Київський національний торговельно-економічний університет
Кафедра інженерії програмного забезпечення та кібербезпеки

ВИПУСКНИЙ КВАЛІФІКАЦІЙНИЙ ПРОЕКТ

на тему:

«Розробка аналітичної системи розпізнавання фішингових сайтів на основі штучного інтелекту»

Студента 2м курсу, 6з групи, спеціальності 121 «Інженерія програмного забезпечення» спеціалізації «Інженерія програмного забезпечення» _____ підпис студента — Філяєва Максима Владиславовича

Науковий керівник кандидат технічних наук, професор кафедри інженерії програмного забезпечення та кібербезпеки _____ підпис керівника — Пашорін Валерій Іванович

Гарант освітньої програми доктор технічних наук, професор кафедри інженерії програмного забезпечення та кібербезпеки _____ підпис гаранта — Криворучко Олена Володимирівна

КИЇВ - 2020

Київський національний торговельно-економічний університет

Факультет інформаційних технологій
Кафедра інженерії програмного забезпечення та кібербезпеки
Освітній ступінь магістр
Спеціальність 121 «Інженерія програмного забезпечення»

Затверджую
Зав. кафедри інженерії програмного забезпечення та кібербезпеки
Криворучко О. В.
8 листопада 2019 р.

Завдання
на випускний кваліфікаційний проект студентові

Філяєву Максиму Владиславовичу
(прізвище, ім'я, по батькові)

1. Тема випускного кваліфікаційного проекту «Розробка аналітичної системи розпізнавання фішингових сайтів на основі штучного інтелекту»
Затверджена наказом ректора від "24" грудня 2019 р. № 4440

2. Строк здачі студентом закінченої проекту 01 грудня 2020 р.

3. Цільова установка та вихідні дані до проекту
Мета дослідження полягає у аналізі актуальних методів класифікації даних та відбору характеристик, притаманних фішинговим сайтам, побудові більш досконалої моделі класифікатора для розпізнавання фішингових сайтів, що буде прогнозувати з більш високою точністю чи є сайт фішинговим.
Об'єкт дослідження фішингові сайти
Предмет дослідження дослідження методи машинного навчання для розпізнавання фішингових сайтів

4. Консультанти проекту із зазначенням розділів, які консультують:

Розділ	Консультант (прізвище, ініціали)	Підпис, дата	
		Завдання видав	Завдання прийняв

5. Зміст випускного кваліфікаційного проекту (перелік питань за кожним розділом)

ВСТУП

РОЗДІЛ 1. АНАЛІЗ ПРЕДМЕТНОЇ ОБЛАСТІ ФІШИНГА, ЯК ЗАГРОЗИ ІНФОРМАЦІЙНІЙ БЕЗПЕЦІ

1.1. Дослідження поняття «фішингу»

