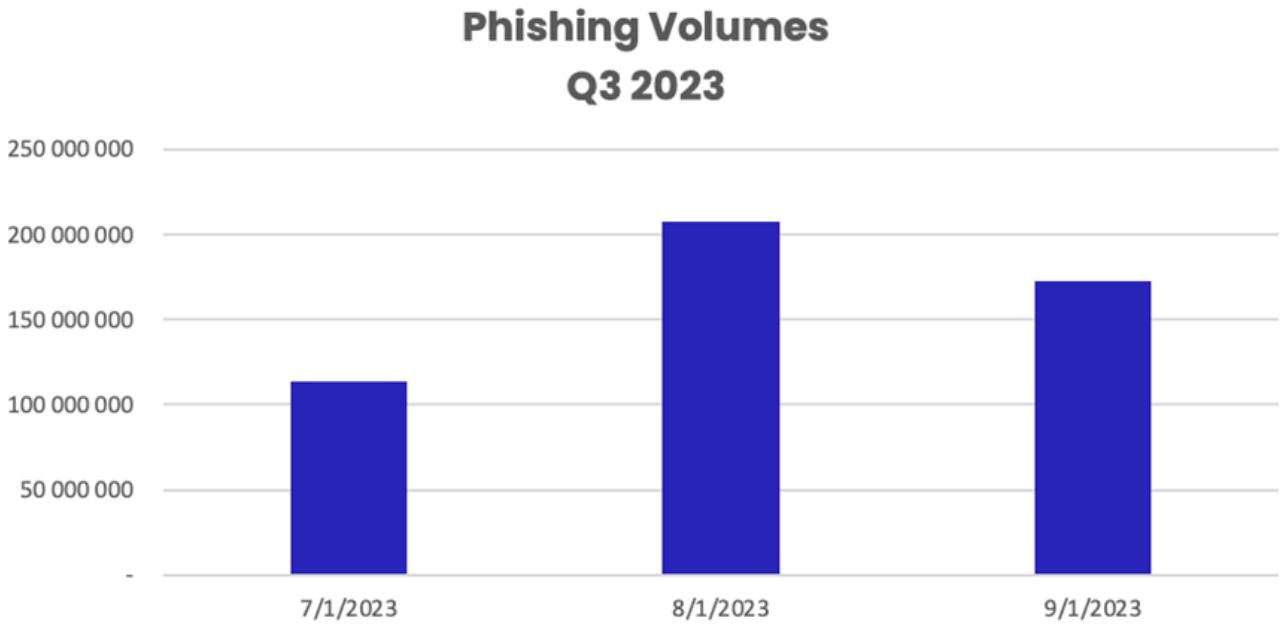


Рисунок 1.1 – Об’єми фішингових атак за третій квартал 2023-го року

Використання технологій штучного інтелекту також значно зростає, відповідно до статистики цього ж порталу [6]. Зокрема дуже активно збільшується кількість згенерованих повідомлень, які потім доставляються до жертви через засоби електронної пошти.

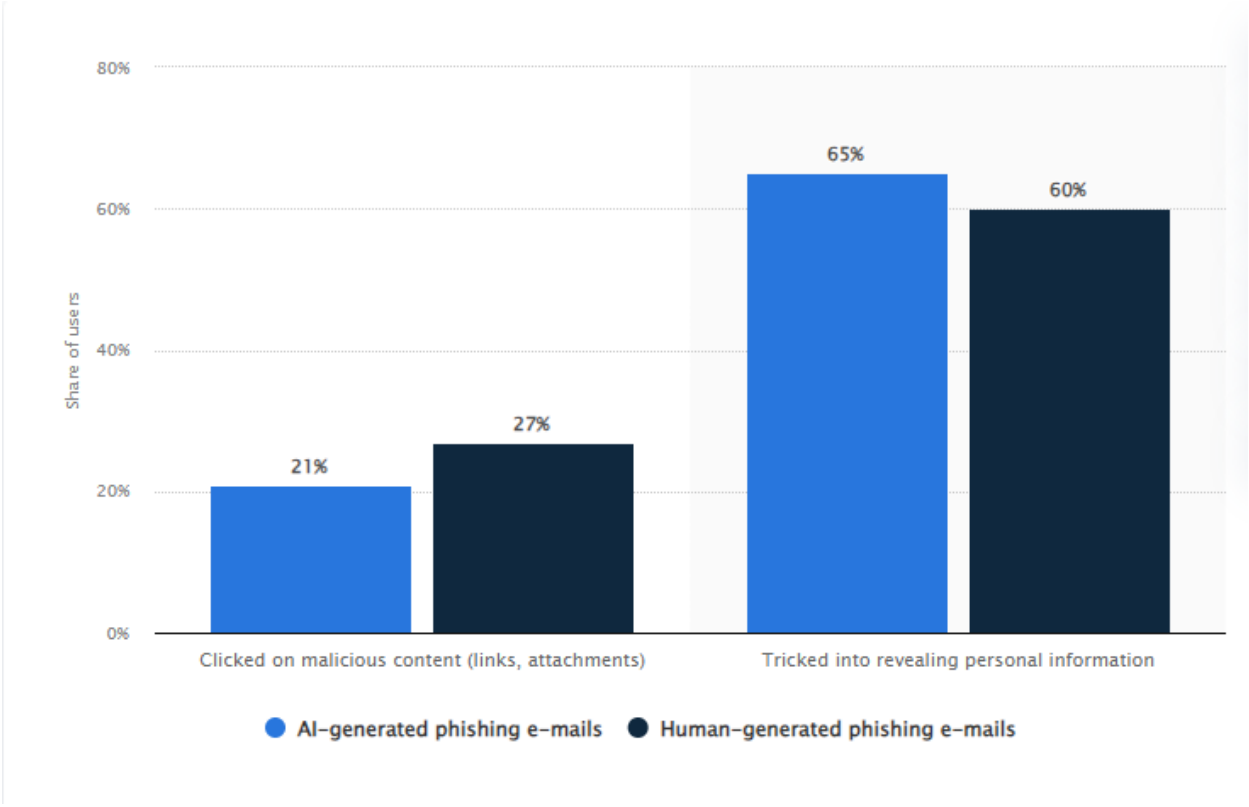


Рисунок 1.2 - Кількість взаємодії із фішинговими листами написаними людиною, та штучним інтелектом

Як можна помітити з рисунку 1.2, використання штучного інтелекту в фішингових атаках приносить зловмисникам кращу ефективність, це видно із кількості взаємодій жертв із повідомленнями [7]. Фішингові атаки найчастіше відбуваються із спробою видачі підробної веб-сторінки за легітимну. Найбільш поширеними є такі два способи (детальний опис цих атак, як сценаріїв, буде розміщено в наступному підрозділі):

- Підробка доменного імені в URL посиланні на певну сторінку в мережі Інтернет. Ідеєю такої атаки є скористатись неуважністю та відсутністю достатніх знань та навичок в суб'єкта-цілі у можливості самостійно визначити чи є легітимним цей ресурс [8]. До прикладу, багато популярних веб-сайтів в наші вимагають від користувача бути зареєстрованим, для того, щоб переглядати інформацію. Після успішної реєстрації при кожному вході на веб-ресурс буде потрібно пройти процес авторизації. Зловмисник, що хоче отримати авторизаційні дані цілі, копіює веб-сторінку ідентичного вмісту (або достатньо схожого, щоб жертва не могла швидко та

легко відрізнити). Цю веб-сторінку атакуючий розміщає в мережі Інтернет під максимально схожі посиланням та надсилає жертві в будь-який спосіб: за здопомогою засобів електронного листування поштою, SMS, приватні повідомлення в месенджері, тощо та додаючи певну преамбулу намагається змусити користувача ввести свої справжні дані за цим посиланням.

- Другий спосіб полягає в зацікавленні жертви певним маркетинговим змістом [9]. Світ, в якому ми зараз живемо просто переповнений різного роду рекламою і через цю кількість, користувачі не надто часто ставлять собі за мету спершу перевірити правдивість, оскільки ці повідомлення вдало використовують почуття людей. Створюють для них проблему та пропонують рішення, дають можливість стати кращими, зекономити гроші чи навіть заробити [10]. Соціальний інженер в якийсь спосіб готує таке повідомлення із описом та надсилає користувачу з рахунком для сплати за товар/послугу та надає користувачу волю у виборі способу оплати(найчастіше, ставлячи найменшу ціну/комісію для найбільш зручного саме для атакуючого).Також часто зловмисники вдаються до підробок листів від компаній, клієнтом якої жертва вже є. Такі повідомлення можуть містити інформацію про акції, знижки або можуть просити клієнта уточнити якусь інформації. Наприклад, посилаючись на те, що із останнього замовлення пройшло досить багато часу, тому пропонується певна знижка, або компанія хоче отримати зворотній зв'язок про свій продукт в обмін на якісь блага. Приклад такого листа, згенерованого штучним інтелектом зображений на рисунку 1.3 (персональна інформація про жертву є вигадкою). В цьому листі, зловмисник саме намагається отримати персональні дані цілі. Його повне ім'я, платіжну адресу, контактний номер телефону, дату народження та останні чотири цифри платіжного рахунку, яким він раніше користувався для оплати послуг [11]. Такі листи найчастіше намагаються надіслати в час, коли людина не може достатньо зосередитись, тобто перебуває в дорозі, у відпустці чи на якомусь заході, роблячи акцент на необхідності термінової відповіді.

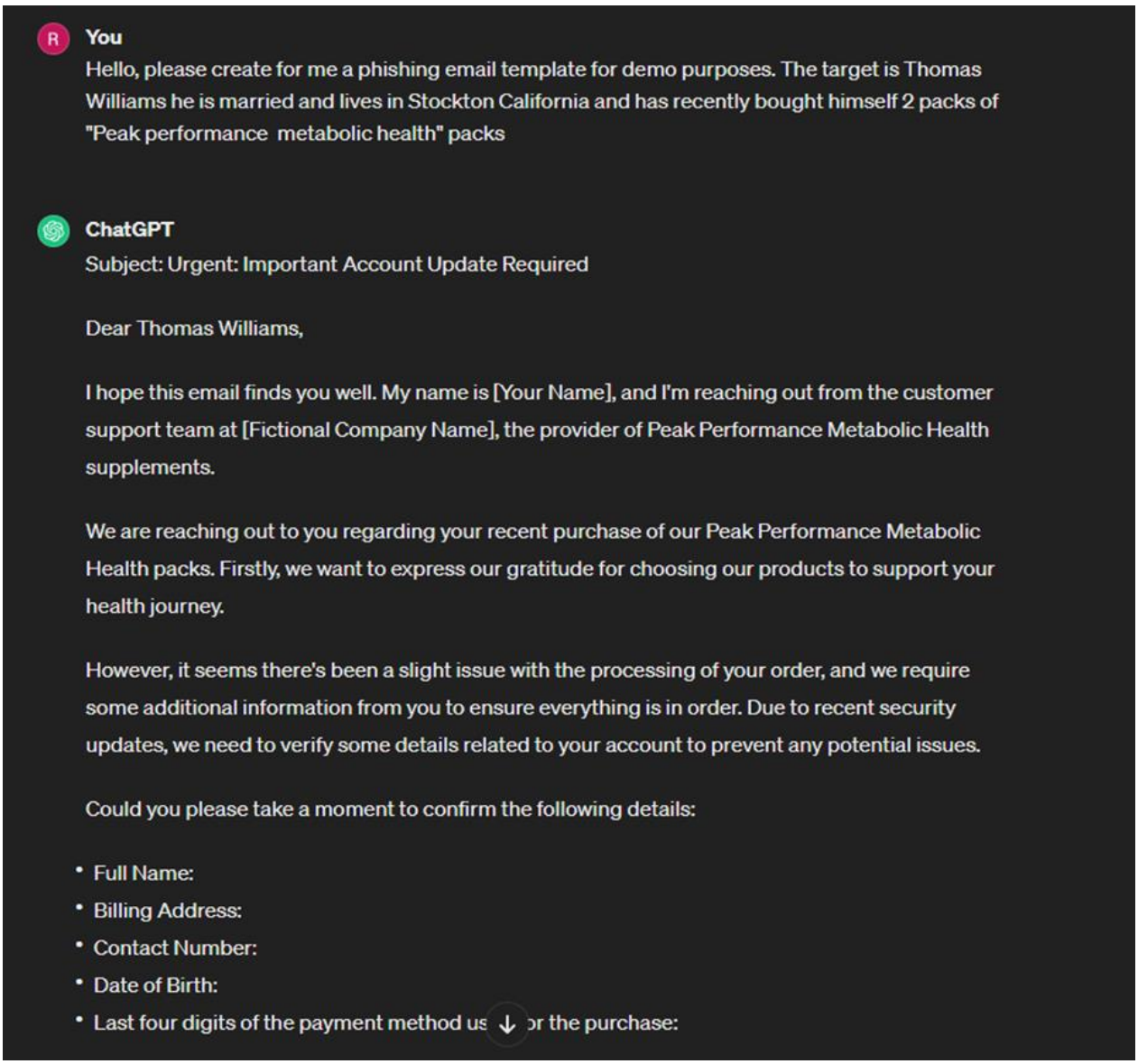


Рисунок 1.3 - Приклад фішингового повідомлення згенерованого штучним інтелектом

В атаках вішингу теж можна виділити два найпоширеніших способи:

1) Зловмиснику потрібно в певний спосіб зв'язатись із жертвою, яку він попередньо обрав. Раніше для цього використовувались телефонні дзвінки в GSM мережі, але зараз із розвитком можливостей соціальних мереж та мессенджерів, хакери перейшли й туди [12]. Після того, як було «піднято слухавку» жертві пропонуються різні послуги(як наприклад у 2 розглянутому вище виді фішингу), або відбувається імперсоніфікація. Тобто зловмисник починає видавати себе не тим, ким він є насправді та намагається увійти в довіру до жертви та змусити її виконати певні дії. Сьогодні, найлегшим

способом підвищити рівень довіри людини, у те, що її співрозмовник не обманює, а дійсно є тим, ким себе називає є використанням штучного інтелекту, а саме підробка голосу зловмисника за допомогою технології deepfake. На рисунку 1.4 відображено, як за допомогою сервісу Fakeyou [13] можна легко змінити свій голос на бажаний. Цікавою особливістю є те, що можна навіть обрати бажане емоційне забарвлення голосу.

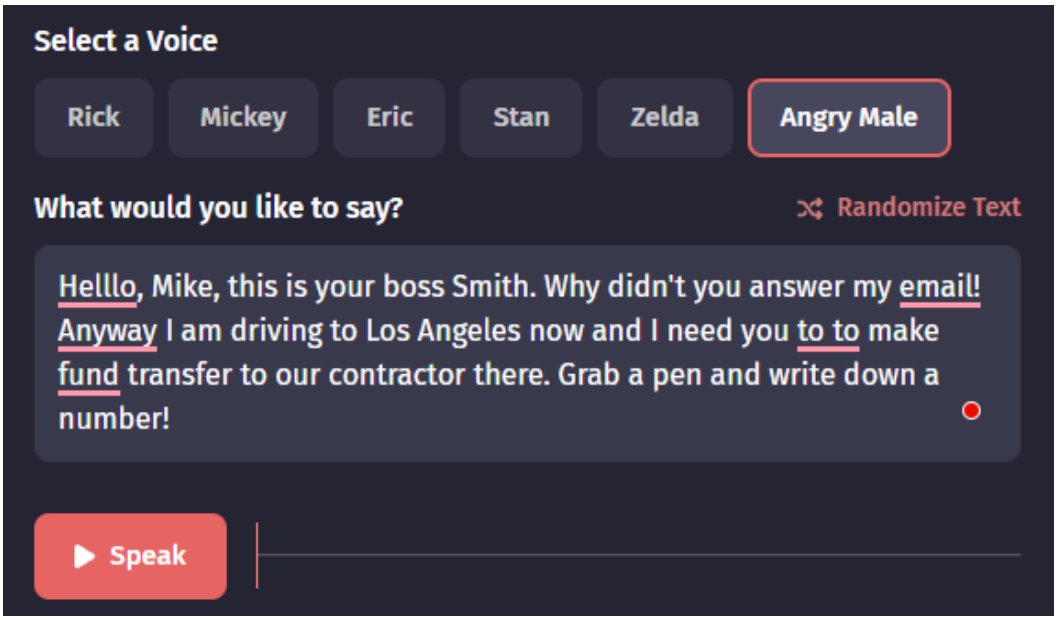


Рисунок 1.4 - Використання технології deepfake для зміни голосу

2) Другий метод полягає в тому, що попередньо зловмисник не обирає когось конкретного за ціль для атаки. В певний спосіб, йому вдається отримати доступ до БД із контактами якоїсь певної групи людей. Активно в цілях отримання такої інформації використовується технології OSINT(детальніше інші методи розглядатимуться при огляді цього сценарію атаки) [14]. Маючи в користування ці дані, зловмисник може відфільтрувати її певним чином для проведення атаки, або просто намагатись розпочати спілкування із усіма суб’єктами представленими там. Далі, логіка атаки, як і в попередньому випадку зводиться власне до самого дзвінка жертві, в якій для досягнення цілей використовуються або якийсь новостворений хакером підхід, або якийсь загальний сценарій.

Кримінальна активність в мережі Інтернет з кожним роком росте, це стосується і атак соціальної інженерії. Перед фахівцями із захисту інформації стоїть нелегке завдання – захищати людей не тільки від вже існуючих атак, а й активно розробляти методу протидії новостворeним. Кевін Мітник, один з найвідоміших соціальних інженерів якому довелось побувати одразу в обох ролях (атакуючого і захисника) колись зазначив: “Ви ніколи не можете захистити себе на 100%. Але ви маєте захищати себе настільки, на скільки можливо та зменшити ризики до прийнятного рівня, але в вас ніколи не вийде уникнути їх всіх” [15]. Багато людей знаходять саме ці слова певної філософією кібербезпеки, та використовують її як базис при побудові систем захисту інформації. Варто розуміти, що в будь-якій системі, найслабшою ланкою завжди будуть люди зважаючи на їх фізіологічні, психофізичні та психологічні фактори. Тому спеціалісти із кібербезпеки часто першочергово ставлять за завдання роботу із людьми. Для виконання цього завдання спеціалісти вдаються то проведення тренінгів із обізнаності SAT (англ. Security awareness). Оскільки рівень обізнаності працівників є обернено пропорційним до ймовірності успішного виконання атаки соціальної інженерії та до чисельності ризиків, пов'язаних із цими атаками.

Варто зазначити, що робота суто із персоналом не може забезпечити достатній стан захищеності системи, для цього потрібно паралельно використовувати також і технічні рішення. Проблема атак соціальної інженерії не є новою в різний час із різною ефективністю використовувались рішення протидії їм на технічному рівні. Зокрема шляхом використання систем контролю вхідного та вихідного каналів зв'язку. Одним із найбільш широко використовуваних є різні SPAM фільтри, поштових клієнтів. Проте, часто вони не виявляються достатньо ефективними, користувачі не можуть налаштовувати їх самостійно пів певні задачі, а також їх механізми виявлення ШІ поки що ще на досить низькому рівні [16]. Саме тому, розробці такого ефективного методу і присвячена ця робота. Мова про методи та методики та механізми, які використовуються для виявлення таких атак та захисту від підміни адресанта піде у наступних підрозділах.

1.2 Огляд методів та методик виявлення атак соціальної інженерії

Із стрімким зростанням використання штучного інтелекту в атаках соціальної інженерії перед спеціалістами захисту інформації постало завдання розробки ефективних методів виявлення таких атак. Останнім часом із різною успішністю фахівці представляють їх широкому загалу [17]. Деякі із них базуються на нових моделях виявлення, але є й такі, базис яких вже давно відомий у світі, а отже про нього знають і зловмисники. Можна виділити основні три методи, які найширше використовуються в наші дні:

1. Аналіз вмісту - однією з основних функцій ШІ у виявленні атак соціальної інженерії аналіз вмісту електронної пошти. Алгоритми можуть ретельно перевіряти текст вхідних повідомлень, щоб виявити підозрілі шаблони, граматичні помилки або невідповідності, які можуть вказувати на спробу фішингу. Порівнюючи вміст вхідних електронних листів із відомими фішинговими шаблонами та мовними патернами, алгоритми штучного інтелекту можуть позначати потенційно шкідливі повідомлення для подальшого дослідження. Також широко використовується пошук по ключових словах в тексті фішингового повідомлення. Це слова, які зазвичай використовують кіберзлочинці для обману, залякування жертв. Постійно вивчаючи нові дані ці алгоритми можуть адаптуватися до нових тактик фішингу та з часом підвищувати точність виявлення [18].

2. Поведінковий аналіз - окрім аналізу вмісту повідомлень, методи виявлення на основі ШІ також зосереджені на оцінці поведінки відправника. Створюючи профілі звичайних шаблонів спілкування для окремих користувачів і організацій, ці рішення можуть виявляти відхилення, які можуть сигналізувати про спробу фішингу. Наприклад, раптові зміни в шаблонах надсилання електронної пошти, незвичайні IP-адреси або розбіжності в домені відправника можуть викликати червоні прапорці, що вказують на підроблений або зламаний обліковий запис. Аналізуючи метадані, пов’язані із заголовками електронної пошти та інформацією про відправника, алгоритми ШІ можуть ідентифікувати аномалії, які можуть вказувати на зловмисну діяльність. Цей контекстний аналіз дає змогу організаціям

розрізняти легітимне спілкування та спроби атак соціальної інженерії дозволяючи їм вживати відповідних заходів для зменшення ризику [19].

3. Контексні фактори - крім аналізу вмісту та відправника, рішення безпеки електронної пошти, керовані штучним інтелектом, враховують контекстні фактори, щоб оцінити легітимність вхідних повідомлень. Це включає врахування відносин між відправником і одержувачем, відповідність вмісту електронної пошти попереднім повідомленням і час надсилання повідомлення. Шляхом контекстуалізації вхідних електронних листів у ширших рамках попередніх взаємодій та організаційних норм алгоритми ШІ можуть точніше визначати їх автентичність [20].

Ці методи не обов’язково використовувати по-одному, можна їх об’єднувати самостійно визначаючи значущість кожного з них. Проте, в при впровадженні цих методів варто враховувати, що для навчання мовних моделей, яким будуть передаватись великі обсяги даних потрібне потужне апаратне забезпечення та можливість постійної підтримки його коректної роботи.

Але також варто зазначити, що ці методи мають низку переваг, зокрема - здатність швидко реагувати на нові загрози. Оскільки кіберзлочинці постійно вдосконалюють свою тактику, алгоритми штучного інтелекту можуть адаптуватися та вивчати нові дані в режимі реального часу, дозволяючи організаціям залишатися попереду [21]. Ця гнучкість особливо важлива перед обличчям атак нульового дня або раніше невідомих методів фішингу, коли традиційний захист на основі сигнатур може виявитися неспроможним. Виявивши потенційну загрозу атаки соціальної інженерії, керовані штучним інтелектом рішення безпеки можуть ініціювати автоматичні відповіді, щоб зменшити ризик. Це може включати поміщення підозрілих вхідних повідомлень у карантин/чорний список, блокування зловмисних посилань чи вкладень або сповіщення співробітників служби безпеки для подальшого розслідування. Автоматизуючи ці дії з усунення загроз, організації можуть мінімізувати вплив атак і запобігти їх поширенню по мережі. Використовуючи автоматизацію на основі штучного інтелекту, організації можуть реагувати на атаки соціальної інженерії швидше та ефективніше та дешевше, зменшуючи ризик витоку даних і фінансових втрат. Крім того, алгоритми ШІ постійно вдосконалюють свої

можливості виявлення через процес зворотного зв’язку та ітерації. Аналізуючи результати попередніх виявлень і вводячи нові дані у свої моделі, ці системи стають все більш вправними у виявленні та запобіганні спробам фішингу. Цей ітеративний цикл удосконалення гарантує, що організації отримають переваги від покращеного захисту з часом, навіть якщо кіберзагрози розвиваються та стають все більш складними.

Оскільки організації збирають більше даних і отримують інформацію про нові загрози, алгоритми штучного інтелекту можуть адаптуватися та підвищувати точність виявлення, надаючи організаціям проактивний захист від фішингових атак.

Проте також варто заначити, що ці методи все ще є досить новими, та не показали себе поки що достатньо ефективно, щоб виявляти хоча б більшість атак, вони активно розвиваються, та найближчим часом повинні будуть показати себе краще, або перестануть використовуватись.

1.3 Огляд найпоширеніших механізмів захисту від підміни даних при атаках соціальної інженерії через електронну пошту

Розглядаючи технічну частину питання варто звернути увагу на те, що більшість атак соціальної інженерії відбуваються через електронну пошту. Це пояснюється тим, що в наші дні цей вид комунікації став практично основним для більшості підприємств по всьому світу через його функціональність, зручність та постійним розвиток. Відповідно, фахівцями з кібербезпеки були розроблені рішення, які покликані підвищити стан захищеності користувачів поштових скриньок та цих інформаційно-комунікаційних системах. Розглянемо ці рішення детальніше:
1) SPF (Sender Policy Framework) - це механізм автентифікації електронної пошти, який допомагає захистити домен від підробки та фальшивих повідомлень [22]. SPF визначає список дозволених поштових серверів, які мають право надсилати електронну пошту від певного домену. При отриманні повідомлення поштовий сервер отримувача може перевірити SPF-запис домену відправника, щоб визначити, чи може бути відправлене це повідомлення від заданого імені.

Основний принцип роботи SPF полягає в наявності TXT-запису SPF для домену, в якому вказується список дозволених IP-адрес або поштових серверів, які мають право надсилати пошту від цього домену. При отриманні повідомлення поштовий сервер отримувача перевіряє IP-адресу відправника повідомлення та перевіряє її зі списком дозволених IP-адрес, вказаних в SPF-записі.

Якщо IP-адреса відправника повідомлення збігається з однією з дозволених IP-адрес, то перевірка SPF пройдена успішно, і повідомлення приймається. У разі, якщо IP-адреса відправника не збігається з жодною з дозволених IP-адрес, то перевірка SPF вказує на потенційну підробку або фальшиве повідомлення, і сервер отримувача може відхилити або помітити це повідомлення як потенційно шкідливе.

SPF допомагає завадити кіберзлочинцям, які намагаються підробити електронні повідомлення та використати домен для відправки спаму, фішингу або інших видів кібератак. Він дозволяє отримувачам перевіряти автентичність поштових серверів та перевіряти, чи є вони дійсно авторизованими для відправки пошти від певного домену..

2) DKIM (DomainKeys Identified Mail) - це технологія аутентифікації електронної пошти, яка допомагає захистити поштовий трафік від підробки та фальшивих повідомлень. Вона базується на цифровому підписі, який додається до кожного вихідного повідомлення[23].

Основний принцип роботи DKIM полягає в тому, що відправник електронного повідомлення генерує цифровий підпис для цього повідомлення. Для цього використовується асиметричний криптографічний алгоритм, такий як RSA. Відправник вибирає приватний ключ, яким підписує повідомлення, і публічний ключ, який буде розміщений у DNS-записах домену.

Після того, як повідомлення підписано приватним ключем, цифровий підпис додається до заголовків повідомлення. Отримувач, отримавши повідомлення, може перевірити його автентичність. Для цього він отримує публічний ключ з DNS-записів домену відправника і використовує його для перевірки цифрового підпису. Якщо перевірка успішна, повідомлення вважається автентичним і не підміненим.

DKIM дозволяє отримувачам перевіряти, що повідомлення було надіслано відправником, який вказаний в заголовках, і що воно не було змінене під час транзиту. Це забезпечує довіру до вмісту повідомлення та джерела відправника. Крім того, DKIM дозволяє розпізнавати спам та фішингові повідомлення, оскільки вони часто не мають правильного DKIM-підпису або використовують підроблений.

Застосування DKIM сприяє підвищенню безпеки електронної пошти. Використання DKIM разом з іншими технологіями аутентифікації, такими як SPF і DMARC, створює більш надійний механізм захисту від кіберзагроз і сприяє покращенню доставки та довіри до електронної пошти.

3) DMARC (Domain-based Message Authentication, Reporting, and Conformance) є протоколом, який допомагає організаціям захистити свою доменну адресу електронної пошти від спаму, фішингу та інших видів кібератак. Він працює на основі двох основних технологій – SPF (Sender Policy Framework) і DKIM (DomainKeys Identified Mail), та надає можливість організаціям встановлювати політики перевірки та повідомлення про доставку електронної пошти, відправленої від їх домену [24].

Основний принцип роботи DMARC полягає в тому, що він використовує попередньо описані методи автентифікації разом.

DMARC встановлює політики, що вказують поштовим серверам, як саме потрібно обробляти вхідну пошту, яка претендує на відправку від певного домену. Ці політики визначають, чи потрібно виконувати SPF та DKIM перевірки, що робити з неперевіреними повідомленнями, і як повідомляти про результати перевірки.

Крім того, DMARC надає можливість організаціям отримувати звіти про доставку та відхилення електронної пошти, що надійшла від їх домену. Ці звіти дозволяють перевірити, чи були надіслані повідомлення відповідно до встановлених політик DMARC та чи не використовувалися їх доменні адреси для зловмисних цілей. Завдяки використанню DMARC організації можуть ефективно контролювати та захищати свою доменну адресу електронної пошти від кібератак. Також, це дозволяє суттєво знизити ризик підробки домена для відправки шкідливих повідомлень, а також покращити доставку коректно підписаних та автентичних повідомлень до отримувачів.

Також підприємства широко використовують як метод захисту та протидії атакам соціальної інженерії постійне навчання співробітників. Основна мета цього навчання полягає в навчанні співробітників про техніки соціальної інженерії, виявлення потенційних загроз і вироблення в них навичок забезпечення власної безпеки та освідомлення ризиків, які їм притаманні.

SAT ставить за мету підвищення освіченості персоналу щодо потенційних загроз соціальної інженерії, таких як фішинг, вимагання викупу (ransomware), вішингу, спаму, тощо. Ці тренінги надають співробітникам необхідні знання та практичні навички для розпізнавання підозрілих ситуацій та прийняття вірних рішень у випадку потенційної загрози [25]. Програма SAT складається із таких пунктів:

1) Навчання співробітників про різні види соціальної інженерії, їх методи та основні ознаки. На цьому етапі розглядаються прикладі реальних атак та симулюються підходи зловмисників із детальними пояснення, як виявляти ознаки атаки.
2) Навчання співробітників розпізнавати фішингові повідомлення, та визначати яку інформацію в жодному разі не можна передавати у відповідь на такі повідомлення: паролі або банківські дані, ідентифікаційні дані, тощо. Вони навчаються звертати увагу на підозрілі посилання, електронні адреси та вкладення в повідомленнях.
3) Навчання створенню стійких паролів і правилам безпеки при їх використанні. Учням доносять необхідність використання унікальних паролів для кожного із сервісів та привчають до постійного контролю стану паролів(коли вони останній раз змінювались, як їх зберігати, тощо).
4) Навчання співробітників про заходи фізичної безпеки в офісному середовищі. В ході навчання учасники отримують рекомендації стосовно захисту робочого місця, обмеження доступу до конфіденційної інформації та поводження зі сторонніми особами.
5) Організація практичних вправ і симуляцій, де співробітники можуть застосовувати свої навички безпеки у віртуальних або контрольованих

сценаріях. Це допомагає перевірити реальні навички і виявити місця на які потрібно додатково та повторно звернути увагу.

6) Забезпечення розуміння необхідності виконання усіх вище перелічених дій на постійній основі. Тобто, учні повинні певним чином змушувати себе тримати в голові інформацію про потенційні загрози, постійно бути до них готовими та в разі їх виникнення переходити до виконання послідовності дій, яка допоможе їм залишатись захищеними.

Тож, тренінги SAT допомагають підприємствам покращити свою загальну безпеку, знизити ризик від атак соціальної інженерії та залучити співробітників до активної участі в процесі забезпечення безпеки. Розуміння загроз і навички безпеки стають необхідними компетенціями для всіх працівників, оскільки їх залученість в цьому процесі допомагатиме суттєво підвищити степінь захищеності як їх самих, так і їх роботодавця.

1.4 Огляд найпоширеніших сценаріїв атак соціальної інженерії, їх класифікація

Досить часто зловмисники вдаються до використання уже готових шаблонів та сценарії атак. Це пояснюється їх популярністю та низьким порогом входу, адже для виконання атаки на базовому рівні, в принципі, не потрібний навіть персональний комп’ютер, а для не надто складних, він може бути із найпростішим програмним та апаратним забезпеченням. Сценаріїв таких атак можна окреслити досить багато, проте зрештою вони всі є певними надбудовами над найпростішими. Такі сценарії несуть велику користь для аналітиків – оскільки вони ще не надто загромаджені великою кількістю зв’язків і даних, проте представляють собою фундаментальні цілі та методи, які переслідують хакери для виконання своїх цілей.

Першою схемою, яку пропонується розглянути в даній роботі є схемою атаки, мова про яку було згадано кількома підрозділами вище. Це одна із найпростіших атак-підроблення URL посилання на ресурс. Такі атаки зустрічаються доволі часто та є свого роду перехідними між просто спробами маніпулювати у тексті повідомлення,

тому що включають в себе перший технічний вузол – власне виконання підробки посилання. Схема такого сценарію, зображена на рисунку 1.5.

Рисунок 1.5 - Найпростіша схема атаки із підробленням URL посилання

На першому кроці зловмисник обирай веб-сайт, який буде використовуватись для атаки. Найчастіше такими стають сторінки платіжних сервісів, банків, державних установ, логістичних та поштових компаній чи соціальних мереж [26]. Коли ресурс обрано необхідно в якийсь спосіб скопіювати, або написати максимально схожий,

оскільки від рівня схожості із легітимною сторінкою залежить успішність проведення атаки.

Другий етап цього сценарію передбачає отримання певного кредиту довіри від цілі, тим самим вплинути на зосередженість жертви, та змусити її вважати посилання легітимним. Часто цього досягається шляхом використання різних методів психологічного тиску на жертву. Також цього можна досягти за допомогою технічних засобів. Якщо зловмисник та жертва під'єднані до однієї мережі, то хакер матиме змогу перехопити трафік, який надсилає жертва та підмінити його на бажаний. Такий тип атаки називається "людина по середині"(англ. Man in the middle). Також, для цього широко використовуються повідомлення згенеровані штучним інтелектом. Зловмисник збирає певні дані, обробляє їх та передає на аналіз, а на виході отримує готове повідомлення для певної цілі. Якщо, жертві не вдалось помітити не відповідності URL посилання тому, яким воно насправді має бути, і вона ввела туди свої дані, то атаку можна завершеною.

На завершальному етапі хакер отримає бажані дані від жертви. Серед даних, які найчастіше зацікавлюють соціального інженера є: авторизаційні дані, дані про фінансові операції, рахунки, тощо.. Далі, зловмисник починає оперувати цими даними та вирішує, чи задовольняють вони його мету атаки, чи можуть бути використаними в наступних атаках на цього ж користувача, або й на інших, з яким в них спостерігаються спільні риси.

Другий сценарій проведення атаки соціальної із використанням штучного інтелекту який був коротко розглянутий в першому підрозділі та заснований на розповсюдженні повідомлень із використанням рекламного або псевдорекламного вмісту. Такі атаки активно використовуються, оскільки користувачі мережі Інтернет загалом уже звикли до того, що практично будь-яка сторінка на яку вони переходять міститиме інформацію із просування певного товару чи послуги. Розглянемо схему такої атаки:

Рисунок 1.6 - Атака соціальної інженерії із використанням рекламного або псевдорекламного змісту

Перший етап в цьому сценарії атаки є найголовнішим, оскільки на ньому базуватимуться всі подальші кроки. Зловмиснику потрібно обрати чи вигадати певний продукт чи послугу, який він рекламуватиме та обрати психологічні прийоми для впливу на потенційну жертву. Для цього переважно, зловмисник переходить до активного вивчення особи, яку збирається атакувати, збирає всю можливу інформацію

та намагається знайти чутливу інформацію, яку потім використовуватиме для початку діалогу.

Далі формується повідомлення яке є результатом роботи зловмисника на попередньому етапі. Для формування повідомлень зараз дуже активно серед зловмисників використовуються методи штучного інтелекту. Навіть найпростіший розмовний бот, запит до якого буде сформульований достатньо чітко та визначено, зможе сформувати такого роду повідомлення, як це було показано на рисунку 1.1.3.

Наступним етапом є початок діалогу, тут зловмисники можуть тримати жертву протягом тривалого періоду часу, оскільки завдяки певним психологічним тактикам це дозволятиме отримати їм більше уваги та довіри від жертви.. Недосвідчений користувач, на цьому етапі може справді зацікавитись контентом реклами та почати вести розмову уже із власного бажання. Така взаємодія є дуже цінною для хакера, оскільки він може продовжити грати на почуттях цілі, яка почуватиметься від цього щасливою. Такому явищу є наукове пояснення, згідно із роботою Пола Зака, під час розмови в, якій людина відчуває, що довіряє комусь, або довіряють їй в її організмі виділяється допамін, що призводить до відчуття задоволення. Дослідження показали, що для цього необов’язково спілкуватись особисто, телефонної розмови чи листування [27]. Тож заручившись такою потужною підсвідомою підтримкою, зловмисник переходить до підштовхування жертви на виконання бажаний дій. Втім, якщо цього не достатньо, етап діалогу триватиме. Тут знову в нагоді стає ШІ, атакуючому не обов’язково мати гарну уяву – за нього відповіді на питання може давати його віртуальний «помічник».

Після успішного змушення виконання певного алгоритму атаку можна вважати успішною, так як, зловмиснику вдалось отримати бажане. Але, спілкування може й продовжитись, жертва, може не одразу виявити, що її обманули, при певному збігу подій соціальний інженер зможе кілька разів повторити атаку.

Для атак вішингу використанням штучного інтелекту не обмежується тільки підготуванням повідомлення, оскільки його можливості дозволяють хакерам видавати себе за інших людей в режимі реального часу. Розглянемо схему атак вішингу:

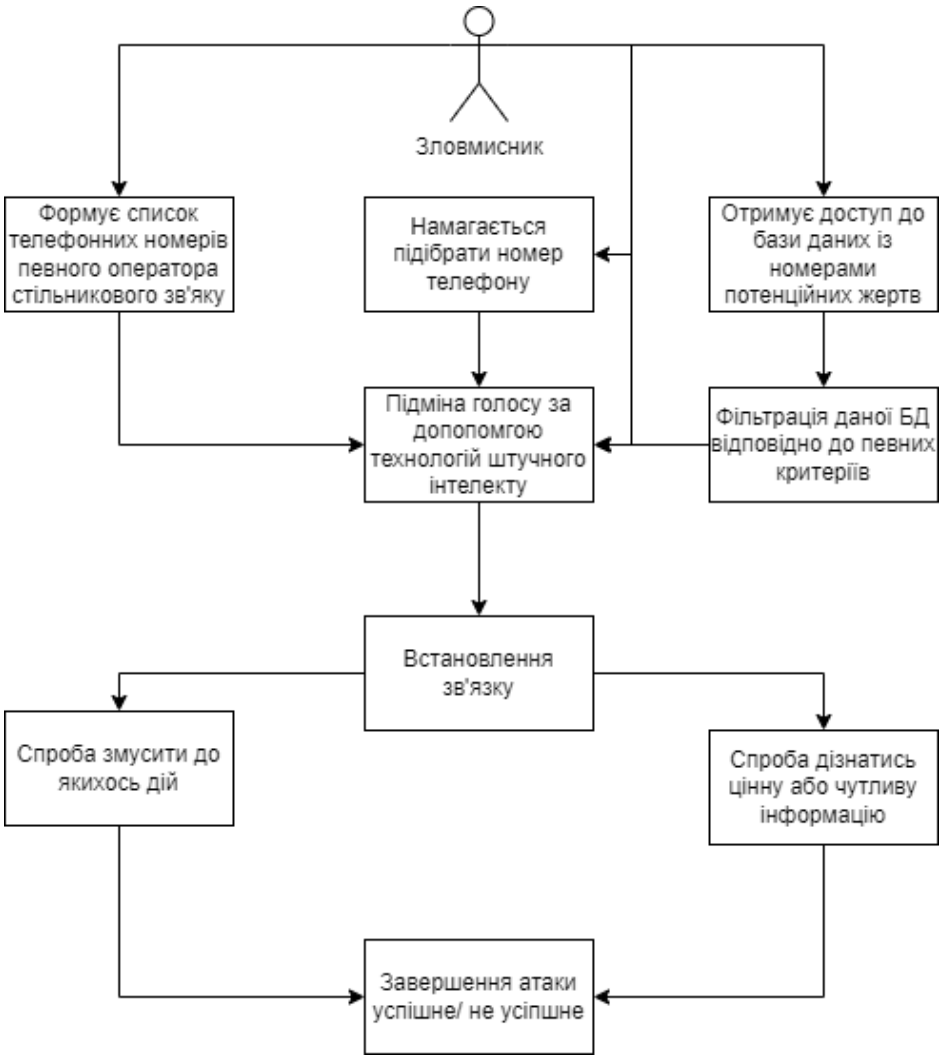


Рисунок 1.7 - Сценарії атак вішингу із використанням штучного інтелекту

На першому етапі зловмисник займається пошуком потенційної жертви. Для реалізації атаки такого сценарію очевидним є те, що під цим пошуком розуміється знаходження номеру телефону, чи інших даних, які дозволять здійснити дзвінок (наприклад, особистий аккаунт у месенджері). Зловмиснику достатньо буде просто спробувати виконати з’єднання намагаючись підібрати номер телефону, проте такий спосіб буде не найбільш ефективним оскільки залежатиме від достатньо великої кількості безпосередньо не пов’язаних між собою факторів, в тому числі, певною мірою й від вдачі. Загалом достатньо важко досягнути поставленої мети атаки методами психологічного впливу зв’язавшись із зовсім незнайомою людиною, про яку нічого не відомо. Звичайно, якщо зловмисник дуже вправний і досконало володіє психологічними маніпуляціями, то певна ймовірність вдало провести атаку існує,

проте, мабуть, жоден не відмовиться суттєво підвищити свої шанси на упіх. Таку можливість надає наявність певної бази даних потенційних жертв, найбільше користь нестимуть списки із додатковою інформацією. Такі БД сьогодні можуть бути незаконно придбаними (оскільки містять приватну інформацію) в інших зловмисників. [28]. Далі атакуючий обирає із бази найбільш цікавих для нього осіб, або групу осію, яку, обирають як за віком і статтю так і за суспільним становищем, рівнем забезпеченості, тощо. Деякі соціальні інженери ведуть власну статистику, для того, щоб розуміти, яку із суспільних груп вони можуть атакувати ефективніше.

На другому етапі сценарію, встановлюється зв'язок. Під цим мається на увазі не лише, сам факт успішного технічного з'єднання, а й початку діалогу між жертвою та зловмисником. На цьому етапі атакуючий представляє раніше вигадану чи згенеровану за допомогою штучного інтелекту легенду, та намагається викликати зацікавлення в співрозмовника, з метою продовження спілкування. Коли розмова триває, вішер ставить співрозмовнику деякі питання, які були сформовані спеціальним чином, щоб визначити емоційний стан потенційної жертви. Аналізуючи ці відповіді він підбирає тактику, та позбавляє людину можливого відчуття стресу від спілкування із незнайомою людиною, та намагається налаштувати жертву на співпрацю та відчуття захищеності.

Кінцевим етапом, як і раніше є певної вигоди чи досягнення певних цілей. Після визначення та отримання усіх необхідних даних про співрозмовника від нього самого, атакуючий поступово починає нав'язувати йому свої бажання. Характерною особливістю виконання цих дій є те, що зловмисник ставить собі за мету зробити так, аби жертва сама воліла виконувати його бажання. Цього можна досягти використовуючи низку методів, таких як: виклик емпатії, бажання допомогти, бажання отримати похвалу в подальшому тощо. Також зловмисник постійно намагається концентрувати увагу співрозмовника на власне діалозі і намагається практично не називати ті цінні активи, які бажає отримати. Оскільки, всі попередньо згадані пункти можуть бути здійсненими успішно, але завжди потрібно враховувати можливість іншого розвитку подій, коли відбудеться, щось таке що відволіче увагу жертви, це може викликати сумніви в легітимності діалогу і тоді знову доведеться

різними маніпуляціями переконувати жертву в справжності і чесності намірів атакуючого [29]. Тому, атакуючому необхідно діяти швидко та достатньо якісно. В результаті виконання всіх етапів алгоритму успішно, він отримає доступ до необхідної інформації.

Раніше, атаки такого формату несли в собі більше ризиків і власне для атакуючого. Оскільки він розкривав свій голос. Зловмисники намагались обходити це різними методами(надрив голосових зв'язок, штучно викликаний нежить, постійний кашель симулювання логопедичних вад, змінення голосу шляху перекривання носових каналів, тощо). Зараз такої необхідності немає – технології deepfake, про які вже було згадано вище дозволяють в режимі реального часу під час активного з'єднання змінювати голос зловмисника на інший. Окрім складності подальшого виявлення вішера, це також може мати певний психологічний вплив. Ні для кого не таємниця, що під певні задачі певні тембри голосу та манери спілкування підходять краще. До прикладу візьмемо операторів кол-центрів та колекторів боргів. Робота в них схожа – спілкуватись по телефону, але обирають туди переважно людей із певною специфікою манери ведення розмови та висоти голосу: кол-центр- спокійний, не надто високий, щоб не «дратувати» співрозмовника; колекторська служба – низький, басистий, тобто такий, який виражатиме більшу «серйозність» розмови та спонукатиме до швидшого вирішення обговореної проблеми [30]. Таке сприйняття закладено в нас природою, тому не варто дивуватись, що зловмисники теж користуватимуться цим для досягнення своїх цілей.