1.2. Аналіз проблематики фішингу

1.3. Дослідження механізму фішингу

1.4. Класифікація типів фішингу

1.5. Дослідження циклу життя фішингової атаки

1.6. Аналіз мотивації та мети фішингових атак

1.7. Дослідження статистики фішингових атак

Висновки до розділу 1

РОЗДІЛ 2. ДОСЛІДЖЕННЯ ІСНУЮЧИХ МЕТОДІВ РОЗПІЗНАВАННЯ ФІШИНГОВИХ САЙТІВ

2.1. Підходи до розпізнавання фішингу

2.2. Програмний підхід до розпізнавання

2.3. Методи розпізнавання з машинним навчанням

2.4. Огляд існуючих рішень

2.5. Алгоритм розпізнавання фішингового сайту

2.6. Попередня обробка

2.7. Відбір характеристик

2.8. Оцінка классифікаторів

2.9. Запропонована модель классифікації

Висновки до розділу 2

РОЗДІЛ 3. АНАЛІЗ РЕЗУЛЬТАТІВ ДОСЛІДЖЕННЯ

3.1. Інструментарій та набір даних

3.2. Критерії оцінки

6. Календарний план виконання проекту

№ пор.	Назва етапів випускного кваліфікаційного проекту	Строк виконання етапів проекту	
		за планом	фактично
1	2	3	4
1.	*Вибір теми випускного кваліфікаційного проект*	*20.09.2019*	*20.09.2019*
2.	*Розробка та затвердження завдання на проект магістра*	*08.11.2019*	*08.11.2019*
3.	*Вступ та перелік літературних джерел*	*24.01.2020*	*24.01.2020*
4.	*Наукова стаття*	*01.09.2020*	*01.09.2020*
5.	*Технічне завдання*	*28.02.2020*	*28.02.2020*
6.	*Розділ 1. Аналіз предметної області фішинга, як загрози інформаційній безпеці*	*25.06.2020*	*25.06.2020*
7.	*Розділ 2 Дослідження існуючих методів розпізнавання фішингових сайтів*	*07.09.2020*	*07.09.2020*
8.	*Розділ 3. Аналіз результатів дослідження*	*19.10.2020*	*19.10.2020*
9.	*Програма та методика тестування*	*21.10.2020*	*21.10.2020*
10.	*Керівництво користувача*	*23.10.2020*	*23.10.2020*
11.	*Висновки та пропозиції*	*30.10.2020*	*30.10.2020*
12.	*Здача випускного кваліфікаційного проекту на кафедру (перша перевірка)*	*05.11.2020*	*05.11.2020*
13.	*Підготовка автореферату та презентації доповіді*	*05.11.2020*	*05.11.2020*
14.	*Попередній захист випускного кваліфікаційного проекту*	*25.11.2020-27.11.2020*	*25.11.2020-27.11.2020*
15.	*Зовнішнє рецензування випускного кваліфікаційного проекту*	*01.12.2020*	*01.12.2020*
16.	*Підготовка до публічного захисту випускного кваліфікаційного проекту*	*10.12.2020-11.12.2020*	

7. Дата видачі завдання «8» листопада 2019 р.

8. Науковий керівник випускного кваліфікаційного проекту Пашорін В. І.
(прізвище, ініціали, підпис)

9. Гарант освітньої програми Криворучко О.В.
(прізвище, ініціали, підпис)

10. Завдання прийняв до виконання студент Філяєв М. В.
(прізвище, ініціали, підпис)

11. Відгук керівника випускного кваліфікаційного проекту

_____
_____
_____
_____
_____
_____
_____
_____
_____
_____
_____
_____
_____
_____
_____
_____
_____
_____
_____
_____
_____
_____
_____
_____
_____
_____

Науковий керівник випускного кваліфікаційного проекту

_____

(підпис, дата)

Відмітка про попередній захист _____Пашорін В. І._____

(ПІБ, підпис, дата)

12. Висновок про випускний кваліфікаційний проект

Випускний кваліфікаційний проект студента __Філяєв М. В._____

(прізвище, ініціали)

може бути допущена до захисту екзаменаційній комісії.

Гарант освітньої програми _____Криворучко О. В._____

(прізвище, ініціали, підпис)

Завідувач кафедри _____Криворучко О. В._____

(підпис, прізвище, ініціали)

«_____»_____20 ___ р.

АНОТАЦІЯ

У даній роботі розглянуто проблематику загрози фішингу та розроблено класифікатор для протидії даній загрози.

Запропонований у роботі класифікатор бізується на аналізі роботи п’яти алгоритмів класифікації даних: Neural Network, Naive Bayes, Logistic Regression, Decision Tree, k-nearest neighbours. Для експерименту з відбором ознак були використані дані алгоритмів машиного навчання: Wrapper subset evaluation, Consistency subset evaluation і Correlation-based feature subset evaluation. П’ять вище зазначених методів були треновані з різним набором даних для виявлення плюсів та мінусів процесу відбору функцій. Дослідження виявило деякий набір ознак з найбільш ефективною точністю класифікації і дало змогу відібрати три алгоритми із запропонованих п’яти для розробки нової схеми класифікації.

Впровадження даного класифікатора дозволить з більшою точністю виявляти фішингові сайти за рядом притаманних їм ознакам, що в свою чергу може використатися як один із засобів протидії фішингу.

Ключові слова: фішинг, фішинговий сайт, розпізнавання фішингових сайтів, відбір характеристик, класифікація, машинне навчання..

ABSTRACT

This paper considers the issue of phishing threat and develops a classifier to counter this threat.

The proposed classifier is based on the analysis of five data classification algorithms: Neural Network, Naive Bayes, Logistic Regression, Decision Tree, k-nearest neighbors. For the experiment with the selection of features, the data of machine learning algorithms were used: Wrapper subset evaluation, Consistency subset evaluation and Correlation-based feature subset evaluation. The five methods

mentioned above were trained with different data sets to identify the pros and cons of the feature selection process. The study identified some set of features with the most effective classification accuracy and allowed us to select three algorithms from the proposed five to develop a new classification scheme.

Keywords: phishing, phishing site, recognition of phishing sites, selection of characteristics, classification, machine learning.

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ, СИМВОЛІВ, ОДИНИЦЬ, СКОРОЧЕНЬ І ТЕРМІНІВ

- J48 Decision tree J48, дерево прийняття рішень J48
- APWG Anti-phishing working group, сервіс боротьби з фішингом
- kNN k-nearest neighbors, К найближчих сусідів
- NB Naive Bayes, наївний Байєсівський метод
- NN Neural network, нейронна мережа

ЗМІСТ

					КНТЕУ 121 06з-16.МР		
Зм.	*Аркуш*	*№ докум.*	*Підпис*	*Дата*			
Зав. каф.		Криворучко О.В.		*19.10.20*	*Розробка аналітичної системи розпізнавання фішингових сайтів на основі штучного інтелекту*	*Стадія*	*Аркуш*	*Аркушів*
Керівник		Пашорін В.І.		*19.10.20*		*З*	*3*	*55*
Гарант		Криворучко О.В.		*19.10.20*		*Факультет інформаційних технологій 2м курс, 6з група*		
Розробив		Філяєв М.В.		*19.10.20*	*Зміст*			

ВСТУП

Проблема атак соціальної інженерії в Інтернеті з кожним роком стає все більш розповсюдженою. Якщо говорити про організації, що мають конфіденційну інформацію таку як дані клієнтів, дані про співробітників, власні розробки, документацію і т.д., то з великою впевненістю можна сказати про те, що маючи захищену комп`ютерну систему всередині організації, не всі турбуються про витік інформації через своїх співробітників, а ті або через незнання або через неуважність стають жертвами соціальної інженерії.

У цій роботі буде розглянуто певний вид соціальної інженерії, який називається фішинг. Фішинг – це атака на основі соціальної інженерії, яка здійснюється через недоліки в кібербезпеки для обману користувачів з метою крадіжки їх логінів, паролів і грошових коштів. Техніки фішингу досить численні і складні, серед яких такі, які передбачають перехід за посиланням на зловмисний сайт.

Оскільки фішинг є досить розповсюдженим явищем, то від нього потерпають як і підприємства так і звичайні люди, методи захисту від фішингу не є досконалими і завжди актуальним є пошук нових рішень в цій сфері. Мій пошуквідбувається у сфері машинного аналізу даних, де все ще не існує досконалого, що означає точного і швидкого, алгоритму розпізнавання фішингових сайтів.

РОЗДІЛ 1.

АНАЛІЗ ПРЕДМЕТНОЇ ОБЛАСТІ ФІШИНГА, ЯК ЗАГРОЗИ ІНФОРМАЦІЙНІЙ БЕЗПЕЦІ

1.1. Дослідження поняття «фішингу»

З розвитком мережі Інтернет закономірно з'явилася сфера злочинності, що використовує особливості побудови Інтернет та взаємодії людей у ньому. У той час як деякі злочинні атаки використовують виключно програмний підхід, інші комбінують взаємодію з жертвою злочину та спеціальне програмне забезпечення для досягнення своїх цілей. У кінці минулого століття вперше почали писати про фішинг. Фішингом є атака на основі соціальної інженерії, яка здійснюється через слабкості в кібербезпеці для обману користувачів з метою крадіжки їх паролів, логінів та фінансових даних. Техніки фішингу постійно змінюються та модифікуються, підлаштовуючись під сучасні засоби комунікації. Дослідження показали, що основна мета більшості фішингових атак – змусити користувача перейти за посиланням на фішинговий сайт. Через значну кількість підходів до побудування фішингових атак неможливо виділити одне універсальне визначення, яке наддасть вичерпний опис цього поняття. Аналіз виявив, що вже сам факт існування різних визначень фішингу показує актуальність питання захисту від цієї загрози.