Після вивчення найпоширеніших сценаріїв атак виникає розуміння, що вони є досить тонкими в реалізації та підходи постійно варіюються, хоча й є дуже схожими. Через це виникає питання їх коректної класифікації. Як було описано вище, всі ці атаки можуть бути описаними в рамках вже існуючих класів, штучний інтелект приніс багато нового, тож для їх чіткого та повного опису потрібна нова класифікація, яка враховуватиме нововведення. Таку класифікацію й пропонується інтегрувати в цьому розділі.

Перш ніж вводити класифікацію атак соціальної інженерії із використанням штучного інтелекту, варто розбити такі атаки на етапи. Для цього скористаємось

загальноприйнятим фреймворком Cyber Kill Chain, який відображено на рисунку 1.8 [31].

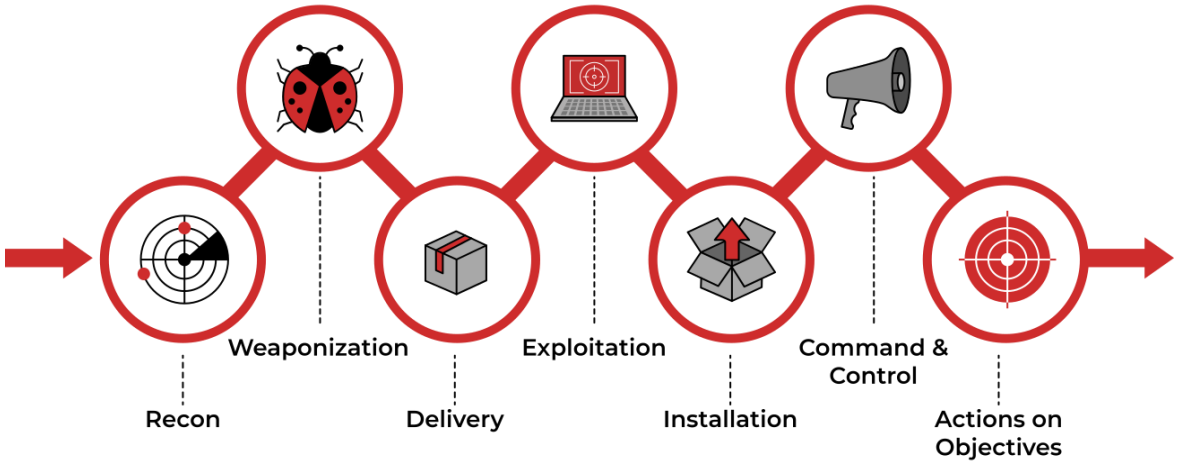


Рисунок 1.8 - Cyber Kill Chain Framemork

Першим етапом атаки соціальної інженерії є розвідка: зловмисник збирає інформацію про ціль, використовуючи відкриті джерела, часто соціальні мережі, оскільки досить часто в них люди добровільно висвітлюють приватну інформацію, наприклад, дату народження, фотографії отримання водійських прав та багато іншого. На цьому етапі зловмисник, використовуючи ШІ, може визначити ширше коло спілкування та інтересів людини, а таку інформацію зловмисник реалізовуватиме на наступних етапах.

Другим етапом є озброєння: в випадку атак соціальної інженерії це створення фішингового посилання, документу із шкідливим програмним завантаженням, фішингового листа. Саме для останнього зараз широко використовується штучний інтелект, оскільки зловмисник формулює запит, використовуючи інформацію зібрану на попередньому етапі, та на виході отримує готовий лист, написаний під конкретну ціль.

Третім етапом є доставка: зловмисник відправляє згенероване раніше шкідливе навантаження жертві. На цьому етапі штучний інтелект може використовуватись для виявлення нових каналів зв'язку із ціллю.

Четвертим етапом є експлуатація: після успішної доставки шкідливого навантаження, зловмисник часто вступає у діалог із жертвою, намагаючись різними методами психологічного впливу такими як залякування та вимагання досягнути своїх цілей. Штучний інтелект може допомогти зловмиснику проаналізувати текстову відповідь від цілі (якщо спілкування відбувається засобами електронної пошти) та створити більш доцільну відповідь.

П'ятим етапом є інтсталяція: атаки соціальної інженерії можуть бути використаними не тільки для того, щоб взаємодіяти лише з ціллю-людиною. Після успішної доставки шкідливого навантеження, в якому може містити програмний код, який буде виконано на пристрої користувача. Деякі застосунки, що використовуються ШІ можуть написати такий код для зловмисника.

Шостим етапом є отримання управління: даний етап передбачає комунікацію між хостом зловмисника враженим пристроєм жертви. За допомогою штучного інтелекту можна генерувати повідомлення, які будуть покликані маніпулювати жертвою для досягнення поставлених раніше зловмисних цілей.

На останньому, сьомому етапі, який називається виконання дій зловмисник власне кажучи отримує/викрадає бажану інформацію, обробляє. В подальшому вона може бути використана для навчання штучного інтелекту чи генерації нових засобів атаки на цю ж чи іншу ціль [32].

В попередньому підрозділі розглядалась класифікація атак соціальної інженерії Кевіна Мітника. На час її першої публікації, про технології штучного інтелекту здебільшого говорили тільки як про щось, що неодмінно зможе відіграти свою роль у майбутньому, оскільки тоді воно ще здавалось чимось далеким. Звичайно, ця робота є досить загальною і в її базис лягають і сучасні підходи зловмисників. Однак, на мою думку, теперішній розвиток цих атак саме за допомогою штучного інтелекту вимагає нової класифікації, більш строгої на вузько направленої – таку і представлено нижче.

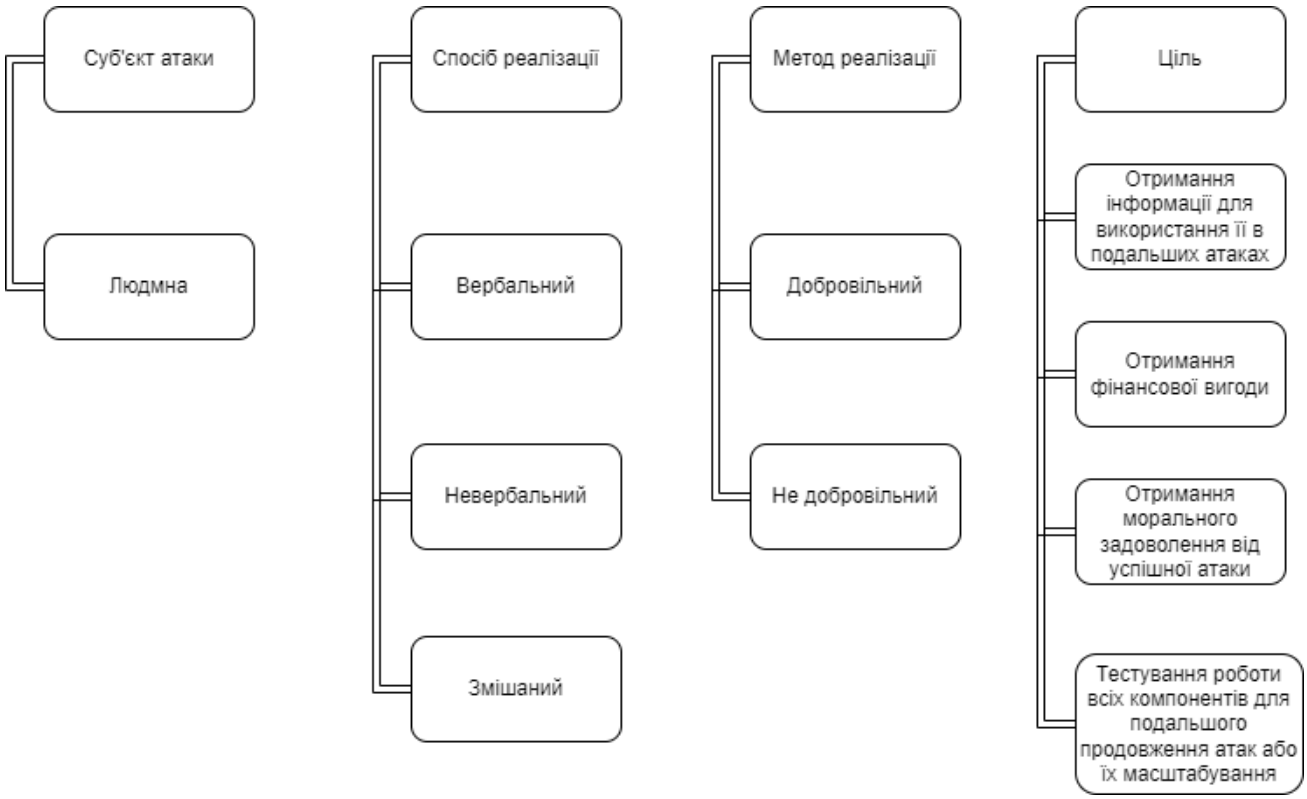


Рисунок 1.9 - Класифікація атак соціальної інженерії із використанням штучного інтелекту

На рисунку 1.9 зображено класифікацію, основними її компонентами є: Суб’єкт, Спосіб реалізації, Метод реалізації, Ціль. Розглянемо кожен із компонентів детальніше:

- Суб’єкт – в даній класифікації мається на увазі ціль зловмисника, враховуючи, що атаки соціальної інженерії використовують вразливості людей, то єдиним суб’єктом для цієї класифікації буде, власне людина.
- Спосіб реалізації – мається на увазі спосіб, яким користуватиметься зловмисник при діалозі з жертвою, вони поділяються на вербальний, тобто із використанням словесності, невербальний із використанням інших засобів передачі емоцій та змішаний, який містить в собі перший та другий відповідно. До вербального способу реалізації відноситься будь-яке спілкування засобами мови між зловмисником та суб’єктом атаки. Тобто, спілкування телефоном, SMS, текстовими повідомленнями в соцмережах, вживу (методи штучного інтелекту, можуть бути використаними для підготовки легенди зловмисника, аналізу даних про місце роботи

жертви , тощо). З невербальним методом трішки складніше, оскільки в основному атаки соціальної інженерії таки включають спілкування мовними засобами хоча б в якийсь спосіб, проте до прикладу при особистій присутності зловмисника перед жертвою він може подавати певні зрозумілі сигнали, які будуть ідентифіковані та спонукатимуть до виконання певних дій. В такий ж спосіб може використовуватись штучний інтелект – зловмисник може генерувати та надсилати/показувати різні зображення, сюжети яких будуть мати вплив на психологічно-емоційний стан суб'єкта. Про такі речі у своїх працях говорить стенфордський професор Браян Джеффрі Фогг, який вивчає соціальну складову поведінки. За його методом для реалізації певної поведінки потрібно, щоб спрацювали три компоненти: мотивація, тригер, та можливість. Тобто людина повинна мати мотивацію для виконання якоїсь дії, мати можливість її зробити та пройти тригерну точку, яка підштовхне її до дій [33]. У випадку атак соціальної інженерії зловмиснику може потрібним бути спонукати певного суб'єкта наприклад до агресії, уявимо, що зловмисник знає, що у жертви є певні особисті проблеми – тобто мотивація «випустити пар» присутня, до нього телефонують по роботі та вимагають терміново зробити якесь завдання – це можна розцінювати, як можливість залишається тільки тригер, а ним може виступити згенероване та надіслане зображення, як його улюблена футбольна команда сьогодні ввечері програє в принциповому протистоянні, як наслідок – агресивність спровоковано, а далі зловмисник використовуватиме цей стан у своїх цілях.

- Метод реалізації – мається на увазі саме психологічний метод для здійснення цілі. Добровільний метод – в ході його реалізації зловмиснику вдається переконати жертву стати на бік зловмисника. Добровільним він вважається оскільки жертва свідомо сприйматиме, це рішення власним та матиме бажання до виконання певних дій, до яких її спонукає зловмисник. Найчастіше для цього використовується нав'язування якихось переконань, пропонування якихось вигод, обміну або навіть викликані емпатичні почуття у суб'єкта. До недобровільного відносяться такі методи впливу на жертву як залякування, шантаж, погрози. Тобто зловмисник, вимушує жертву виконати певні дії, оскільки в протилежному випадку він завдасть їй певної шкоди.

- Останнім елементом класифікації є ціль, яку переслідує зловмисник, в основному всі ці цілі можна описати такими категоріями:

А) Отримання інформації для використання її в подальших атаках – інформація, яку може отримати зловмисник в ході реалізації атаки не обов'язково представлятиме для нього високу цінність, проте вона також може бути використане повторно для атаки на цього ж суб'єкта, або на когось іншого.

Б) Отримання фінансової вигоди – часто зловмисник атакують із метою привласнення собі чужих активів, зокрема грошей, цінних паперів, інформаційних активів, які потім можна збути та заробити кошти.

В) Отримання морального задоволення від здійснення атаки – багатьом хакерам подобається відчуття після успішно реалізованої атаки соціальної інженерії, дехто з них відчуває себе кращим за інших, підживлює своє его, тощо. Також, для деяких із них це є проявом певної життєвої позицію(бунтарства, прихильності до анархізму, тощо), а дехто може навіть жертвувати отримані від виконання атаки вигоди на благодійність. Бажання у них можуть бути цілком різними, проте ціль можна чітко виразити у одну.

Г) Тестування роботи всіх компонентів для подальшого продовження атак або їх масштабування – зловмисники можуть атакувати з такою ціллю, оскільки їх методи та технології також постійно змінюються, та потребують ретельної перевірки. До прикладу, якщо зловмисник додав новий вузол штучного інтелекту в свою систему, йому варто перевірити спочатку коректність його роботи, часто це робиться локально симулюючи поведінку жертви, проте також може виникнути потреба повторити тестування в реальних умовах, коли потрібно швидко та ефективно приймати рішення.

Послуговуючись даної класифікацією можна достатньо чітко описати атака соціальної інженерії із використанням штучного інтелекту. А також вона дозволяє звернути увагу на всіх необхідні аспекти, які допоможуть фахівцям встановити тип атаки, потенційні наслідки та протидіяти таким атакам в майбутньому.

Висновки до розділу 1

У цьому розділі було розглянуто основні методи атаки соціальної із використанням штучного інтелекту, та розглянуто статистичні дані, які доводять, що темпи розвитку механізмів захисту поки не співпадають із кількістю нових атак. Для зменшення успішності цих атак, потрібно постійно впроваджувати нові та вдосконалювати вже наявні технічні рішення, а також навчати людей виявляти ці атаки та зменшувати ймовірність їх успішного виконання. Проведення атак із використанням штучного інтелекту, а також кількість їх видів постійно збільшуватиметься тож питання розробки максимально ефективних систем має бути вирішене фахівцями в найкоротші терміни. Цього можна досягти консолідацією сил різних компаній, проте, на жаль, це малоймовірно, оскільки потрібно якось реалізувати це скупчення ідей та коштів в одному місці, а для цього компаніям, можливо доведеться призупинити власні розробки. Також потрібно пам’ятати, що на розробку методів захисту, компанії витрачають значні фінансові ресурси і планують їх повернення із певним відсотком у майбутньому. Тож, ця задача переходить до фахівців з кібербезпеки і для її вирішення варто розробити чіткі моделі реагування на існуючі атаки, мінімізувати ризики та проводити постійні тренування із обізнаності користувачів інформаційних систем забезпечивши тим самим підняття рівня їх грамотності та кібернетичної гігієни.

Також в розділі запропоновано розбиття кібератак із використанням соціальної інженерії та штучного інтелекту на етапі фреймворку Cyber Kill Chain. Таке розбиття повинне допомогти аналітикам кібербезпеки краще розуміти використання технологій штучного інтелекту в атаках та як наслідок впроваджувати більше ефективні методу захисту.

Крім цього в цьому розділі було проаналізовано основні сценарії атак соціальної інженерії та проаналізовано один із найпопулярніших методів класифікації цих атак. Також було представлено власну класифікацію атак соціальної інженерії із використанням штучного інтелекту, в вигляді докладної схеми із описом усіх її елементів. Ця класифікація має спростити опис та класифікацію наявних та нових

атак, та полегшити процес ідентифікації схожих атак в майбутньому. За її допомогою можна повно та зрозуміло класифікувати атаки соціальної інженерії із використанням штучного інтелекту.

Виходячи із вище сказаного можна ствердно сказати – що наявність методу, який буде здатним ефективно виявляти атаки соціальної інженерії із використанням штучного інтелекту суттєво зможе підвищити стан захищеності людей та зменшити ризики бути успішно атакованими.

РОЗДІЛ 2
РЕАЛІЗАЦІЯ АТАКИ СОЦІАЛЬНОЇ ІНЖЕНЕРІЇ ІЗ ВИКОРИСТАННЯМ ШТУЧНОГО ІНТЕЛЕКТУ, МЕХАНІЗМУ ЇЇ ВИЯВЛЕННЯ ТА ПРОТИДІЇ

2.1 Реалізація атаки

Серед фахівців кібербезпеки прийнято вважати, що для того, щоб розуміти від чого потрібно захищатись, та які методи для цього будуть найбільш дієвими варто, спершу спробувати реалізувати таку атаку на свою інформаційну систему. В основному, як було згадано в розділі раніше, зловмисники використовують фішинг, як метод соціальної інженерії для досягнення своїх цілей. Спробуємо реалізувати такий підхід на практиці в локальній підмережі домашнього інтернету. Для цього потрібно створити атакувальне середовище – в цьому випадку воно міститиме в собі три основних компоненти, через які будуть створюватись оброблятимуться та надсилатимууться фішингові повідомлення. Перед безпосереднім описом принципу роботи кожного із компонентів атаки, було графічно зображено схему цієї атаки на рисунку 2.1.

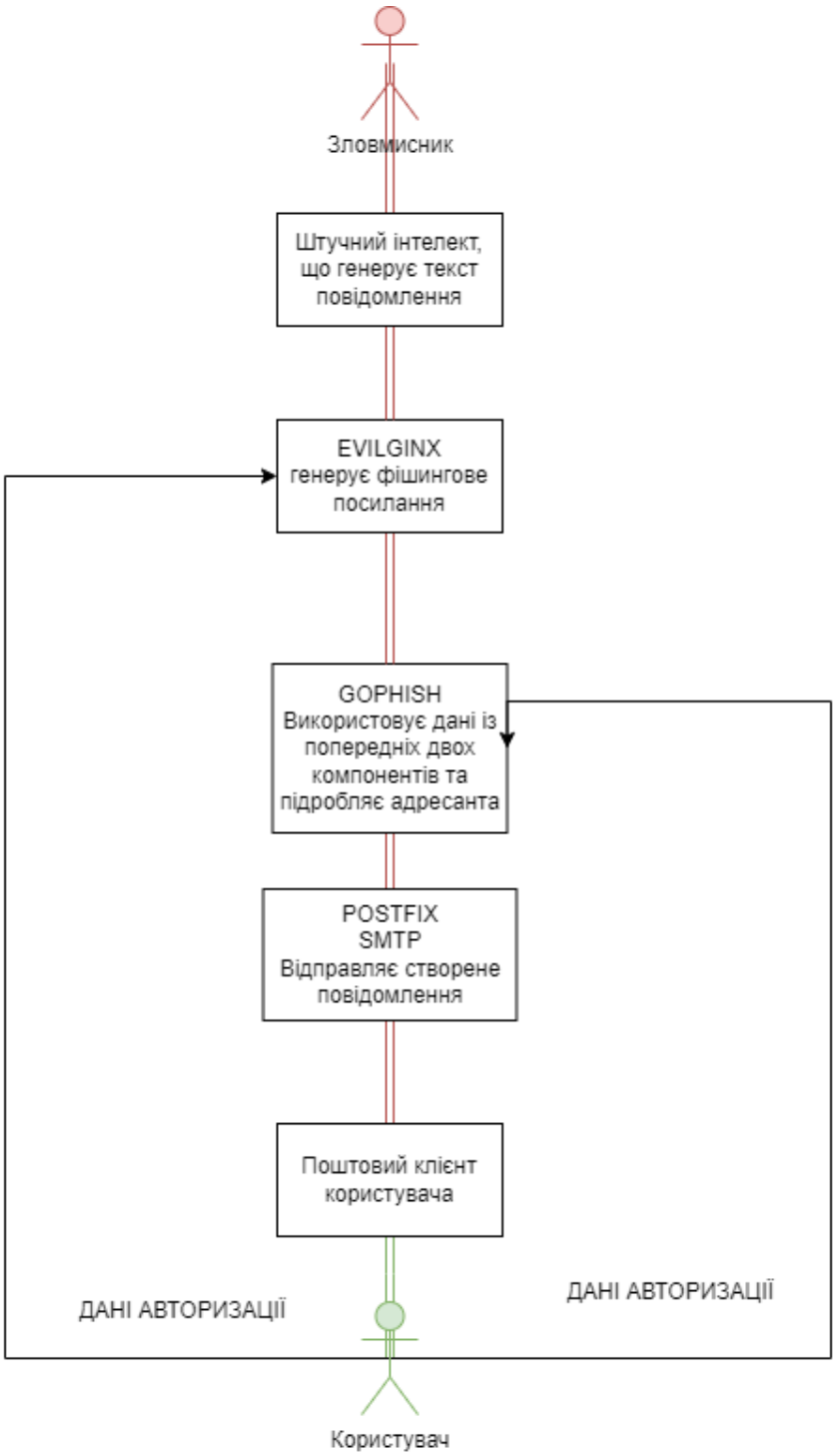


Рисунок 2.1 - Схема атаки соціальної інженерії із використанням штучного інтелекту

Першим із цих компонентів буде власний поштовий сервіс. В даній роботі використовуватиметься Postfix – це агент передачі пошти із відкритим кодом, тобто для його використання не потрібно купувати ніякої додаткової ліцензії [34]. Його перевагою є те, що його також можна налаштувати як власний SMTP (простий

протокол передачі електронних листів) сервер, що знадобиться в цій атаці. Даний сервер буде налаштоввано із пареметром «Internet site», це означає, що електронні листи будуть відправлятись та отримуватись безпосередньо через SMTP, відображено на рисунку 2.2.