Над питанням протидії фішингу працює ряд організацій. За визначенням Anti-Phishing Working Group (APWG) фішингом є злочин на основі використання технічних засобів, соціальної інженерії чи комбінування цих методів для крадіжки даних користувачів, що включає в себе облікові засоби соціальних мереж, фінансових рахунків та іншу конфіденційну

інформацію. У схемах соціальної інженерії використовуються підробні електронні листи, які імітують повідомлення від відомих підприємств і установ, щоб змусити споживачів повірити у реальність повідомлень та перейти на підробні веб-сторінки. За допомогою останніх зловмисники отримують фінансові данні, імена та паролі користувачів.

Схеми технічних засобів впроваджують кримінальне програмне забезпечення для крадіжки облікових даних на пряму на ПК жертви. Використовуються системи для перехоплення імен користувачів і паролів в облікових записах користувачів, а також пошкодження локальної навігаційної інфраструктури, з метою дезорієнтації користувачів та перенаправлення їх на підробні веб-сторінки (або на справжні веб-сторінки через проксі-сервери, які контролюються фішером). [1]

Інший ресурс, PhishTank Company (далі PHT), дає більш загальний опис фішингу. Фішинг - спроба вкрасти особисту інформацію людини, використовуючи для цього електронну пошту. Фішингові атаки часто посилаються на сайти, які є підробкою реальних сайтів організацій, банків. Завдяки цьому у користувача складається враження, що він потрапив на реальний сайт цієї організації. [2]

Такий варіант визначення покриває велику частину сценаріїв фішингових атак, у той же час обмежуючи поняття фішингу лише крадіжкою особистих даних. Якщо зловмиснику вдасться встановити на пристрій жертви шкідливу програму, яка зможе реалізувати атаку Man in the Browser (MITB), це призведе до переказу грошей на банківський рахунок атакуючого, за умови що жертва авторизується в своєму банківському обліковому записі. Цей сценарій відбувається без отримання персональної інформації, що доводить неповноту попереднього визначення.

Якщо виділити спільне з приведених вище визначень то можна приписати таку поведінку типової фішингової атаки: вона може відбуватися з будь-якого електронного каналу зв'язку, метою атакуючого є виконання жертвою якихось дій, та атакуючий потенційно може отримати від цього особисту вигоду.

Тому якщо спробувати задати загальне формулювання дослідження поняття «фішинг», воно буде звучати таким чином:

Фішинг - це комп'ютерна атака, метою якої є переконання користувача виконати певні дії через застосування повідомлень та методів соціальної інженерії, що може призвести до задоволення інтересів атакуючого.

1.2. Аналіз проблематики фішингу

Дослідження задачі виявлення фішингових сайтів потребує багато часу і включає в себе багато різних факторів і критеріїв. На сьогоднішній день дуже легко отримати доступ до комерційних та фінансових послуг через розповсюдження мобільних пристроїв. Мережа Інтернет забезпечує функціональну платформу для фінансових транзакцій користувачів, водночас відкриваючи шляхи для реалізації пов’язаних з цим загроз. Викрадення інформації або фішинг, як вже було зазначено вище, - поширена проблема захисту інформації, що виконується через надсилання спаму або підроблених електронних листів. Цей тип атак передбачає отримання користувачем обманного листа, у якому міститься посилання на зловмисний веб- сайт, який в свою чергу призначений для збору інформації та персональних даних користувача. Фішингові атаки в основному базуються на розсиланні спаму і призводять до значних фінансових втрат – тільки у Центр скарг на злочини в Інтернеті ФБР повідомив, що за один рік люди втрачають приблизно 57 мільйонів доларів через фішингові схеми. [2] Аналіз розсилки спаму та

фішингових атак показує наявність шаблонів проведення атак. Однак незважаючи на простоту концепції фішингової атаки, цей шаблон складно розпізнати з першого погляду. Виявлення шаблонів фішингових атак є ефективним методом при розробці програмного забезпечення для запобігання атакам цього типу. Існує безліч методів збору даних для вилучення потрібної інформації з різних невпорядкованих даних. Використання методу збору даних для аналізу, що були отримані від попередніх атак, дозволяє розробляти ПЗ для ідентифікування шаблонів атак і запобігти майбутнім вторгненням.