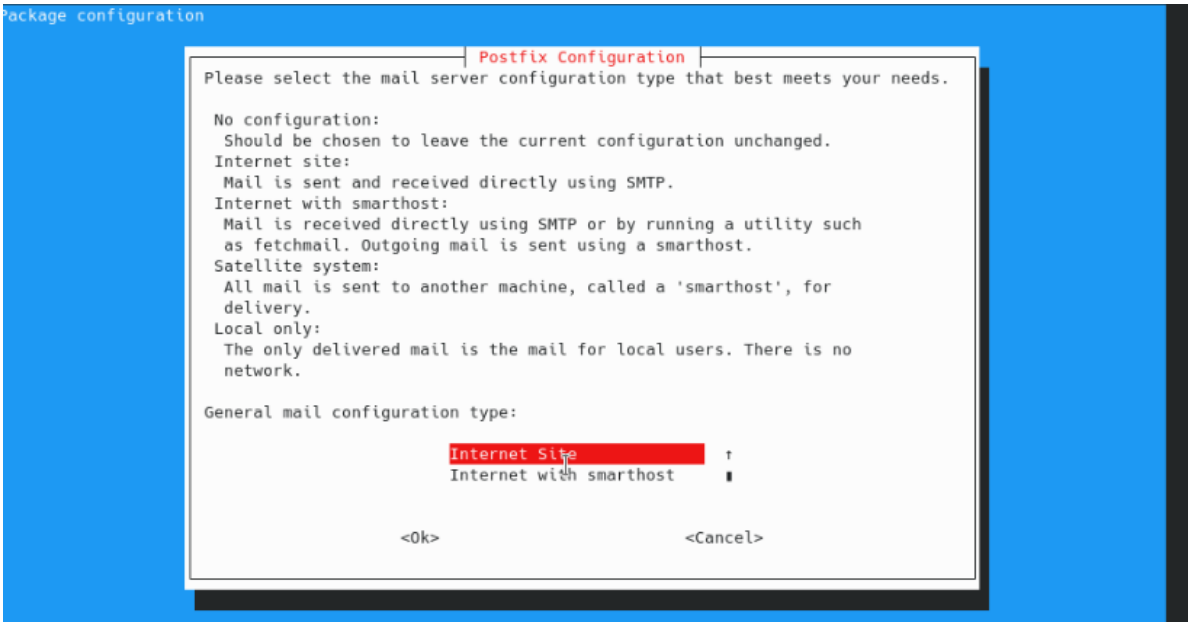


Рисунок 2.2 – налаштування Postfix для роботи через SMTP

Наступним компонентом є Gophish. Gophish – це фреймворк для фішингу із відкритим вихідним кодом. Фахівці із інформаційної безпеки використовують його для тестування рівня обізнаності працівників в компаніях, кількості інтеракцій із фішинговими повідомленнями [35]. Розглянемо основні налаштування цього компоненту:

1) Першим пунктом є «Users and Groups». В цьому підменю налаштовується список користувачів, які в подальшому отримають фішингове повідомлення. Також, для зручності в подальшому аналізі результатів в ньому можна додати не тільки саму електронну пошту, а також прізвище, ім’я та посаду. Приклад додавання такого профілю зображено на рисунку 2.3

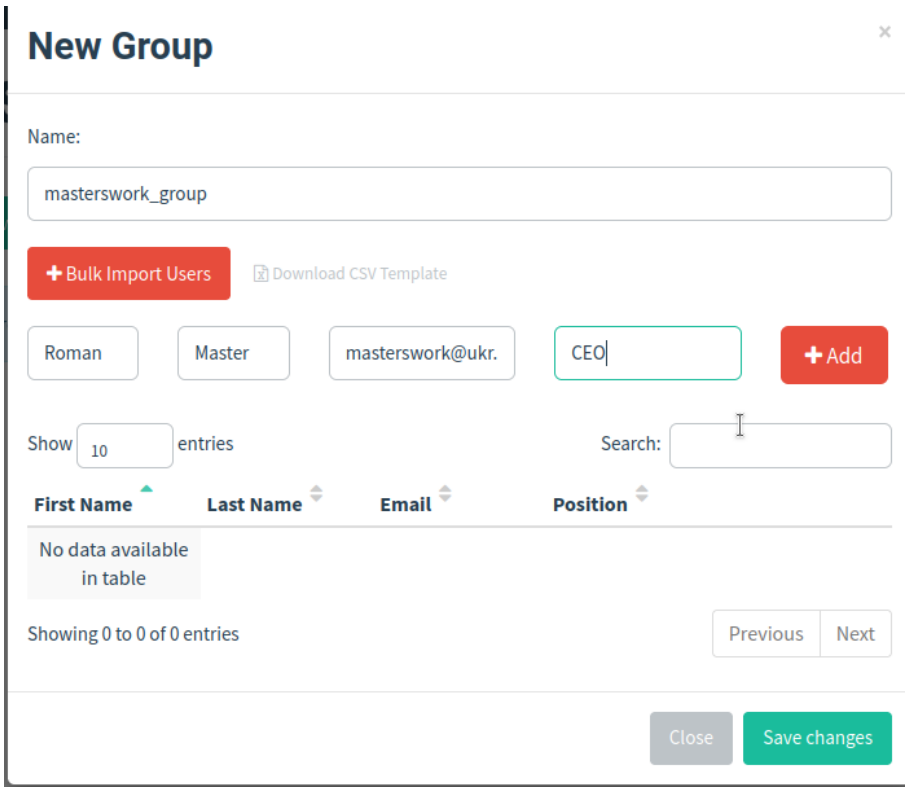


Рисунок 2.3 – Налаштування підпункту меню «Users and groups”

2) Наступним кроком потрібно налаштувати “Email template”. Цей підпункт відповідає за створення шаблону листа, який потім надсилатиметься користувачу. Також в ньому конфігурується підроблена поштова адреса, яку можуть відображати певні поштові клієнти, які потенційно можуть бути основними в жертви. Також можна додати до листа будь-які вкладення, щоб зробити його вигляд ще більш легітимним. Особливу увагу варто звернути на прапорець «Add tracking image» це дозволяє додати до вкладення трекер, який зможе сповістити, що цей файл було відкрито. Текст листа вкладеного в цей шаблон було написано за допомогою штучного інтелекту та з додаванням прохання відповісти на ці запитання із пошти, з якої було зроблено замовлення(припустимо для цієї атаки, що такою була в домені outlook.com) відображено на рисунку 1.3. Приклад конфігурації для демонстраційних можливостей цієї роботи міститься на рисунку 2.4

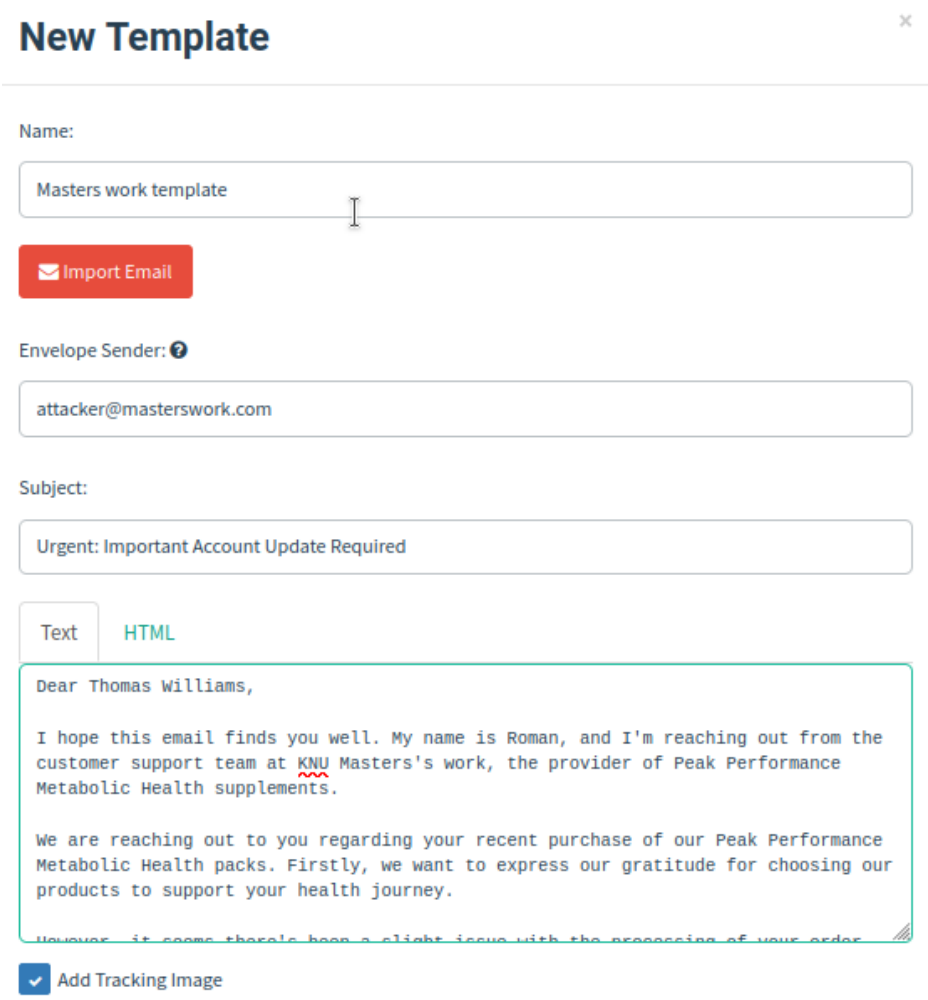


Рисунок 2.4 - Налаштування «Email Template»

3) Наступним в списку необхідних налаштувань є “Sending profiles”. Тут потрібно налаштувати профіль, від якого насправді відправлятимуться повідомлення, а також для цього потрібний доступ до SMTP серверу з якого повідомлення буде відправлятись(в випадку цієї атаки це налаштований Postfix сервер). Також при необхідності можна доналаштувати додаткові заголовки електронних листів. Приклад конфігурації зображено на рисунку 2.5

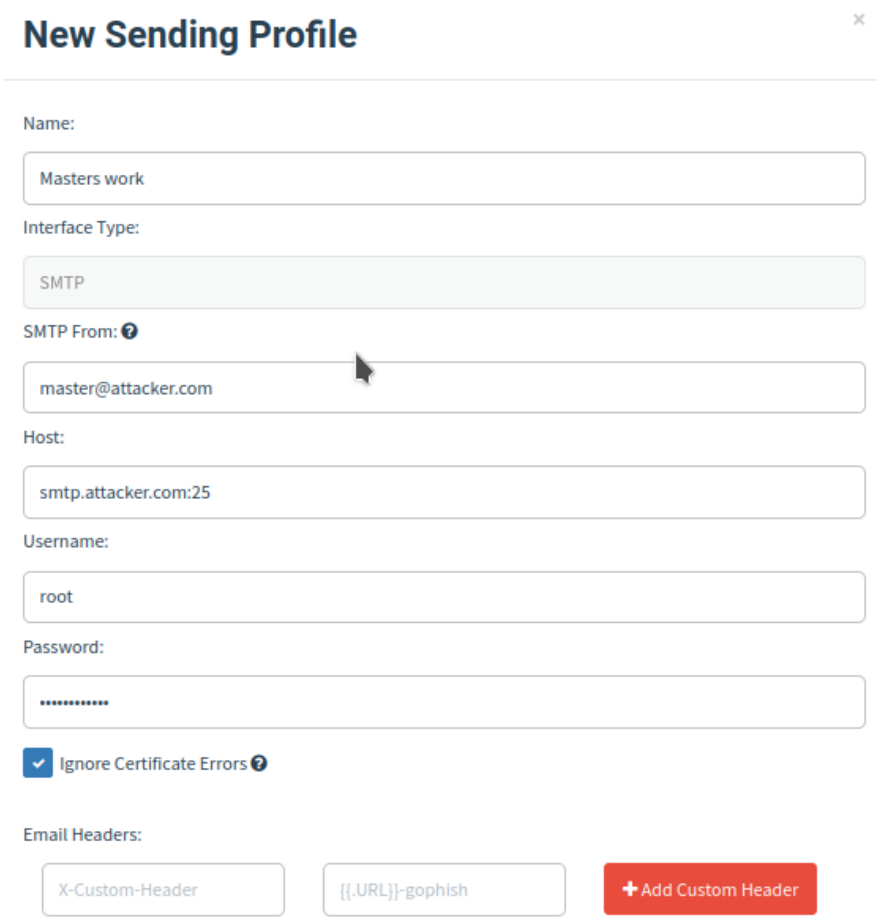


Рисунок 2.5 – Налаштування «Sending Profiles»

4) Останнім налаштуванням в цьому компоненті є «Landing Pages”. Тут налаштовується вигляд фішингової сторінки, з якою буде взаємодіяти отримувач. На цьому кроці можна або вставити посилання на бажану веб-сторінку, яку сервіс може скопіювати сам, або написати мовою гіпертекстової розмітки (HTML). Для цієї атаки буде використано перший варіант, проте для отримання бажаної сторінки необхідно встановити третю компоненту. Після її успішного встановлення потрібно буде повернутись до цього пункту та внести зміни. Також на цьому етапі є можливість обрати чи потрібно буде збирати паролі, та куди перенаправити користувача після закінчення інтеракції із сторінкою. Приклад конфігурації на цьому етапі зображено на рисунку 2.6.

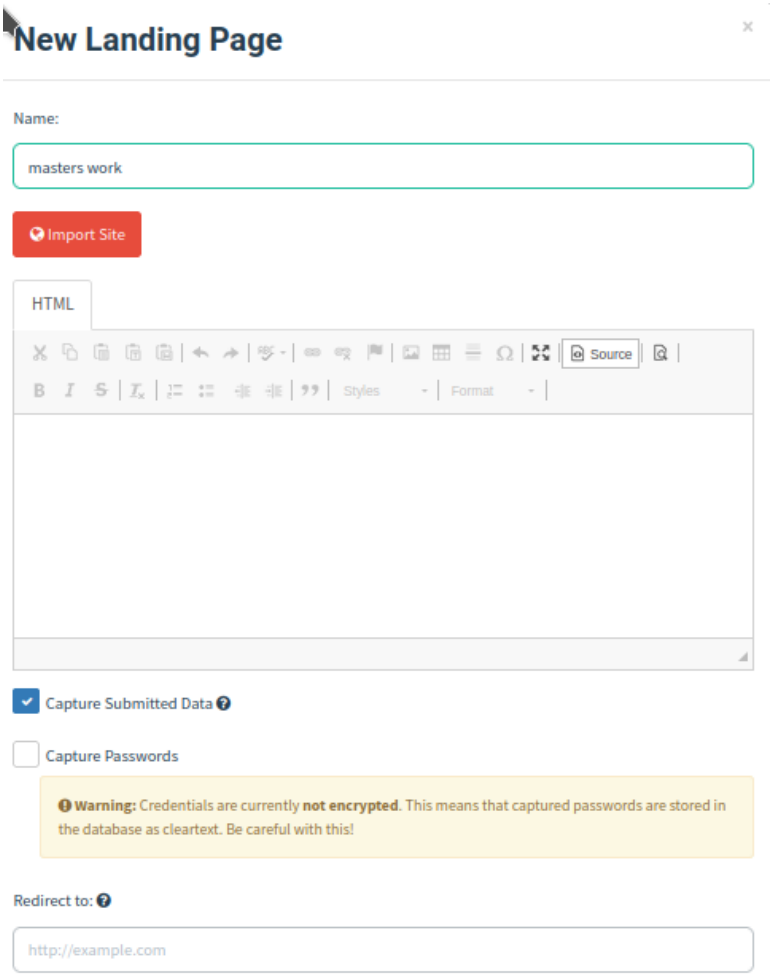


Рисунок 2.6 - Налаштування «Landing Pages»

Після виконання перелічених налаштувань, потрібно перейти до третьої компоненти нею є Evilginx – це фреймворк, який широко використовується не тільки спеціалістами із захисту інформації, але й також зловмисниками [36]. Цей компонент дозволяє створювати фішингові посилання, які матимуть ідентичний вигляд із легітимними, має в собі кілька вже підготовлених шаблонів та дозволяє написати власний використовуючи для цього YAML [37]. На рисунку 2.7 зображено встановлений Evilginx.

Рисунок 2.7 – Встановлений Evilnginx

В даній атаці, як раніше було зазначено буде використовуватись шаблон для пошти outlook.com. Проведемо всі необхідні для цього налаштування:

1) Потрібно сконфігурувати IPv4 адресу, на якій міститиметься фішингове посилання, в цій атаці, оскільки вона проводиться тількит із метою демонстрації не будуть використовуватись куплені доменні імена, тож скористаємось локальною мережею. Налаштування відображене на рисунку 2.8

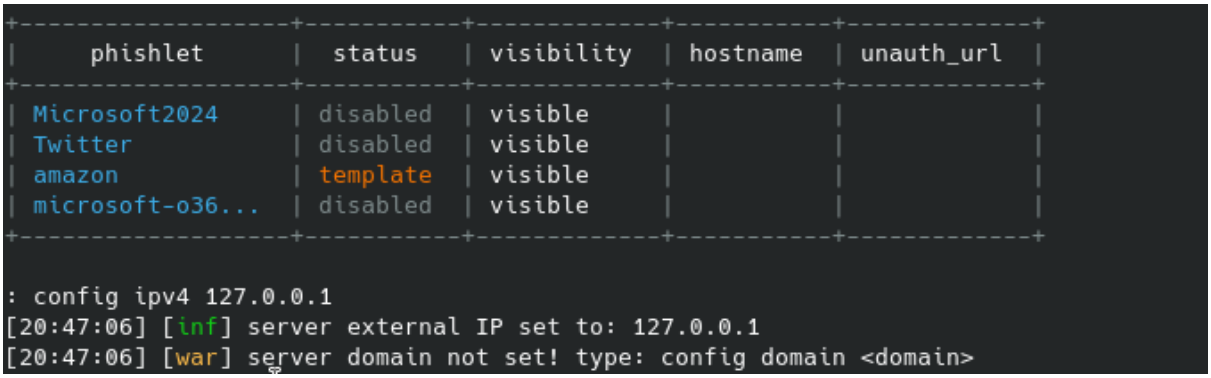


Рисунок 2.8 - Налаштування IP адреси

2) Наступним кроком потрібно налаштувати домен, як можна помітити із підказки після виконання попереднього пункту. Як доменне ім’я для демонтсрації було обрано «romanmastersknu.com», зображено на рисунку 2.9.

Рисунок 2.9 – Налаштування доменного імені

3) Далі потрібно обрати шаблон(в цьому фреймворку вони мають назву «phishlet» та додати його під використання заданого раніше домену, результат відображено на рисунку 2.10.

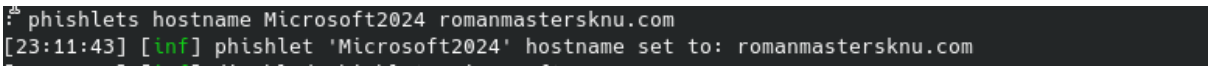


Рисунок 2.10 - Використання доменного імені для певного шаблону

4) Після обрання шаблону необхідно його власне активувати та переглянути список хостів, які були для нього створені. Результат зображено на рисунку 2.11

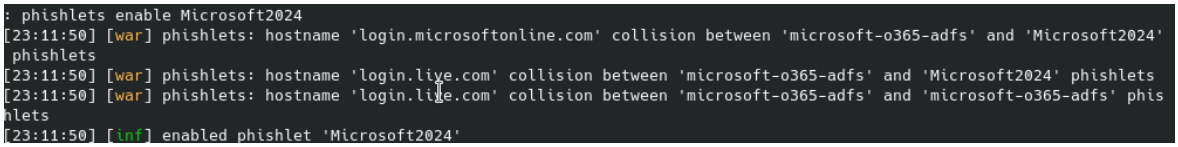


Рисунок 2.11 - Активування шаблону

5) Далі потрібно перейти власне кажучи до створення фішингового посилання, в цьому фреймворку воно має назву «lure». Результат створення фішингового посилання відображено на рисунку 2.12

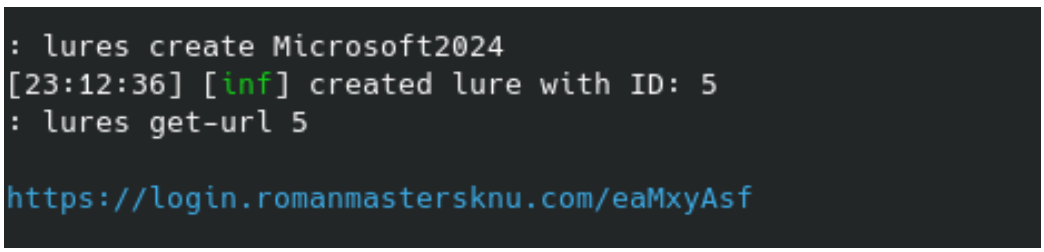


Рисунок 2.12 – Створення фішинговго посилання

6) Далі потрібно додати це посилання, як було згадано раніше до налаштувань «Landing Pages» на Gophish та провести атаку. Після того, як посилання

додане, все що залишається це власне ініціалізація фішингової кампанії, використовуючи всі попередньо задані налаштування. Це робиться у підменю “Campaigns” – відображено на рисунку 2.13

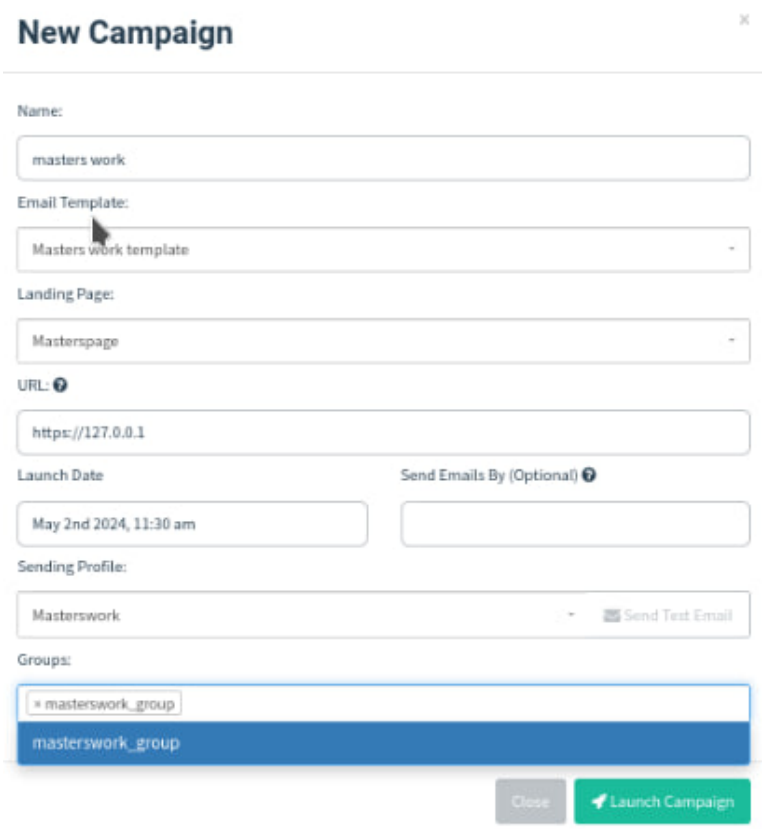


Рисунок 2.13 - Запуск фішинговгої кампанії

Після запуску кампанії відкривається її можна переглядати в режимі реального часу, та бачити як користувач взаємодіяв з листом, чи він був відкритий, чи було здійснено перехід за посиланням, або можливо, атаку було виявлено та повідомлено про неї через опцію «report», яка присутня в більшості клієнтів. Це меню відображено на рисунку 2.14

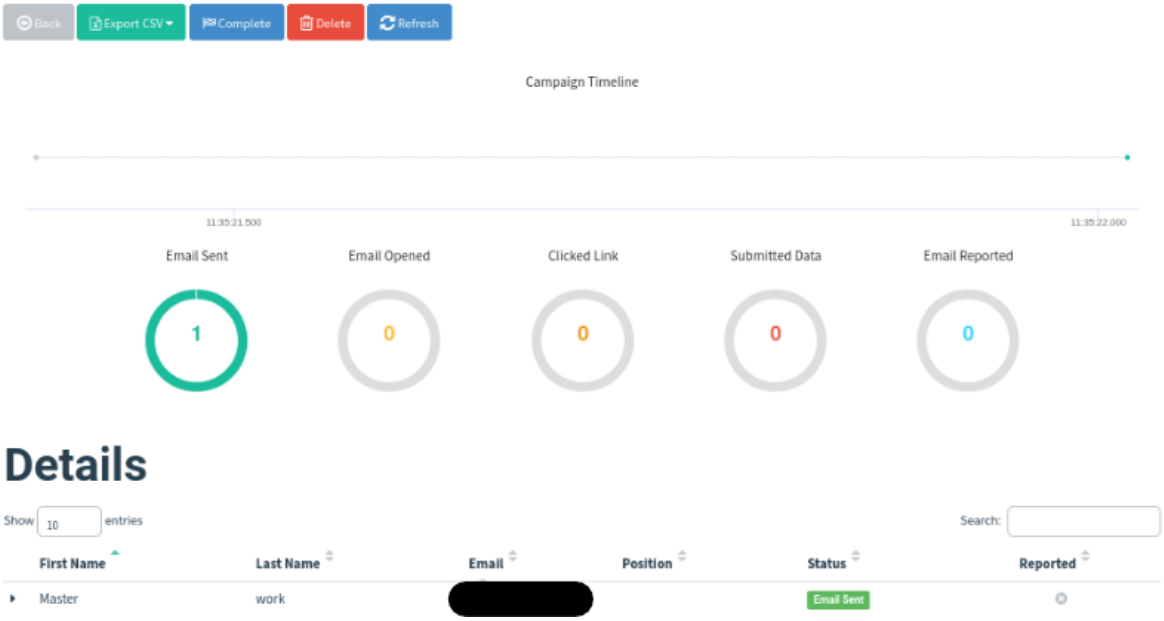


Рисунок 2.14 - Підменю для перегляду інформації про каманію

7) Після виконання усіх цих дій, все що залишається атакуючій стороні – це чекати, та моніторити камапанію, або Evilnginx, оскільки в ньому теж було налаштовано можливість збирати дані. Якщо користувач відкриє лист/посилання, проте не надасть жодної бажаної для зловмисника інформації це в будь-якому випадку дасть йому розуміння того, шо потрібно щось змінити, чи покращити. В демонстраційних цілях, просимулюємо активність користувача, та введемо не справжні авторизаційні дані(якщо ввести пошту, якої насправді не існує, то користувачу виведеться на екран повідомлення, про те, що такого користувацького аккаунту не існує). На рисунку 2.15. зображено перехоплені авторизаційні дані користувача

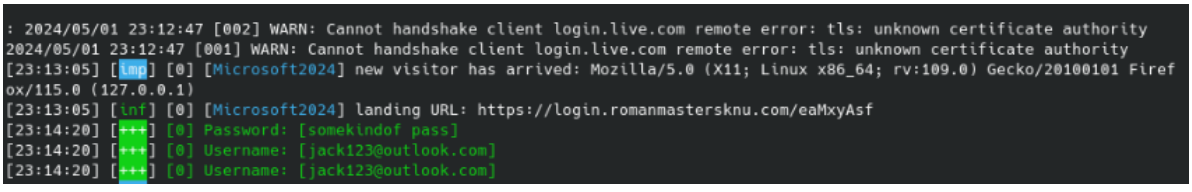


Рисунок 2.15 - Перехоплення даних жертви

Інформацію про те, як саме користувач взаємодіяв з листом відображено на рисунку 2.16.

Рисунок 2.16 - Інформація про перебіг кампанії

Після успішної реалізації атаки можна ствердно сказати, що сучасні методи захисту електронної пошти не є достатньо ефективним. Оскільки фільтр повідомлень не зміг ані сповістити користувача про підозрілий вміст фішингового листа, а ні про посилання, яке очевидно є фішинговим.

2.2. Реалізація методу захисту від атак соціальної інженерії із використанням штучного інтелекту.

В минулому підрозділі було показано приклад атаки соціальної інженерії із використанням штучного інтелекту на користувача, шляхом комунікації із ним засобами електронної пошти. В цьому ж підрозділі мова піде про те, як захиститись від такої атаки.

Реалізація цього методу може використовуватись як для захисту персонального пристрою так і корпоративних, відмінність тільки буде в її масштабності. Розглянемо приклад саме для персонального комп’ютера, на якому встановлена операційна система Ubuntu (linux) та безкоштовний поштовий клієнт Thunderbird. Цей застосунок є достатньо легким в користуванні та адмініструванні, та є досить широко використовуваним по всьому світі, його активними користувачами є близько 20 мільйонів людей [38]. Метод передбачає взаємодію із листами іншими утилітами, які будуть описані згодом, для цього потрібно реалізувати доступ до таких листів, це можна зробити реалізувавши не складний фільтр листів, який копіюватиме їх в теку на пристрої кожні N кількість хвилин. Цей фільтр відображено на рисунку 2.17.

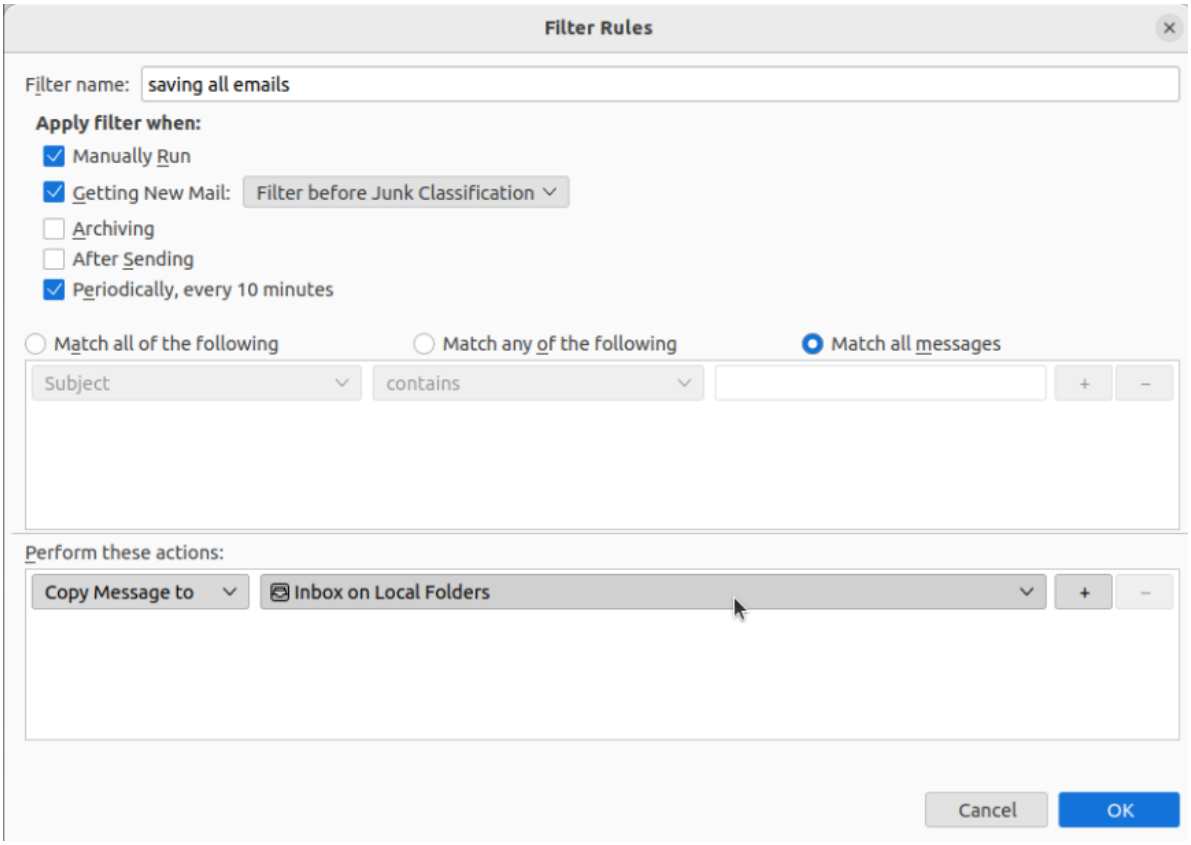


Рисунок 2.17 - Налаштування фільтру для збереження копій листів локально

Даний фільтр буде переглядати всі теки на вміст нових повідомлень кожні 10 хвилин, та копіюватиме їх до локальної директорії на пристрої. Рисунок 2.18 демонструє коректність налаштованого фільтру та можливість перегляду повідомлень, а також формат даних, в якому листи зберігатимуться.

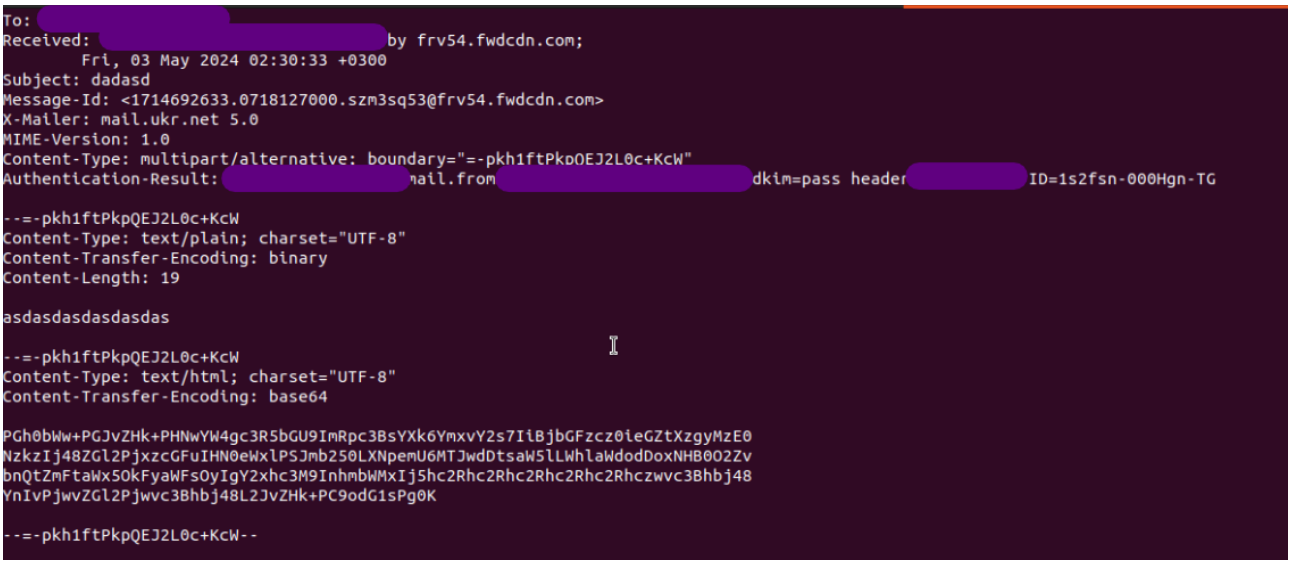


Рисунок 2.18 - Приклад роботи коректно налаштованого фільтру

Наступним кроком потрібно проаналізувати сам формат та виявити корисні в ньому поля. Метод захисту передбачатиме аналіз посилань та текстового наповнення повідомлення, тому ці поля представлятимуть найбільший інтерес та потрібно реалізувати можливість їх обробки окремо. Для цього буде використано ще один продукт із відкритим кодом на основі Python email-analyzer [39], за його допомогою можна відсортувати поля файлу електронного листа в зручному для форматі читання. На рисунку 2.19 зображено приклад обробки електронного листа за допомогою вищеописаного скрипта.

Рисунок 2.19 - Приклад парсингу електронного листа

Наступним компонентом для реалізації захисту користувача є модель штучного інтелекту, яка була навчена вирізняти фішингові посилання. Вона написана із використанням відкритої бібліотеки для машинного навчання Tensorflow та була викладена у відритий доступ для фахівців із інформаційної безпеки студентом Стендфордського університету Заком Петрі. Дана модель має точність 97.68%, зображено на рисунку 2.20 [40]

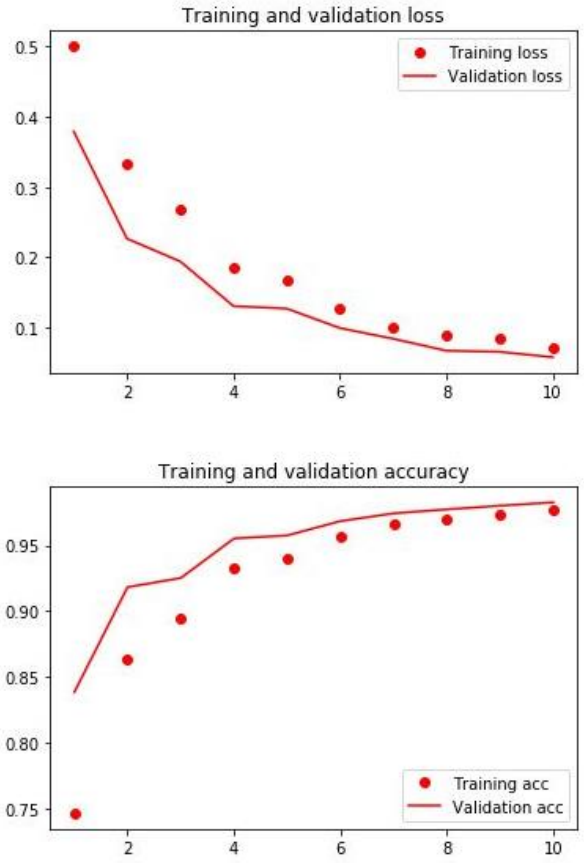


Рисунок 2.20 - Точність навчання моделі для виявлення фішингових посилань

Модель складається із таких основних файлів: (зображено на рисунку 2.21)

Рисунок 2.21 - вміст теки із моделлю для визначення фішингового посилання

Перший файл bi-lstmchar256256128.h5 – це власне кажучи сама модель, натренована на виявлення та викладена в відкритий доступ автором. Його вміст відображено на рисунку 2.22

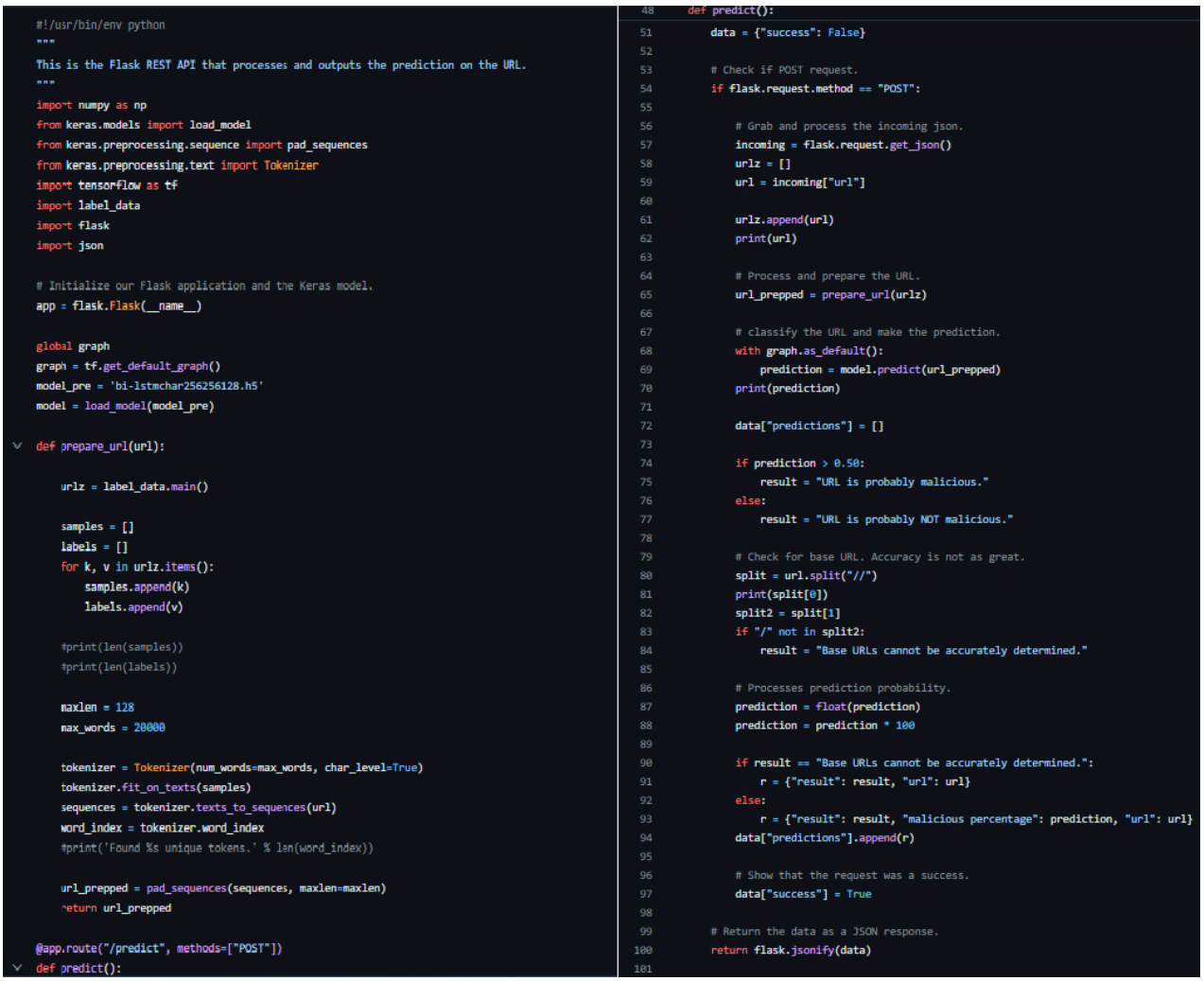
Label_data.py – файл, в якому містить скрипт, який обробляє дані із датасету перед передачею його моделі, відображено на рисунку 2.23.

Code Blame 28 lines (20 loc) · 595 Bytes

```
#!/usr/bin/env python
"""
This file gathers data to be used for pre-processing in training and prediction.
"""
import pandas as pd

def main():

    blacklist = 'phishing_database.csv'
    whitelist = 'whitelist.txt'

    urls = {}

    blacklist = pd.read_csv(blacklist)

    #Assign 0 for non-malicious and 1 as malicious for supervised learning.
    for url in blacklist['url']:
        urls[url] = 1

    with open(whitelist, 'r') as f:
        lines = f.read().splitlines()
        for url in lines:
            urls[url] = 0

    return urls

if __name__ == "__main__":
    main()
```

Рисунок 2.22 - Вміст файлу label_data.py

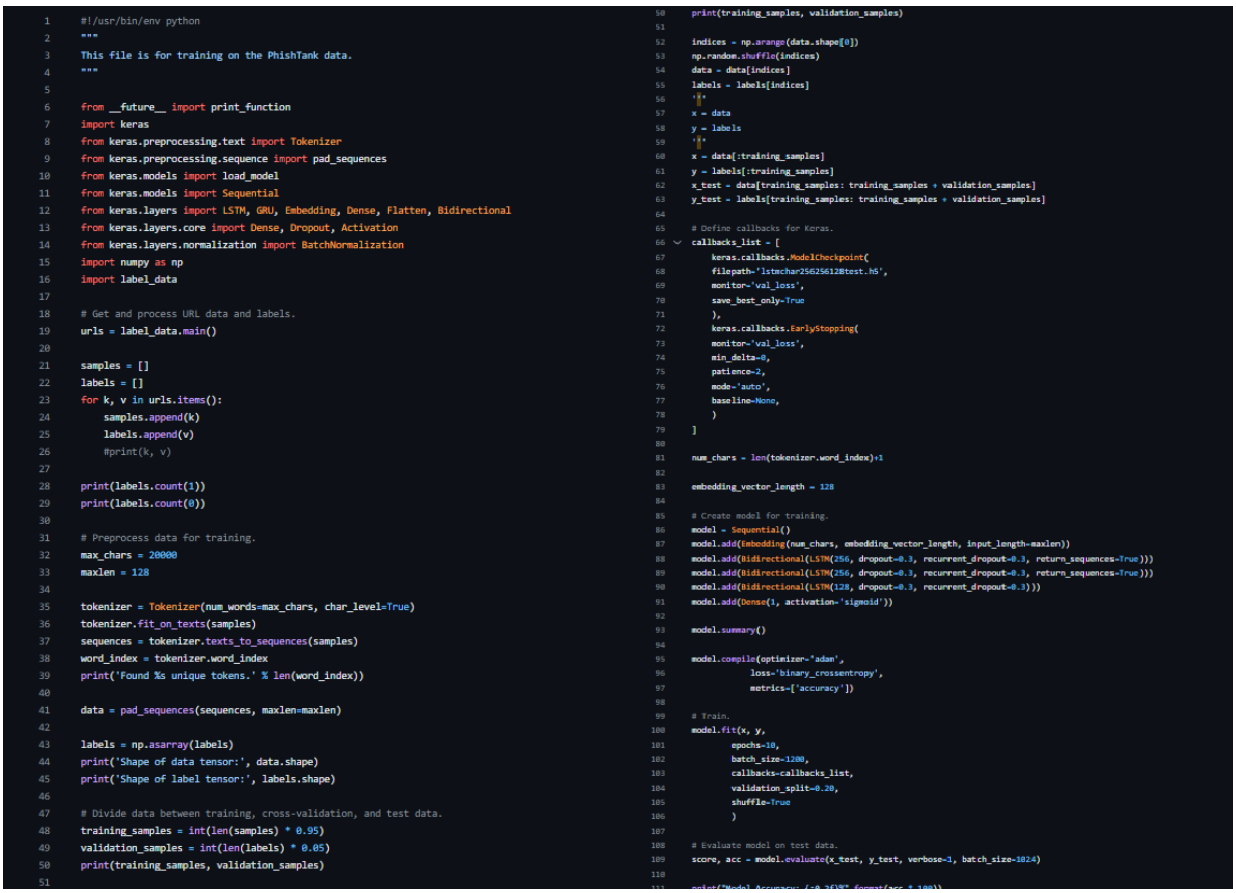
Flaskrestapi.py – це API, який обробляє та виводить результат після того, як модель винесла свій вердикт, відображено на рисунку 2.24.



```
#!/usr/bin/env python
"""
This is the Flask REST API that processes and outputs the prediction on the URL.
"""
import numpy as np
from keras.models import load_model
from keras.preprocessing.sequence import pad_sequences
from keras.preprocessing.text import Tokenizer
import tensorflow as tf
import label_data
import flask
import json

# Initialize our Flask application and the Keras model.
app = flask.Flask(__name__)

global graph
graph = tf.get_default_graph()
model_pre = 'bi-lstmchar256256128.h5'
model = load_model(model_pre)

def prepare_url(url):

    urlz = label_data.main()

    samples = []
    labels = []
    for k, v in urlz.items():
        samples.append(k)
        labels.append(v)

    #print(len(samples))
    #print(len(labels))

    maxlen = 128
    max_words = 20000

    tokenizer = Tokenizer(num_words=max_words, char_level=True)
    tokenizer.fit_on_texts(samples)
    sequences = tokenizer.texts_to_sequences(url)
    word_index = tokenizer.word_index
    #print('Found %s unique tokens.' % len(word_index))

    url_prepped = pad_sequences(sequences, maxlen=maxlen)
    return url_prepped

@app.route("/predict", methods=["POST"])
def predict():
    data = {"success": False}

    # Check if POST request.
    if flask.request.method == "POST":

        # Grab and process the incoming json.
        incoming = flask.request.get_json()
        urlz = []
        url = incoming["url"]

        urlz.append(url)
        print(url)

        # Process and prepare the URL.
        url_prepped = prepare_url(urlz)

        # classify the URL and make the prediction.
        with graph.as_default():
            prediction = model.predict(url_prepped)
        print(prediction)

        data["predictions"] = []

        if prediction > 0.50:
            result = "URL is probably malicious."
        else:
            result = "URL is probably NOT malicious."

        # Check for base URL. Accuracy is not as great.
        split = url.split("//")
        print(split[0])
        split2 = split[1]
        if "/" not in split2:
            result = "Base URLs cannot be accurately determined."

        # Processes prediction probability.
        prediction = float(prediction)
        prediction = prediction * 100

        if result == "Base URLs cannot be accurately determined.":
            r = {"result": result, "url": url}
        else:
            r = {"result": result, "malicious percentage": prediction, "url": url}
        data["predictions"].append(r)

        # Show that the request was a success.
        data["success"] = True

    # Return the data as a JSON response.
    return flask.jsonify(data)
```