Кількість активних користувачів інтернету зросла неспинно продовжує зростати. Одна з причин такого росту – доступність мережі, яка забезпечується значною кількістю мобільних пристроїв. Смартфони, планшети, ноутбуки і навіть годинники – всі ці пристрої забезпечують зручний доступ до мережі, тому використання інтернет протоколів стало невід'ємною частиною життя мільйонів людей по всьому світу. Водночас з користувачами, які є потенційним джерелом доходу, в інтернет прийшли фінансові та бізнес-організації, банки, онлайн-магазини т інші види бізнесу, що використовують інтернет платформи для надання своїх послуг. Водночас з перевагами такого сервісу з’явилися і недоліки у вигляді загрози від користувачів, що можуть наражати на небезпеку фінансові операції інших користувачів і обманювати їх через фішингові атаки для крадіжки цінної інформації. Інтернет є не лише простором для розваг та фінансових операцій, а й спілкування та самовираження. Жага людей до визнання та фінансової вигоди іноді ставить під загрозу інтернет безпеку інших користувачів. Крадіжка інформації або фішинг відноситься до сфери інформаційної безпеки.

На сьогоднішній день, використання електронної пошти стало невід'ємною частиною життя. Електронне листування включає в себе ділову, фінансову або особисту переписку. Оскільки електронна пошта потенційно

містить багато конфіденційної інформації, вона стає основною ціллю у крадіжці призначених для користувача даних.

Один з типів електронних листів з якими користувачі стикаються майже щодня – спам-листи від різних сервісів. Зазвичай це реклама, однак іноді на обліковий запис користувача надходить спам від фішерів або хакерів. Спам може розглядатися як засіб обману інтернет-користувачів в якому зловмисник представляється довіреною особою від організації (рисунок 1.1) або обіцяє виграш у лотереї або будь-який інший трюк, що змушує користувача перейти на підроблений сайт і ввести особисту інформацію у відповідне поле.

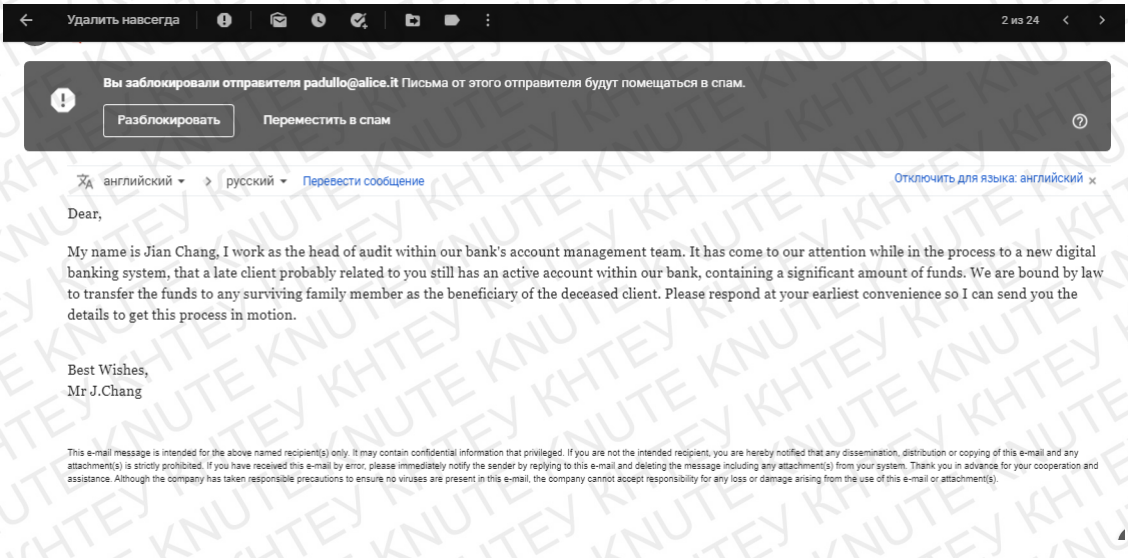

Рис. 1.1. Приклад фішингового листа

Мені надійшло повідомлення без вказання імені, у якому згадується про померлого родича та перерахування значної суми коштів з його банківського рахунку на мій. Для цього я повинен зв’язатися з відправником для отримання деталей. Поштова система вже позначила такий лист як «спам» та помістила його у відповідну папку.