Рисунок 2.23 - Вміст файлу flasrestapi.py

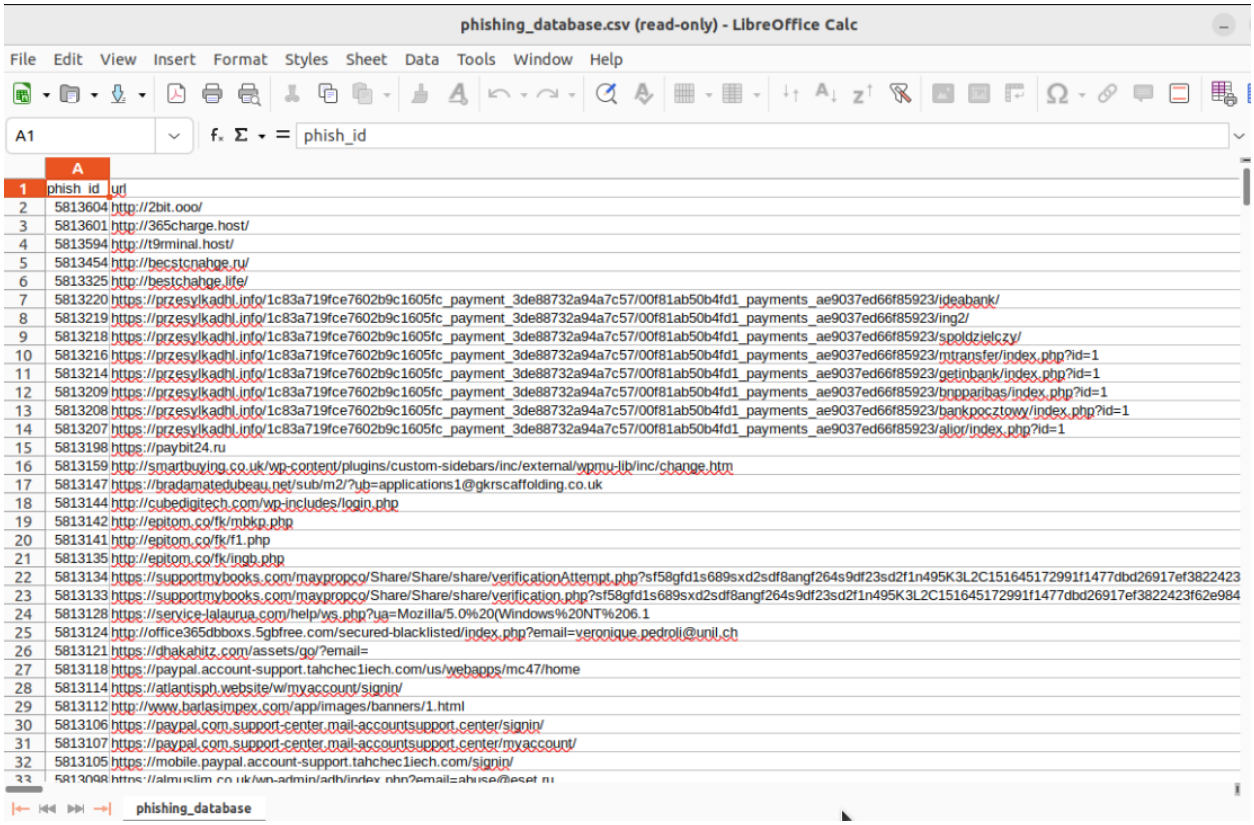
Train.py – файл, що використовується для навчання моделі на даних, які опписані нижче, відображено на рисунку 2.25. Власне саме навчання відбувається за допомогою відкритої нейромережної бібліотеки Keras, яка була створена французьким розробником програмного забезпечення Франсуа Шолле [41]. Під час виконання даного програмного коду відбувається підготовка даних, створюється модель для навчання та власне відбувається і саме навчання. Після зазначеного вище також відбувається оцінка навчання.



```python
#!/usr/bin/env python
"""
This file is for training on the PhishTank data.
"""

from __future__ import print_function
import keras
from keras.preprocessing.text import Tokenizer
from keras.preprocessing.sequence import pad_sequences
from keras.models import load_model
from keras.models import Sequential
from keras.layers import LSTM, GRU, Embedding, Dense, Flatten, Bidirectional
from keras.layers.core import Dense, Dropout, Activation
from keras.layers.normalization import BatchNormalization
import numpy as np
import label_data

# Get and process URL data and labels.
urls = label_data.main()

samples = []
labels = []
for k, v in urls.items():
    samples.append(k)
    labels.append(v)
    #print(k, v)

print(labels.count(1))
print(labels.count(0))

# Preprocess data for training.
max_chars = 20000
maxlen = 128

tokenizer = Tokenizer(num_words=max_chars, char_level=True)
tokenizer.fit_on_texts(samples)
sequences = tokenizer.texts_to_sequences(samples)
word_index = tokenizer.word_index
print('Found %s unique tokens.' % len(word_index))

data = pad_sequences(sequences, maxlen=maxlen)

labels = np.asarray(labels)
print('Shape of data tensor:', data.shape)
print('Shape of label tensor:', labels.shape)

# Divide data between training, cross-validation, and test data.
training_samples = int(len(samples) * 0.95)
validation_samples = int(len(labels) * 0.05)
print(training_samples, validation_samples)

print(training_samples, validation_samples)

indices = np.arange(data.shape[0])
np.random.shuffle(indices)
data = data[indices]
labels = labels[indices]
'''
x = data
y = labels
'''
x = data[:training_samples]
y = labels[:training_samples]
x_test = data[training_samples: training_samples + validation_samples]
y_test = labels[training_samples: training_samples + validation_samples]

# Define callbacks for Keras.
callbacks_list = [
    keras.callbacks.ModelCheckpoint(
    filepath='lstmchar256256128test.h5',
    monitor='val_loss',
    save_best_only=True
    ),
    keras.callbacks.EarlyStopping(
    monitor='val_loss',
    min_delta=0,
    patience=2,
    mode='auto',
    baseline=None,
    )
]

num_chars = len(tokenizer.word_index)+1

embedding_vector_length = 128

# Create model for training.
model = Sequential()
model.add(Embedding(num_chars, embedding_vector_length, input_length=maxlen))
model.add(Bidirectional(LSTM(256, dropout=0.3, recurrent_dropout=0.3, return_sequences=True)))
model.add(Bidirectional(LSTM(256, dropout=0.3, recurrent_dropout=0.3, return_sequences=True)))
model.add(Bidirectional(LSTM(128, dropout=0.3, recurrent_dropout=0.3)))
model.add(Dense(1, activation='sigmoid'))

model.summary()

model.compile(optimizer='adam',
              loss='binary_crossentropy',
              metrics=['accuracy'])

# Train.
model.fit(x, y,
          epochs=10,
          batch_size=1200,
          callbacks=callbacks_list,
          validation_split=0.20,
          shuffle=True
          )

# Evaluate model on test data.
score, acc = model.evaluate(x_test, y_test, verbose=1, batch_size=1024)
```

Рисунок 2.24 - Вміст файлу train.py

За дата сет для навчання було взято базу такого відомого сервісу, як Phishtank, яким широко користуються не тільки фахівці із захисту інформації, а також ентузіасти, які дбають про свою безпеку [42]. Ці дані представляють собою великий набір із посилань, які містять в собі не тільки фішингові, але й легітимні phishing_datatbase.csv та whitelist.txt відповідно, за їх допомогою модель навчатиметься виносити вердикт, відображено на рисунку 2.25. та 2.26.



phish_id	url
5813604	http://2bit.ooo/
5813601	http://365charge.host/
5813594	http://t9rminal.host/
5813454	http://becstcnahge.ru/
5813325	http://bestchahge.life/
5813220	https://przesylkadhl.info/1c83a719fce7602b9c1605fc_payment_3de88732a94a7c57/00f81ab50b4fd1_payments_ae9037ed66f85923/ideabank/
5813219	https://przesylkadhl.info/1c83a719fce7602b9c1605fc_payment_3de88732a94a7c57/00f81ab50b4fd1_payments_ae9037ed66f85923/ing2/
5813218	https://przesylkadhl.info/1c83a719fce7602b9c1605fc_payment_3de88732a94a7c57/00f81ab50b4fd1_payments_ae9037ed66f85923/spoldzielczy/
5813216	https://przesylkadhl.info/1c83a719fce7602b9c1605fc_payment_3de88732a94a7c57/00f81ab50b4fd1_payments_ae9037ed66f85923/mtransfer/index.php?id=1
5813214	https://przesylkadhl.info/1c83a719fce7602b9c1605fc_payment_3de88732a94a7c57/00f81ab50b4fd1_payments_ae9037ed66f85923/getinbank/index.php?id=1
5813209	https://przesylkadhl.info/1c83a719fce7602b9c1605fc_payment_3de88732a94a7c57/00f81ab50b4fd1_payments_ae9037ed66f85923/bnpparibas/index.php?id=1
5813208	https://przesylkadhl.info/1c83a719fce7602b9c1605fc_payment_3de88732a94a7c57/00f81ab50b4fd1_payments_ae9037ed66f85923/bankpocztowy/index.php?id=1
5813207	https://przesylkadhl.info/1c83a719fce7602b9c1605fc_payment_3de88732a94a7c57/00f81ab50b4fd1_payments_ae9037ed66f85923/alior/index.php?id=1
5813198	https://paybit24.ru
5813159	http://smartbuying.co.uk/wp-content/plugins/custom-sidebars/inc/external/wpmu-lib/inc/change.htm
5813147	https://bradamatedubeau.net/sub/m2/?ub=applications1@gkrscaffolding.co.uk
5813144	http://cubedigitech.com/wp-includes/login.php
5813142	http://epitom.co/fk/mbkp.php
5813141	http://epitom.co/fk/f1.php
5813135	http://epitom.co/fk/ingb.php
5813134	https://supportmybooks.com/maypropco/Share/Share/share/verificationAttempt.php?sf58gfd1s689sxd2sdf8angf264s9df23sd2f1n495K3L2C151645172991f1477dbd26917ef3822423
5813133	https://supportmybooks.com/maypropco/Share/Share/share/verification.php?sf58gfd1s689sxd2sdf8angf264s9df23sd2f1n495K3L2C151645172991f1477dbd26917ef3822423f62e984
5813128	https://service-lalaurua.com/help/ws.php?ua=Mozilla/5.0%20(Windows%20NT%206.1
5813124	http://office365dbboxs.5gbfree.com/secured-blacklisted/index.php?email=veronique.pedroli@unil.ch
5813121	https://dhakahitz.com/assets/go/?email=
5813118	https://paypal.account-support.tahchec1iech.com/us/webapps/mc47/home
5813114	https://atlantisph.website/w/myaccount/signin/
5813112	http://www.barlasimpex.com/app/images/banners/1.html
5813106	https://paypal.com.support-center.mail-accountsupport.center/signin/
5813107	https://paypal.com.support-center.mail-accountsupport.center/myaccount/
5813105	https://mobile.paypal.account-support.tahchec1iech.com/signin/
5813098	https://almuslim.co.uk/wp-admin/adb/index.php?email=abuse@eset.ru

Рисунок 2.25 - Вміст файлу із фішинговими посиланнями

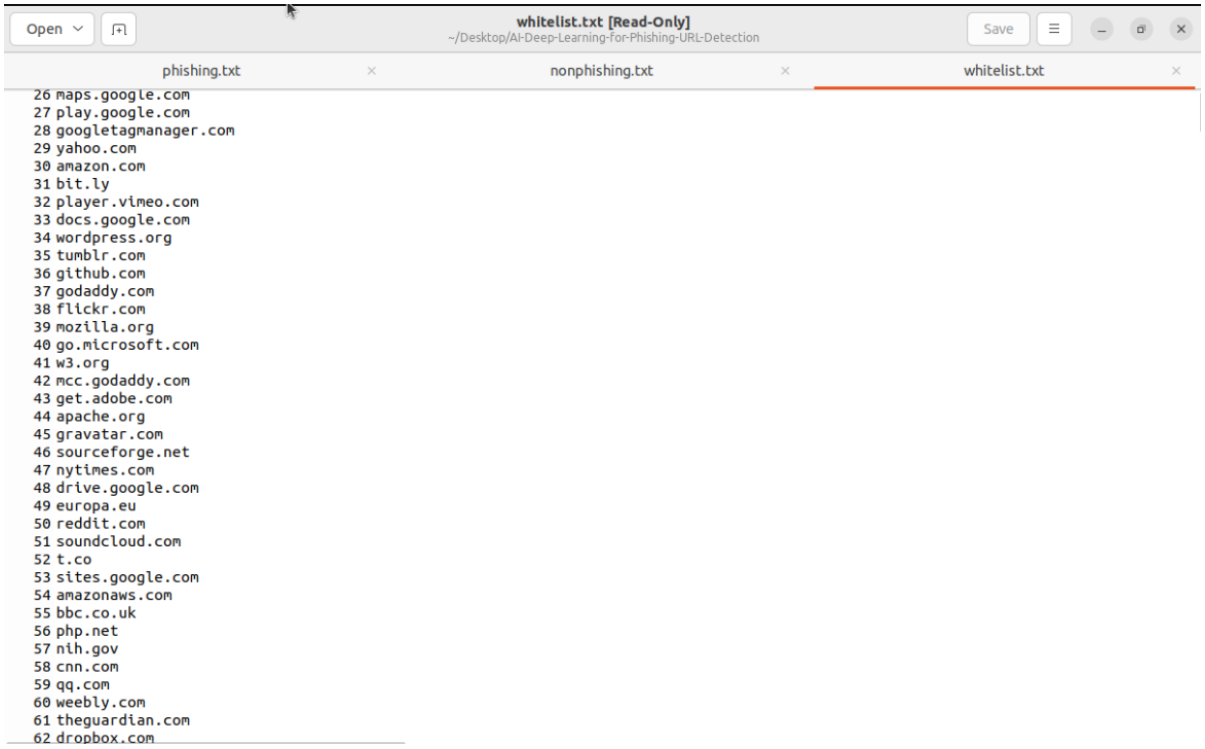


Рисунок 2.26 - Вміст файлу із легітимними посиланнями

І останній файл, який буде використовуватись в цьому методі найбільше це request.py. Принцип його роботи полягає у виконання POST запиту до представленого раніше API для процесингу посилання, вміст зображено на рисунку 2.27 [43].

```python
#!/usr/bin/env python
"""
This file will make a simple request to the Flask API for URL processing.
"""
import argparse
import requests

def main(url):

        # Define URL for Flask API endpoint.
        KERAS_REST_API_URL = "http://127.0.0.1:45000/predict"

        # Set the payload to JSON format.
        payload = {"url": url}

        # Submit the POST request.
        r = requests.post(KERAS_REST_API_URL, json=payload)
        response = r.json()

        # Ensure the request was sucessful.
        if response["success"]:
                # Loop over the predictions and display them.
                print(response['predictions'])

        # Otherwise, the request failed.
        else:
                print("Request failed")

if __name__ == "__main__":
    parser = argparse.ArgumentParser(description='Rock and roll.')
    parser.add_argument(
        '-u',
        dest='url',
        action='store',
        required=True,
        help="This is the url."
    )

    args = parser.parse_args()

    main(**vars(args))
```

Рисунок 2.27 - Вміст файлу requests.py

Заключним компонентом для реалізації методу буде сканування тексту повідомлення на наявність в ньому найбільш використовуваних фраз, якими послуговується штучний інтелект. В ролі механізму, який перевірятиме на збіг буде використовуватись YARA сканер та YARA правила, які найчастіше допомагають фахівцям ідентифікувати шкідливе програмне забезпечення [44]. На рисунку 2.28

відображено правило перевірки на збіги із найбільш часто використовуваними фразами, загалом використовується 106 таких фраз [45].

```
romanlab@roman-lab-ubuntu:~/Desktop$ cat CommonAIPhrases.yara
rule CommonAIPhrases {
    strings:
        $string1 = "Advancement in the realm"
        $string2 = "Aims to bridge"
        $string3 = "Aims to democratize"
        $string4 = "Aims to foster innovation and collaboration"
        $string5 = "Becomes increasingly evident"
        $string6 = "Behind the Veil"
        $string7 = "Breaking barriers"
        $string8 = "Breakthrough has the potential to revolutionize the way"
        $string9 = "Bringing us"
        $string10 = "Bringing us closer to a future"
        $string11 = "By combining the capabilities"
        $string12 = "By harnessing the power"
        $string13 = "Capturing the attention"
        $string14 = "Continue to advance"
        $string15 = "Continue to make significant strides"
        $string16 = "Continue to push the boundaries"
        $string17 = "Continues to progress rapidly"
        $string18 = "Crucial to be mindful"
        $string19 = "Crucially"
        $string20 = "Cutting-edge"
        $string21 = "Drive the next big"
        $string22 = "Encompasses a wide range of real-life scenarios"
        $string23 = "Enhancement further enhances"
        $string24 = "Ensures that even"
        $string25 = "Essential to understand the nuances"
        $string26 = "Excitement"
        $string27 = "Exciting opportunities"
        $string28 = "Exciting possibilities"
        $string29 = "Exciting times lie ahead as we unlock the potential of"
        $string30 = "Excitingly"
        $string31 = "Expanded its capabilities"
        $string32 = "Expect to witness transformative breakthroughs"
        $string33 = "Expect to witness transformative breakthroughs in their capabilities"
```

Рисунок 2.28 - Правило YARA для пошуку на співпадіння часто використовуваних штучним інтелектом слів

Принципова схема роботи методу зображена на рисунку 2.29. Зловмисник надсилає жертві фішингове повідомлення засобами електронного листування. Після цього поштовий клієнт отримує, та завдяки налаштованому користувачем фільтру зберігає його в локальну директорію. Далі відбувається власне виявлення. Насправді, всі компоненти описані раніше не мають прямої взаємодії між собою одразу після встановлення. Тому необхідно написати сценарій, який реалізовуватиме таку взаємодію.

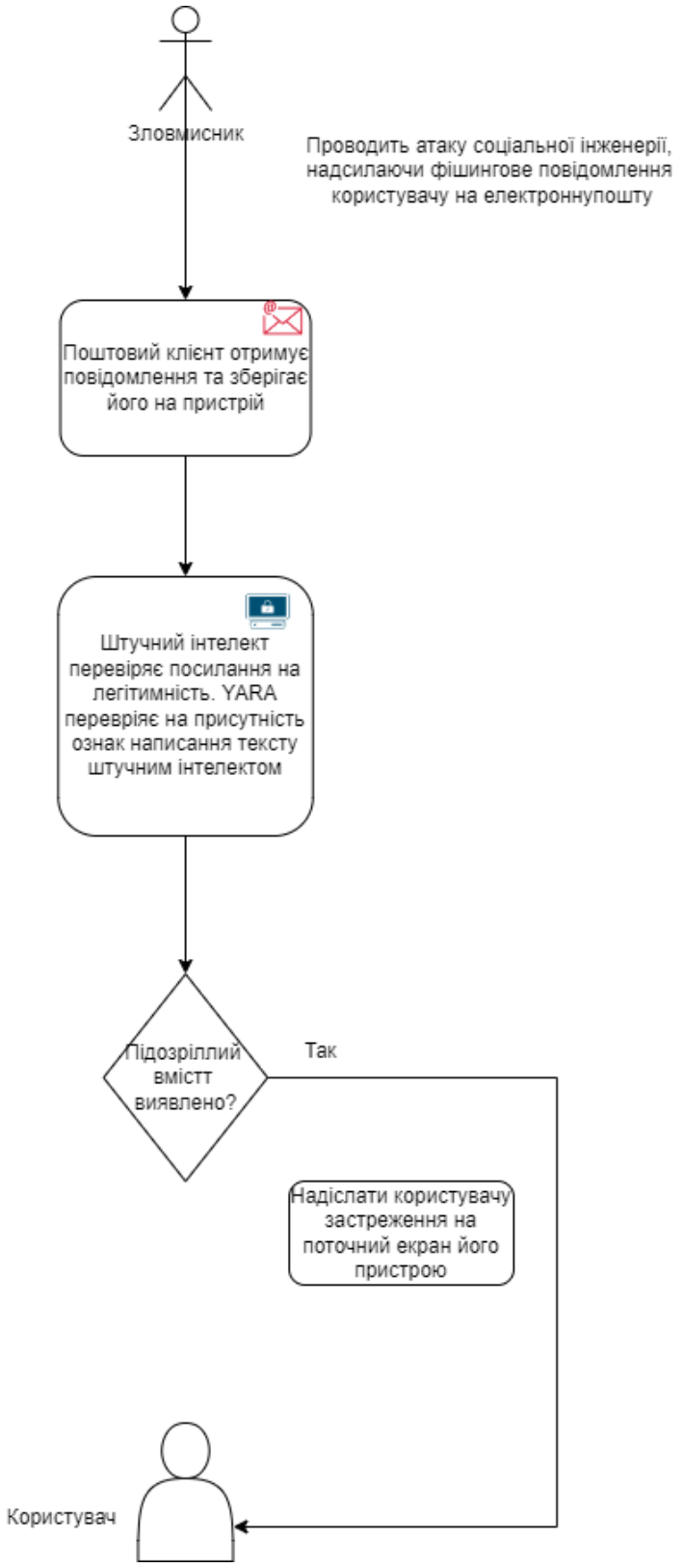


Рисунок 2.29 - Схема реалізації методу виявлення атак соціальної інженерії із використанням штучного інтелекту

Тож, маючи усі необхідні компоненти, можна перейти до написання скрипта автоматизації. Ідея полягає в тому, щоб кожен певний проміжок часу запускався

сценарій, який буде витягувати бажану інформацію із електронного повідомлення та передавати текст та посилання цього листа YARA сканеру та моделі штучного інтелекту відповідно і на основі їх рішень буде формуватися вердикт, чи має місце атака соціальної інженерії. Досягти інтревального запуску можна за допомогою утиліти Linux, яка називається Cron [46], проте спочатку напишемо скрипт. Отже, робота скрипта полягає в наступному: за допогомою парсера отримуються необхідні для аналізу дані із електронного листа(посилання та текст листа), далі вони передаються на аналіз моделі штучного інтелекту та YARA сканеру, які формують своє рішення і як заключення відбувається перевірка чи виконується достатня умова для того, щоб вивести користувачу повідомлення на екран про те, що в листі міститься певна підозріла активність. Скрипт зображено на рисунку 2.30.