Аналіз змісту подібних листів або інтернет спаму показує що зловмисники використовують поведінковий шаблон в написанні електронних листів. Оскілки лист може перекладатися на декілька мов, у ньому зазвичай присутні граматичними помилки, сліди машинного перекладу та часто зустрічаються слова-тригери, такі як «допомога», «гроші», «віза», тощо.

Дані та інформація представлені в спамі і листах відправлених фішерами дають розуміння, яке потрібне для розпізнавання таких атак, а аналіз їх структури допомагає у розвитку утиліт безпеки. Інтелектуальний аналіз даних – один з важливих методів для отримання прихованих шаблонів і корисної інформації з величезної кількості невпорядкованих даних.

Інтелектуальний аналіз даних включає різні техніки для виділення корисної інформації, такі як дерево рішень, кластеризація, класифікація, машинне навчання та штучна нейронна мережа.

1.3. Дослідження механізму фішингу

Дослідження показало, що першим етапом механізму фішингу є етап підготовки - атакуючий збирає дані про організацію яку збирається підробляти. Детальна інформація на сайті організації і вдала імітація дозволяють створити, викликаючий довіру, підробний сайт. Загальний механізм фішингу продемонстровано на рисунку 1.2.

Рис. 1.2. Механізм фішингу

Наступний етап атаки полягає у створенні підробного електронного листа. До листа зловмисник прикріплює посилання на фішинговий сайт і розсилає його базі користувачів. Якщо йде цілеспрямований фішинг, то

розсилка іде кільком людям. Зазвичай такі листи відрізняться від масових розсилок своєю якістю та підходом до попереднього опрацювання інформації про користувачів. Поширення шкідливого посилання відбувається не тільки за допомогою електронної пошти – з цією метою використовують форуми, соціальні мережі, коментарі на стрімінгових сервісах, тощо.

Коли користувач відкриває підроблений сайт, він бачить на ньому форму для введення даних, теж підроблену, найчастіше його просять ввести свої логін і пароль. Якщо користувач їх введе і підтвердить відправку форми – зловмисник отримає доступ до даних жертви. Як результат, фішер використовує дані користувача в злочинних цілях. Роблячи завершення дослідження механізму фішингу, злочинними цілями фішера може бути крадіжка особистості (для впровадження фішингу від імені реального облікового запису) чи використання кредитної картки для покупок, шантаж користувача, тощо.

1.4. Класифікація типів фішингу

Як вже було зазначено, класифікувати всі типи фішингових атак дуже складно, однак класифікація [5] виділяє типи (приведені деякі з описаних):

- *Whaling*. Атака націлена на керівника підприємства, «великих риб». Інформація від керівника буде завжди більш цінною ніж від звичайного співробітника. Ціллю зловмисника є люди на керівних посадах, проводиться комплексна підготовка до атаки.
- *Spear phishing(направлений фішинг)*. Цей вид фішингу має конкретного одержувача чи аудиторію одержувачів. Для виконання такої атаки фішер збирає загальну інформацію про аудиторію атаки. Цільові атаки мають успіх, оскільки підготовлені (наприклад, розсилка спаму від імені ВОЗ під час пандемії).

- *Business email compromise(BEC).* Атака спрямована на співробітників бухгалтерії і відділу фінансів компаній. Зловмисник видає себе за ключових людей у компанії і віддає накази про перекази коштів на несанкціоновані рахунки. Як правило, зловмисники компрометують обліковий запис електронної пошти керівника або фінансового директора, використовуючи різні методи.