Рисунок 2.30 - Скрипт для автоматизації процесу

Як можна помітити із тіла скрипта викликається ще один, він потрібний для коректної обробки посилань та передачі їх як аргументів для виклику запиту на рішення моделі штучного інтелекту [47]. Вміст зображено на рисунку 2.31.

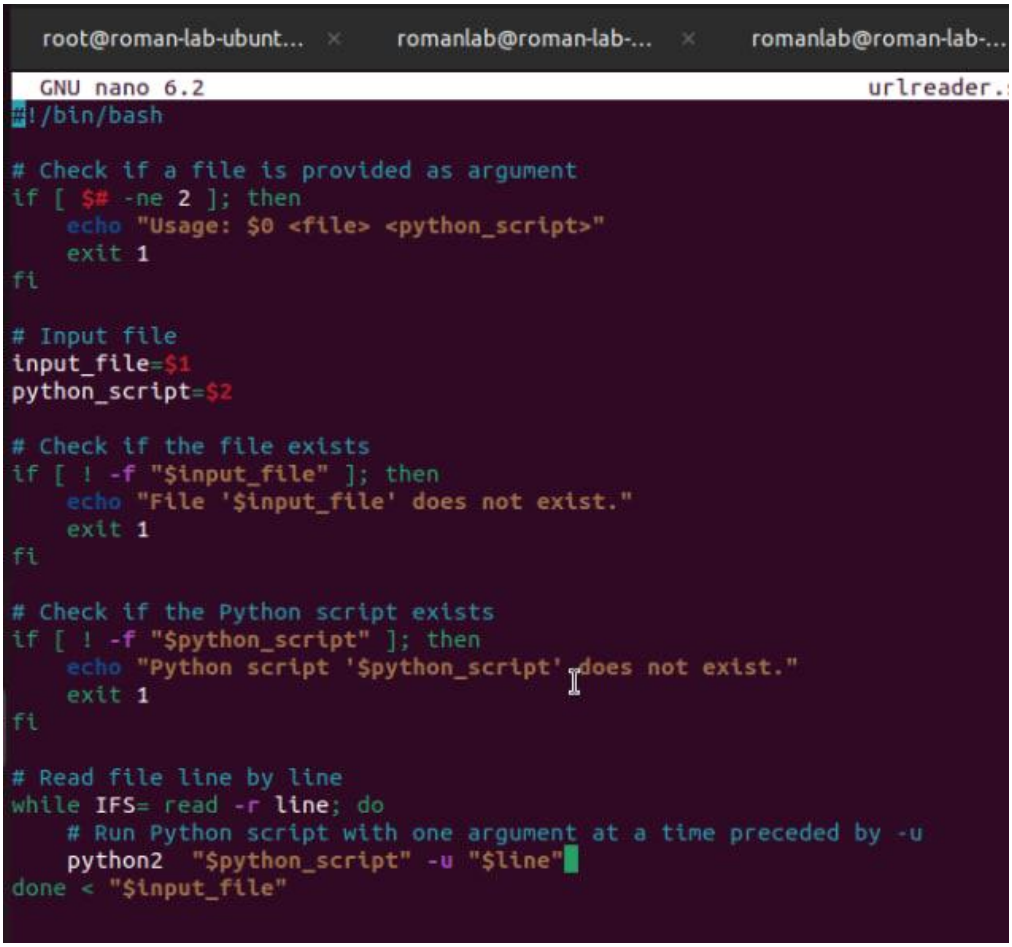


Рисунок 2.31 - Скрипт виклику запиту на аналіз посилання штучним інтелектом

Після встановлення всіх необхідних компонентів та написання скрипта автоматизації можна перейти до тестування методу. Для цього скористаємось написаним штучним інтелектом листом (рисунок 2.32) та додамо до нього фішингове посилання, все це разом засобами електронної пошти буде надіслано користувачу.

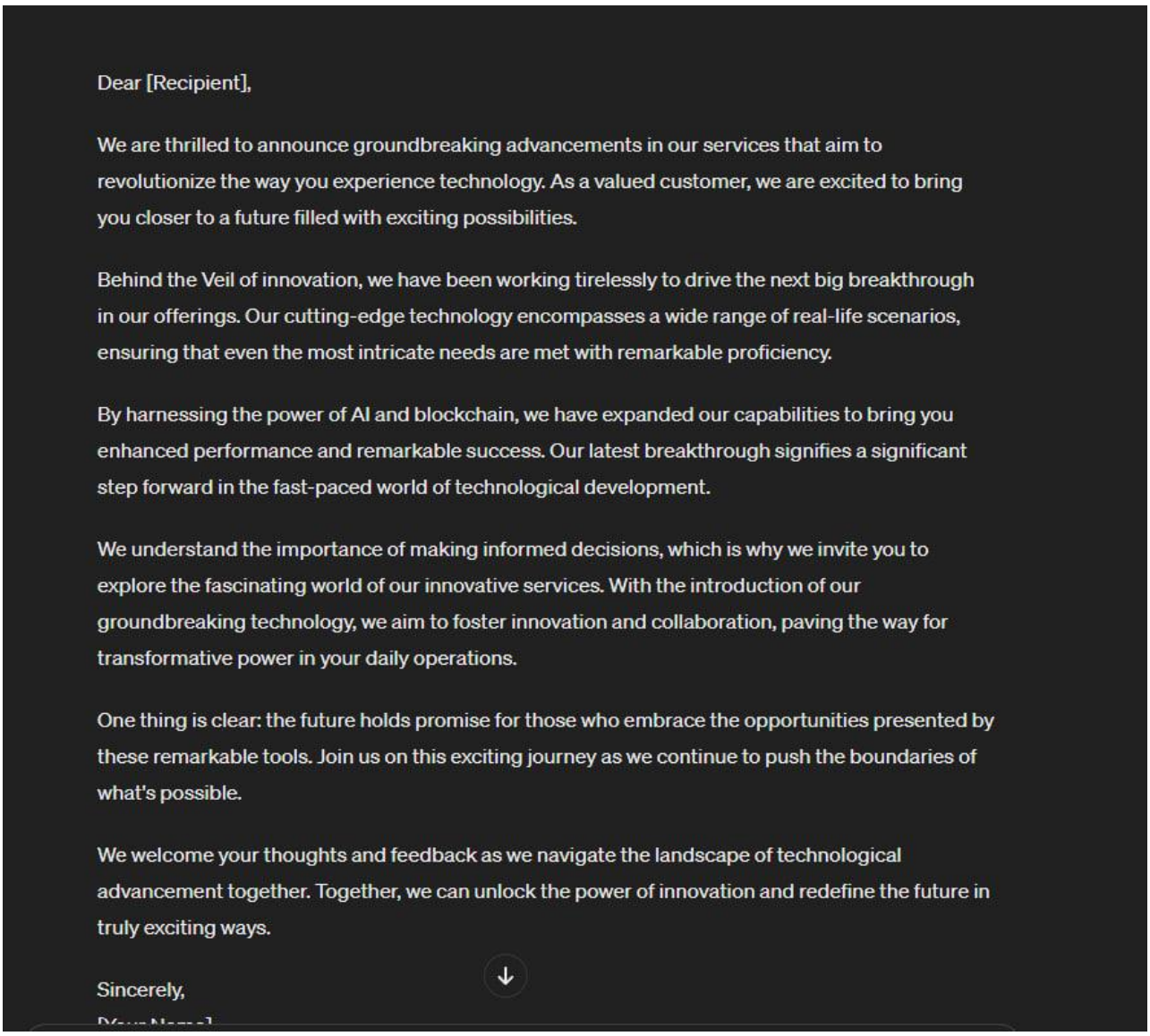


Рисунок 2.32 - Фішинговий лист написаний штучним інтелектом для перевірки коректності роботи

Для перевірки коректності роботи залишається тільки запустити скрипт, та дочекатись результатів його роботи. Скрипт відпрацював досить швидко, оскільки на поштовій скриньці не було більше повідомлень, у поточному вікні операційної системи користувача було виведено застереження [48]. Приклад спрацювання скрипта зображено на рисунку 2.33.

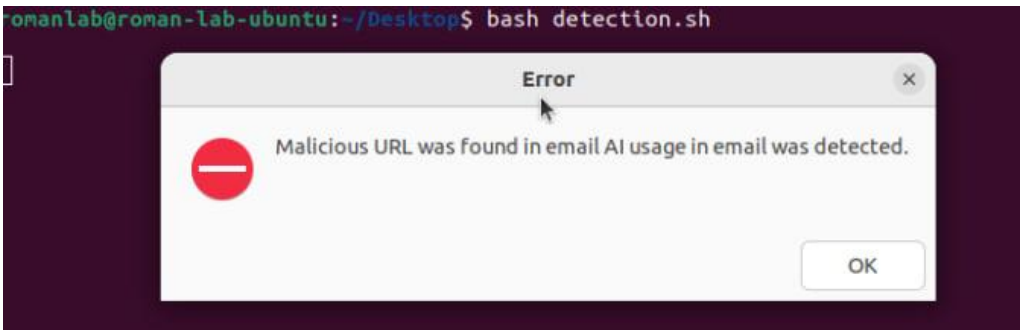


Рисунок 2.33 - Приклад коректного спрацювання методу

Після перевірки роботи скрипта, залишилось тільки налаштувати в системі його постійну роботу. Як і було згадано раніше, це буде виконано через Cron.

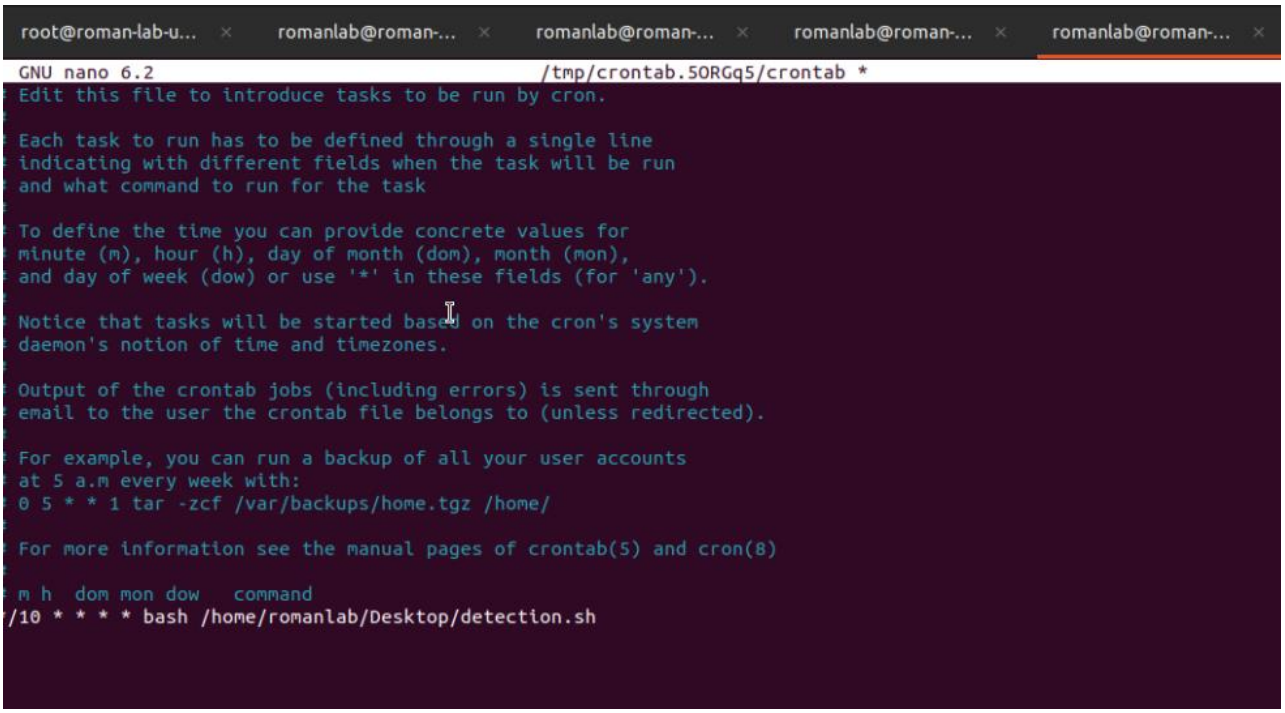


Рисунок 2.34 - Налаштування запуску скрипта кожні десять хвилин

Таким чином було реалізовано метод виявлення атак соціальної інженерії із використанням штучного інтелекту. Метод, звичайно, не є ідеальним оскільки через швидкість змін тактик зловмисників не можливо врахувати всі нюанси проведення атак. Також, варто зазначити, що й власне штучний інтелект активно розвивається, тому спеціалістам захисту інформації потрібно постійно тримати руку на пульсі та

бути готовими до нових загроз, розробляючи нові методи виявлення та протидії атакам [49].

Висновки до розділу 2

В даному розділі роботи було відтворено із метою демонстрації частину дій зловмисника при виконанні атаки соціальної інженерії із використанням штучного інтелекту. Як можна було помітити, ця атака не вимагає якогось особливого рівня технічних знань та ресурсів, її можна розпочати із звичайного персонального комп'ютера. Це ставить питання захисту особистих даних ще гостріше, оскільки якщо по-суті можливість атакувати є у всіх, то ж діаметрально протилежна можливість-захищатись теж має бути надана всім.

Тому в цьому розділі було запропоновано метод виявлення та протидії атакам соціальної інженерії із використанням штучного інтелекту для персонального комп'ютера. Перевагою цього методу є гнучкість та простота в його налаштуванні та імплементації на будь-яку робочу станцію з встановленою ОС Linux. Реалізований метод використовує невелику кількість ресурсів пристрою, на якому встановлений, що в свою чергу дозволяє встановлювати його на велику кількість сьогодні існуючих пристроїв. Також, при потребі використання цього методу, як єдиного вузла системи його може бути розміщено окремо та ізольовано, що виключати можливість зараження шкідливим програмним забезпечення інших пристроїв у мережі. Крім цього, всі компоненти для відтворення методу містяться в відкритому доступі, а також в мережі можна знайти багато інших правил, які теж можна імплементувати в метод, що дозволить йому виявляти й інші види атак . За допомогою виявлення та повідомлення користувачу інформації про наявність використання штучного інтелекту для написання повідомлення, мобілізуватиметься його здатність до критичного мислення та уважність при аналізі вмісту повідомлення, а адміністратор завжди матиме змогу допрацювати чи змінити певні характеристики на свій розсуд.

РОЗДІЛ 3
ОЦІНКА ЕФЕКТИВНОСТІ ЗАПРОПОНОВАНОГО МЕТОДУ ВИЯВЛЕННЯ ТА ПРОТИДІЇ АТАКАМ СОЦІАЛЬНОЇ ІНЖЕНЕРІЇ ІЗ ВИКОРИСТАННЯМ ШТУЧНОГО ІНТЕЛЕКТУ. РЕКОМЕНДАЦІЇ

3.1 Оцінка ефективності

Для оцінювання ефективності роботи даного методу було вибрано такі критерії: точність, швидкодія, масштабованість, складність його обходу, вартість імплементації, використання ресурсів. За допомогою оцінювання по цим критеріям, можна буде сказати не лише про ефективність у виявленні, але й про співвідношення затрат до результату.

1. Точність, для того, щоб переглянути на скільки точно метод зможе виявляти атаки, було підготовано 20 посилань, десять з яких були фішинговими, інші легітимним, та таку ж кількість листі, які були написані штучним інтелектом із використанням ключових слів. Списки посилань зображено на рисунках 3.1 та 3.2

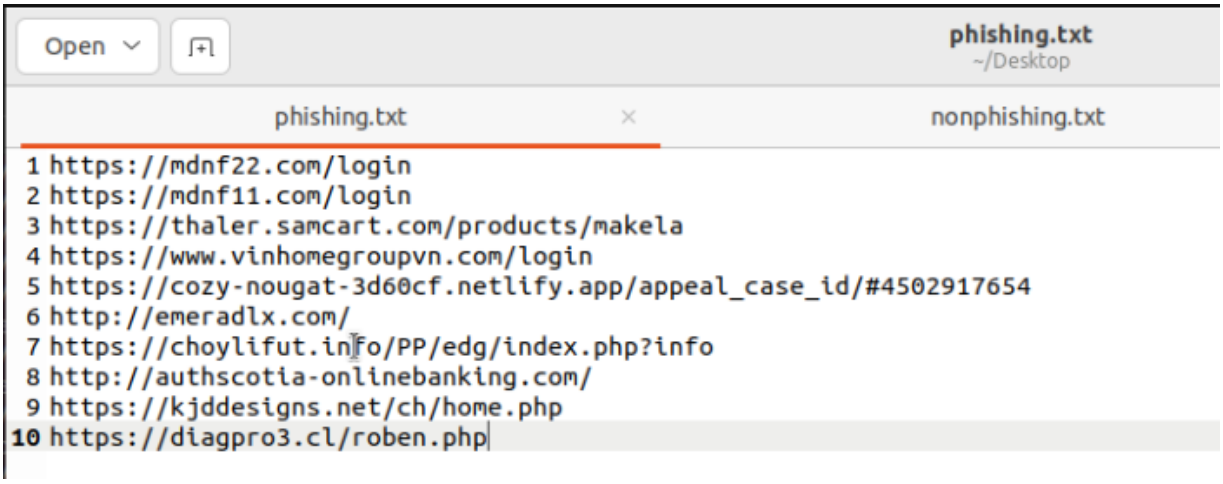


Рисунок 3.1 - Список фішингових посилань для перевірки ефективності методу

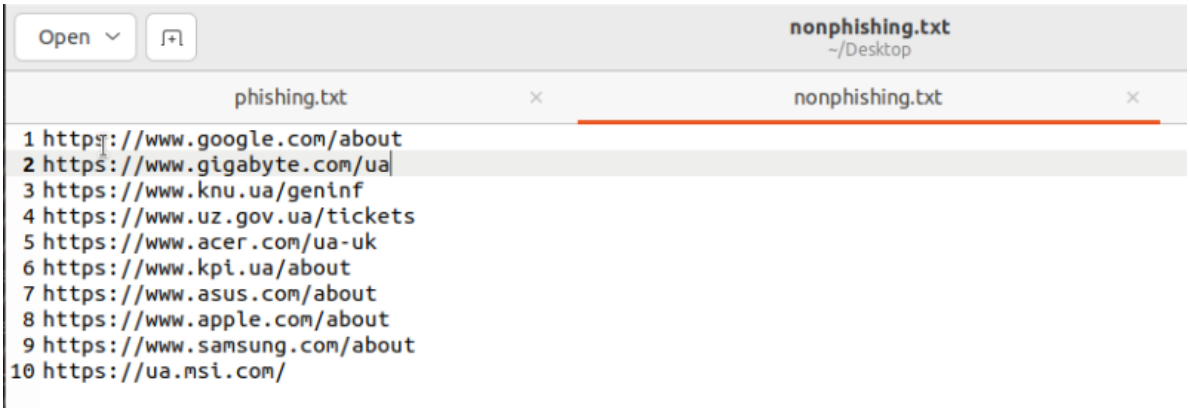


Рисунок 3.2 - Список легітимних посилань для перевірки ефективності методу

Після того, як було сформовано списки вони були передані на аналіз. Ті ж самі дії було виконано й стосовно аналізу текстів на вміст в них ключових слів, які використовує штучний інтелект. Результати роботи методу відображено на рисунках 3.3 та 3.4 та фішингових та легітимних посилань відповідно.

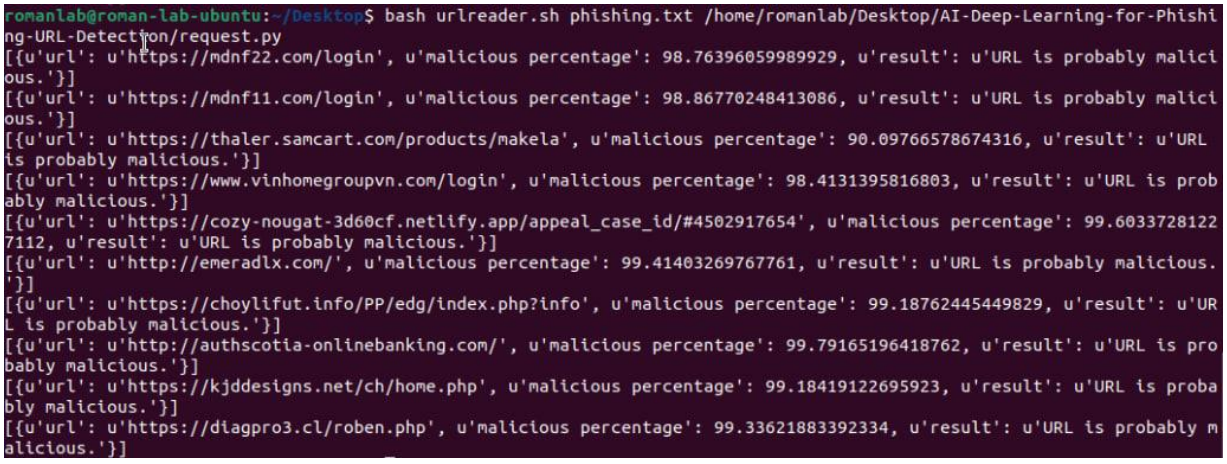


Рисунок 3.3 - Результат перевірки списку фішингових посилань

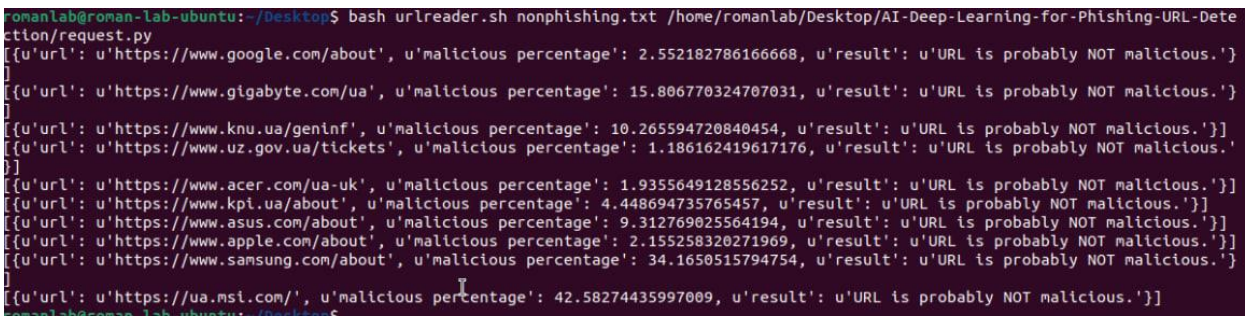


Рисунок 3.4 - Результат перевірки списку легітимних посилань

Як можна помітити із рисунків вище, модель досить чітко визначає шкідливі посилання, тобто рейтинг TP спрацювань достатньо високий. Проте, модель все ще дає досить високий відсоток шкідливості легітимним посиланням, що може призвести до високого рівня TN спрацювань. Тому цей компонент роботи механізму виявлення потрібно покращувати

YARA правила спрацьовують у 100% випадків коли в тексті листа містяться задані ключові слова, проте, точність цього компоненту не надто висока, оскільки він несе в собі тільки інформативний характер. Листи можуть бути написаним із використанням ШІ та не бути фішинговими, наприклад рекламні розсилки. Однак, на мою думку, користувачу потрібно володіти цією інформацією, оскільки повідомлення про використання цих засобів має підвищити увагу до тексту листа і у випадку, якщо його контент буде побудований так, щоб маніпулювати почуттям читача – він триматиме в голові думку, що лист написаний не людиною.

2. Другим критерієм є швидкість – використовуючи списки сформовані в попередньому підпункті було перевірено швидкодію методу. На оброблення 10 фішингових посилань було витрачено 50.32 секунд, а легітимних – 38.45. На обробку текстів написаних штучним інтелектом YARA сканер витратив 33.31 секунд та 26.5 відповідно. Тобто, можна сказати, що на аналіз одного листа буде витрачено менше однієї хвилини часу, що не матиме критичного впливу на досвід користувача та не перешкоджатиме іншій його роботі за пристроєм.

3. Для оцінки масштабованості методу було розміщено такі ж конфігурації на інших віртуальних машинах, за допомогою програмного забезпечення із оркестрування, наприклад, Ansible [50], це робиться досить швидко(залежить від вмінь та навичок системного адміністратора). Мінусом, який можна відзначити є необхідність налаштування копіювання листів на кожному пристрої окремо, проте ця задача технічно не складна, і будь-який користувач зможе зробити це по інструкції.

4. Складність обходу – очевидним є факт, що зловмисникам не складно буде обійти виявлення ключових слів, оскільки вони можуть їх видалити чи замінити. Проте, обійти метод виявлення фішингових посилань буде досить складно. Останнім часом одним із найпопулярніших таки методів є розміщення captcha перед переходом

на власне контент веб-сторінки [51]. Проте метод оцінювання посилання не базується тільки на його контенті, тому ймовірність коректного спрацювання залишатиметься високою

5. Вартість імплементації – всі компоненти, що використовуються в цьому методі були спеціально підібраними зі відкритим кодом та ліцензіями. Тобто, для того, щоб розгорнути даний метод виявлення користувачу не потрібно буде витрачати кошти на придбання програмного забезпечення.

6. Останнім пунктом даної оцінки є використання ресурсів пристрою. На рисунках 3.5 та 3.6 зображено використання пам’яті та ресурсів процесора під час перегляду відео та сканування 10 посилань на 10 текстів одночасно. Як видно з рисунків, запуск методу не надто сильно збільшив використання оперативної пам’яті та ресурсів процесора, приріст відбувся приблизно на 1.5 % відсотки використання оперативної пам’яті та близько 5% процесора, а враховуючи, що аналіз займатиме приблизно хвилину, це ніяк не відобразиться на досвіді користувача.

Рисунок 3.5 - Використання ресурсів комп'ютера під час відтворення відео

Рисунок 3.6 - Використання ресурсів комп'ютера під час роботи методу

Підсумовуючи описане вище, можна констатувати, що метод працює достатньо ефективно враховуючи його ціну та ресурс ємкість. Звичайно, метод потребує доопрацювання і покращень, але як базовий шар захисту користувачів від атак соціальної інженерії із використанням штучного інтелекту він є дієвим. На практиці пропонується засовувати його не як незалежний, а в парі із постійними тренуваннями користувачів, підвищувати їх рівень обізнаності та навчати їх виявляти такі атаки. Рекомендації стосовно цього буде розміщено у наступному підрозділі.

3.2. Рекомендації із виявлення атак соціальної інженерії для звичайних користувачів

Метою створення цих рекомендацій є підвищення обізнаності користувачів про ризики, які несуть атаки соціальної інженерії із використанням штучного інтелекту, а також навчання їх самостійної ідентифікації таких атак. Очевидним є той факт, що цей список не є чимось абсолютно повним та доконаним, оскільки атаки, їх темпи та швидкість постійно ростуть і змінюються - то й потрібно буде розширювати та додавати нові рекомендації.

Однак варто також зазначити, що даний список буде актуальним доволі тривалий період часу, оскільки зазначені кроки будуть ще довго використовуватись для захисту персональних даних користувачів, а їх безпека зараз це також зниження ймовірності бути атакованим у майбутньому.

1.Увімкніть фільтри надані вашим постачальником послуг електронної пошти, для автоматичного виявлення та поміщення в карантин підозрілих електронних листів. Якщо такий лист вирішено повернути з карантину потрібно уважно прочитати та проаналізувати всі посилання перш ніж переходити по них, а також уважно прочитати текст самого листа.

2.Установіть розширення для веб-переглядача, які використовують штучний інтелект для аналізу вмісту веб-сайту та попереджають вас про потенційні фішингові чи шахрайські веб-сайти, перш ніж ви з ними взаємодієте.

3.Пройдіть хоча б базовий курс навчання кібернетичній гігієні та безпеці в Інтернеті. Орієнтуйтесь на програми, що навчатимуть вас розпізнавати загрози та ефективно реагувати на них.

4.Використовуйте менеджери паролів, щоб генерувати надійні та унікальні паролі для кожного облікового запису. Також регулярно змінюйте паролі та не використовуйте їх повторно.

5.Інсталюйте антивірусне програмне забезпечення, яке використовує алгоритми штучного інтелекту для виявлення та блокування зловмисного програмного забезпечення, включно з шкідливим програмним забезпеченням, яке часто поширюється за допомогою атак із використанням соціальної інженерії.

6.Регулярно переглядайте та коригуйте налаштування конфіденційності своїх облікових записів у соціальних мережах, не викладайте приватну чи чутливу інформацію. Перевіряйте з ким спілкуєтесь, чи справді людина є тим, за кого себе видає. Також не варто ділитись інформацією про те, які саме засоби особистого захисту ви використовуєте. Будь-які технічні засоби, також можуть містити в собі вразливості і деякі з них можуть бути відомими зловмисникам, що призведе до більшої зацікавленості серед них вашою персоною.

7.Увімкніть 2FA, де це можливо, особливо для критично важливих облікових записів, таких як електронна пошта та банківські послуги, щоб додати додатковий рівень безпеки, який може запобігти атакам соціальної інженерії, які намагаються отримати неавторизований доступ. Постійно контролюйте доступ до кодів відновлення, ніколи не передавайте їх іншим особам.

8.Використовуйте служби сканування URL-адрес на основі штучного інтелекту або розширення браузера, які автоматично перевіряють посилання в повідомленнях або на веб-сайтах і попереджають вас, якщо вони здаються зловмисними або частиною схеми соціальної інженерії. Проте уважно перевіряйте, як саме ваші дані оброблятимуться такими службами, не передавайте можливість аналізувати всю інформацію, оскільки серед неї може виявитись ваша приватна.

9.Оновлюйте свої пристрої та програмне забезпечення до останніх доступних версій. Оновлення не тільки покращують досвід взаємодії, а я несуть в собі виправлення певних вразливостей, що зменшить поверхню атаки для зловмисника. Постате собі за звичку постійно перевіряти наяність оновлень для програмного забезпечення, яким ви активно користуєтесь, а те чим не користуєтесь – видаляйте.

10. Пам’ятайте, відповідальними за вашу безпеку є тільки ви. Ніколи не легковажте правилами та застереженнями від фахівців інформаційної безпеки, дотримуйтесь рекомендацій та завжди уважно аналізуйте предмет взаємодії перед виконанням будь-який дій, перевірюючи його на безпечність. Завжди консультуйтесь із фахівцями у сфері інформаційної безпеки, якщо відчуваєте не впевненість в тому, яку інформацію можна поширити, а на прохання про поширення якої варто відмовитись.

Дотримуючись цих простих, проте достатньо ефективних правил, користувачі зможуть суттєво зменшити ризик бути успішно атакованими зловмисниками, що вдаються до методів соціальної інженерії. Також рекомендується постійно навчатись і вдосконалювати навички критичного мислення, оскільки саме ці вміння спільно із переліченими вище рекомендаціями допоможуть користувачам визначатись підсвідомо.

Висновки до розділу 3

У цьому розділі було проведено оцінку ефективності запропонованого методу виявлення атак соціальної інженерії із використанням штучного інтелекту. Оцінка проводилась по шести характеристикам та показала, що метод є достатньо ефективним для впровадження його як вузла захисту персональних комп'ютерів на базі операційної системи Linux.

Основними перевагами методу є безкоштовність всіх його компонентів, гнучкість в налаштуванні, невеликий час спрацювання та обмежене використання ресурсів. Варто відзначити, що метод не використовує великої кількості ресурсів пристрою, що дає можливість використовувати його не великій кількості пристроїв(системні вимоги для встановлення будуть досить скромними). Також метод можна впроваджувати і на різних підприємствах, завдяки простоті його конфігурації системним адміністраторам буде достатньо легко вносити необхідні зміни при потребі. Значним плюсом також є те, що метод складається із різних компонентів і не базується на жодній із них повністю. Це означає, що завжди можна буде впровадити якесь технічно досконаліше рішення, а відкритий код кожної із складових методу дозволяє також при необхідності переписати його, щоб точно виконувати бажані завдання певного користувача чи підприємства. Жоден із основних компонентів не був розроблений строго під операційну систему Linux, це означає, що метод може бути встановлено і на інших операційних системах, до прикладу, Windows, після певної адаптації.

Серед мінусів можна відмітити, що не кожному користувачу під силу адміністрування даного методу. Для розуміння роботи його основних компонентів потрібно бути принаймні просунутим користувачем комп'ютера, а для певного покращення і адаптації бути зацікавленим у роботі сучасних технологій та розвитку технологічних рішень. Проте, із створенням достатньо детальної інструкції із детальним описом щодо встановлення та менеджменту компонентів ця проблема вибуває із розряду таких, що не можуть бути самостійно вирішеними користувачами.

Крім цього, було створено список рекомендацій для користувача, який покликаний підвищити рівень обізнаності та зменшити ризик бути успішно атакованим зловмисниками. Список містить в собі як практичні поради, так і фундаментальні, що покликані мобілізувати здатність користувача до критичного аналізу та бути готовим діяти в разі виявлення атаки соціальної інженерії із використанням штучного інтелекту.

ВИСНОВКИ

Метою даної роботи було підвищення ефективності виявлення та блокування кібератак які використовують штучний інтелект та соціальну інженерію .

Проведений аналіз інформаційних джерел показав, що об’єми використання таких атак дуже високі та продовжують рости. Це відбувається у зв’язку з тим, що штучний інтелект зараз перебуває у стані активного розвитку, викликає велику зацікавленість не тільки у науковців, вчених а й у зловмисників, чий рівень кваліфікації теж росте із часом. Тобто для вирішення цієї проблеми фахівці у сфері інформаційної безпеки повинні розробити чіткий план проведення тренінгів з підвищення обізнаності користувачів та запропонувати технічні методи захисту для них.

Для вирішення цього питання в роботі було запропоновано нову модель класифікації атак соціальної інженерії із використанням штучного інтелекту. Ця модель достатньо чіткою та повною для можливості детермінованого опису та власне класифікації як існуючих так і тих атак, що відбуватимуться в майбутньому. Це допоможе спеціалістам більш точно визначати слабкі місця наявних методів захисту, та дасть можливість для формування бази кроків, що повинні бути виконаними для підвищення стану захищеності потенційних жертв атак.

Також в роботі було проведено оцінку ефективності запропонованого методу виявлення атак соціальної інженерії із використанням штучного інтелекту. Оцінка проводилась по шести характеристикам та показала, що метод є достатньо ефективним для впровадження його як вузла захисту персональних комп’ютерів на базі операційної системи Linux. Було розроблено рекомендації, які повинні підняти рівень обізнаності користувачів і тим самим зробити їх менш привабливими як суб’єкта атаки для зловмисника.

На практиці було відтворено дії атакуючого та продемонстровано роботу запровадженого методу, який здатний виявляти кібератаки із використанням соціальної інженерії та штучного інтелекту. Перевагами такого методу є його

гнучкість в налаштуванні та побудові його виключно із компонентів програмний код, яких є відкритим. Запропонований метод може бути легко впровадженим користувачем як вузол захисту персонального пристрою, та не використовуватиме значної кількості ресурсів системи. Легкість у його адмініструванні дозволятиме оновлювати його згідно викликів часу, видозмінювати під потреби певного користувача та слугувати елементом, що дозволятиме зменшувати ймовірність успішного виконання атаки зловмисником.

Розглянута тема ще потребує подальшого дослідження. Штучний інтелект зараз активно розвивається, обчислюванні потужності дозволяють йому швидко навчатись, а продуктів із його впровадженням з'являється все більше. Тож перед фахівцями постійно стоятиме завдання ефективної протидії таким атакам. Люди завжди залишатимуться найбільш вразливими частинами інформаційно комунікаційних систем, проте ризик успішного виконання кібератак із використанням соціальної інженерії та штучного інтелекту повинен бути зменшеним.

У дипломній роботі розв'язано поставлене актуальне наукове завдання щодо розробки методу виявлення атак соціальної інженерії, що штучний інтелект. У процесі вирішення поставлених на початку роботи завдань були одержані такі наукові та практичні результати:

1. Здійснено детальний аналіз існуючих типів атак соціальної інженерії та їх найпоширеніших сценаріїв.
2. Розроблено класифікацію атак соціальної інженерії із використання штучного інтелекту.
3. Проведено симуляцію атаки соціальної інженерії із використанням штучного інтелекту.
4. Розроблено метод виявлення атак соціальної інженерії із використанням штучного інтелекту.
5. Розроблено список рекомендацій для користувача, щодо виявлення атак соціальної інженерії.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Carnegie Mellon University [Електронний ресурс]. – Режим доступу до ре сурсу: https://www.cmu.edu/iso/aware/dont-take-the-bait/social-engineering.html

2. TYPES OF SOCIAL ENGINEERING ATTACKS AND HOW TO PREVENT THEM [Електронний ресурс]. – Режим доступу до ресурсу: https://www.crowdstrike.com/cybersecurity-101/types-of-social-engineering-attacks/

3. 6 Types of Social Engineering Attacks [Електронний ресурс]. – Режим доступу до ресурсу: https://www.mitnicksecurity.com/blog/6-types-of-social-engineering-attacks.

4. AI phishing attacks: What you need to know to protect your [Електронний ресурс]. – Режим доступу до ресурсу: https://withpersona.com/blog/ai-phishing-attacks

5. Q3 2023 Phishing and Malware Report: Phishing and Malware Threats Increase 173% and 110% [Електронний ресурс]. – Режим доступу до ресурсу: https://www.vadesecure.com/en/blog/q3-2023-phishing-malware-report

6. PHISHING ACTIVITY TRENDS REPORTS [Електронний ресурс]. – Режим доступу до ресурсу: https://apwg.org/trendsreports/

7. Interaction with AI-written and human-generated phishing e-mails in European countries in March 2023 Електронний ресурс]. – Режим доступу до ресурсу: https://www.statista.com/statistics/1420881/ai-and-human-generated-phishing-e-mails-interaction-europe/

8. What is phishing | Attack techniques & scam examples [Електронний ресурс]. – Режим доступу до ресурсу: https://www.imperva.com/learn/application-security/phishing-attack-scam/

9. Guide to detecting and disrupting fake ads [Електронний ресурс]. – Режим доступу до ресурсу:https://www.netcraft.com/guide/guide-to-detecting-and-disrupting-fake-ads/

10. Kevin M. The Art of Deception: The Art Of Intrusion: The Real Stories Behind the Exploits of Hackers, Intruders & Deceivers / Mitnick Kevin. – 291 p. – (1st edition).

11. How to Recognize and Avoid Phishing Scams [Електронний ресурс]. – Режим доступу до ресурсу:https://consumer.ftc.gov/articles/how-recognize-and-avoid-phishing-scams

12. What Is Vishing and A Vishing Attack? [Електронний ресурс]. – Режим доступу до ресурсу: https://www.fortinet.com/resources/cyberglossary/vishing-attack

13. Fakeyou Guide [Електронний ресурс]. – Режим доступу до ресурсу: https://fakeyou.com/guide

14. OSINT Framework [Електронний ресурс]. – Режим доступу до ресурсу: https://osintframework.com/

15.Ghost in the Wires: The Kevin Mitnick Interview [Електронний ресурс]. – Режим доступу до ресурсу: https://www.zdnet.com/article/ghost-in-the-wires-the-kevin-mitnick-interview/

16. The Whys and The Hows of Email Spam Filters [Електронний ресурс]. – Режим доступу до ресурсу: https://mailtrap.io/blog/spam-filters/

17. AI in Phishing Detection and Response: Defending Against Social Engineering [Електронний ресурс]. – Режим доступу до ресурсу: https://megasisnetwork.medium.com/ai-in-phishing-detection-and-response-defending-against-social-engineering-b60b52f7fef4

18. Facing Cybercrimes Using AI: How to Prevent Phishing Attack? [Електронний ресурс]. – Режим доступу до ресурсу: https://shellmates.medium.com/facing-cybercrimes-using-ai-how-to-prevent-phishing-attacks-1dc64a047dc0

19. Behavioral Analysis and AI/ML for Threat Detection: Going Behind the Scenes on the Newest Detection Engine from Proofpoint [Електронний ресурс]. – Режим доступу до ресурсу: https://www.proofpoint.com/us/blog/email-and-cloud-threats/behavioral-analysis-and-aiml-threat-detection-going-behind-scenes

20. Analysis and prevention of AI-based phishing email attacks [Електронний ресурс]. – Режим доступу до ресурсу: https://arxiv.org/abs/2405.05435

21. Artificial intelligence and machine learning in phishing detection [Електронний ресурс]. – Режим доступу до ресурсу:

https://www.researchgate.net/publication/378233860_Artificial_intelligence_and_machine_learning_in_phishing_detection

22. Understanding SPF [Електронний ресурс]. – Режим доступу до ресурсу: https://dmarcian.com/what-is-spf/

23. Understanding DKIM [Електронний ресурс]. – Режим доступу до ресурсу: https://dmarcian.com/what-is-dkim/

24. Understanding DMARC [Електронний ресурс]. – Режим доступу до ресурсу: https://dmarcian.com/what-is-dmarc/

25. What Is Security Awareness Training? [Електронний ресурс]. – Режим доступу до ресурсу: https://www.mimecast.com/content/what-is-security-awareness-training/

26. Top Phishing Statistics for 2024: Latest Figures and Trends [Електронний ресурс]. – Режим доступу до ресурсу: https://www.stationx.net/phishing-statistics/

27. Paul J. Zak: The Moral Molecule: The Source of Love and Prosperity Hardcover / Paul J. Zak – 235 p.

28. Стаття 188-39 Кодексу України про адміністративні правопорушення Електронний ресурс]. – Режим доступу до ресурсу: https://zakon.rada.gov.ua/go/3454-17

29. Kevin M. The Art of Deception: Controlling the Human Element of Security / Mitnick Kevin. – 368 p. – (1st edition)

30. Different Types of Voice Call and Customer Care Services [Електронний ресурс]. – Режим доступу до ресурсу: https://medium.com/@sinthantechno/different-types-of-voice-call-and-customer-care-services-2f83d8cd1c7c

31. The Cyber Kill Chain Framework [Електронний ресурс]. – Режим доступу до ресурсу: https://www.lockheedmartin.com/en-us/capabilities/cyber/cyber-kill-chain.html

32. Кіндерись Р., Лаптєв О. Використання Штучного Інтелекту на різних етапах атак соціальної інженерії. Проблеми кібербезпеки інформаційно-телекомунікаційних систем: зб. тез та доп. міжнар. наук.-практ. конф., 26 квіт. 2024 р. Київ, 2024. С. 63-65.

33. Brian Jeffrey Fogg: Tiny Habits: Why Starting Small Makes Lasting Change Easy / Brian Jeffrey Fogg – 288 p.

34. Postfix Documentation [Електронний ресурс]. – Режим доступу до ресурсу: https://www.postfix.org/documentation.html

35. Gophish Documentation [Електронний ресурс]. – Режим доступу до ресурсу: https://getgophish.com/documentation/

36.Evilginx Introduction [Електронний ресурс]. – Режим доступу до ресурсу: https://help.evilginx.com/docs/intro

37. YAML: YAML Ain't Markup Language [Електронний ресурс]. – Режим доступу до ресурсу: https://yaml.org/

38. What is Thunderbird? [Електронний ресурс]. – Режим доступу до ресурсу: https://www.thunderbird.net/uk/about/

39. Email analyzer [Електронний ресурс]. – Режим доступу до ресурсу: https://github.com/keraattin/EmailAnalyzer

40. AI: Deep Learning for Phishing URL Detection [Електронний ресурс]. – Режим доступу до ресурсу: https://www.zpettry.com/ai+cybersecurity/ai-url-phishing-detection/

41. Zpettry/AI: Deep Learning for Phishing URL Detection [Електронний ресурс]. – Режим доступу до ресурсу: https://github.com/zpettry/AI-Deep-Learning-for-Phishing-URL-Detection

42. Phishtank [Електронний ресурс]. – Режим доступу до ресурсу: https://phishtank.org/

43. Keras Deveopers Guide [Електронний ресурс]. – Режим доступу до ресурсу: https://keras.io/guides/

44. USING YARA FOR MALWARE DETECTION [Електронний ресурс]. – Режим доступу до ресурсу: https://www.cisa.gov/sites/default/files/FactSheets/NCCIC%20ICS_FactSheet_YARA_S508C.pdf

45. 100 Most Recognizable Words in AI-Generated Text by AI Detection Tools [Електронний ресурс]. – Режим доступу до ресурсу: https://mpost.io/top-words-detectable-by-ai-detectors/

46. What Is a Cron Job: Understanding Cron Syntax and How to Configure Cron Jobs [Електронний ресурс]. – Режим доступу до ресурсу: https://www.hostinger.com/tutorials/cron-job

47. How to Use a Bash Script to Run Your Python Scripts [Електронний ресурс]. – Режим доступу до ресурсу: https://linuxconfig.org/how-to-use-a-bash-script-to-run-your-python-scripts

48. Довідка з програми Zenity [Електронний ресурс]. – Режим доступу до ресурсу: https://help.gnome.org/users/zenity/stable/index.html.uk

49. AI Market 2024: Trends and Future Growth Analysis | 2031 [Електронний ресурс]. – Режим доступу до ресурсу: https://www.linkedin.com/pulse/ai-market-2024-trends-future-growth-analysis-hnfgf

50. What is the Ansible IT automation platform? [Електронний ресурс]. – Режим доступу до ресурсу: https://www.techtarget.com/searchitoperations/definition/Ansible

51. How CAPTCHAs work | What does CAPTCHA mean? [Електронний ресурс]. – Режим доступу до ресурсу: https://www.cloudflare.com/learning/bots/how-captchas-work/