1.5 Дослідження циклу життя фішингової атаки

Дослідження життєвого циклу фішингової атаки широко використовується для розробки технік протидії фішингу. На рисунку 1.3 зображена блок-схема життєвого циклу фішингової атаки при наявності/відсутності методу протидії.

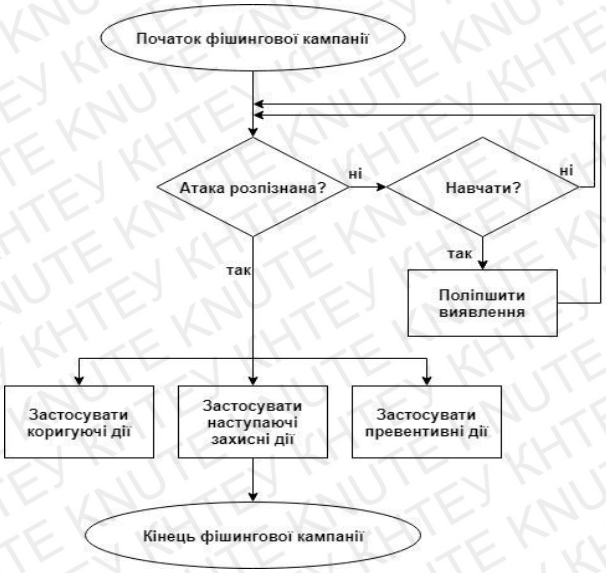

Рис. 1.3. Алгоритм протидії фішинговій атаці

Коли фішингова атака тільки починається першим кроком до протидії є її виявлення. У виявленні фішингових атак виділено 2 основні підходи: клієнтське програмне забезпечення і програми інформування користувачів (приклад інформування можна побачити на рисунку 1.1).

Здатність розпізнавати атаки вдосконалюється з часом шляхом навчання, це навчання може бути проведено або за допомогою користувачів, або алгоритмами машинного навчання. Після виявлення атаки, є кілька шляхів протидії:

- Запобіжні заходи – цей підхід полягає у виконанні користувачем дій, які спрямовані на запобігання атак в майбутньому.
- Наступальний захист - в цьому випадку жертва фішингової атаки атакуватиме компанію, що запустила початкову атаку. Цей підхід може допомогти користувачам, які вже відправили свої персональні дані зловмисникові, однак він скоріш про те, як мінімізувати збитки від фішингу.
- Корекція - в даному випадку користувач повідомляє про відповідний хостинг і видаляє сліди і файли, що залишилися після фішинг-атаки.

Всі перераховані вище методи працюють тільки якщо атака виявлена. Життєвий цикл фішингових атак має 3 етапи: ранній етап, середній етап і після фішінговий етап. [4] На ранньому етапі фішер готується до атаки і формує шаблон, який розсилає жертвам. На середньому етапі жертви отримують повідомлення зловмисника і розкривають персональні дані та цінну інформацію. Результатом буде викрадення інформації на третій стадії.

1.6. Аналіз мотивації та мети фішингових атак

Існують різні мотиви, які спонукають зловмисників до реалізації фішингу. Серед цих мотивів можна виділити 2 основні:

1. Фінансовий. Найпоширеніший мотив для фішингової атаки - в цьому випадку, атакуючий намагається вкрасти кошти з таких установ як банки або рахунки платіжних систем користувачів.
2. Психологічний. Жага до визнання, помсти чи задоволення інших

3. нематеріальних потреб, що принесуть нематеріальному вигоду штовхають зловмисника до фішингу.

1.7. Дослідження статистики фішингових атак

Досліджуючи звіт сервісу APWG (за другий квартал 2020 року) [5] основними цілями для фішингу стали електронна пошта (Webmail), та фінансова індустрія (Financial Institution). Візуалізацію можна побачити на рисунку 1.4:

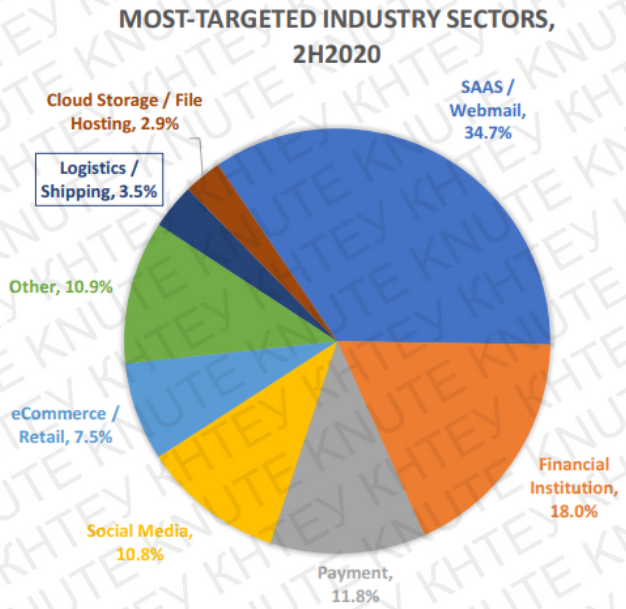

Рис. 1.4. Найбільш цільові галузі промисловості у 2 кварталі 2020 року

Покращується і якість фішингових сайтів. Зловмисники можуть легко створювати безкоштовні сертифікати DV (domain validated), і більше веб-сайтів використовують SSL в цілому. Все більше сайтів використовують SSL, тому що браузер попереджає користувачів, коли SSL не використовується – це є першою ознакою ненадійного сайту. На рисунку 1.5 можна побачити відсоток фішингових сайтів, розміщених на HTTPS:

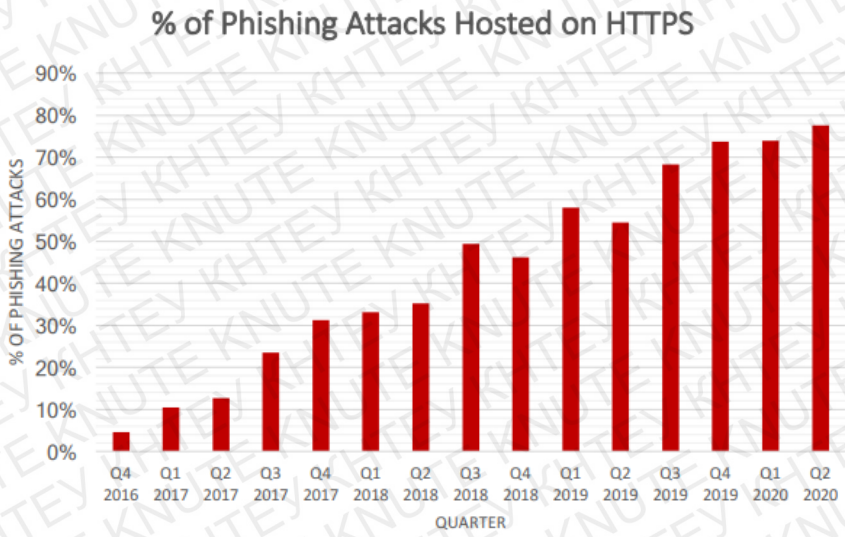

Рис. 1.5. Відсоток фішингових атак на HTTPS

1.8. Висновки до розділу 1

Фішинг є дуже розвинутою областю соціальної інженерії. Різноманітність різновидів фішингу та частота виявлення цього явища у повсякденному житті роблять з цього явища актуальну загрозу. Фішинг застосовують у таких галузях як електронна пошта, прикладне програмне забезпечення, платіжні сервіси, фінансові установи тощо. Незважаючи на кількість фішингових атак, більшість підходів є шаблонними. Таким чином можна дійти до висновку, що ефективне розпізнання фішингової атаки може виступати інструментом у подоланні цієї загрози.

РОЗДІЛ 2.
ДОСЛІДЖЕННЯ ІСНУЮЧИХ МЕТОДІВ РОЗПІЗНАВАННЯ ФІШИНГОВИХ САЙТІВ

2.1. Підходи до розпізнавання фішингу

Є 2 основні підходи до розпізнавання фішингу: навчання користувача і класифікація за допомогою програмних засобів.

1) *Навчання користувача.* Користувачів можна навчати для їх кращого розуміння поняття та природи фішингових атак, що в свою чергу приведе до коректного розпізнаванню фішингових і справжніх сайтів. Навчання користувачів направлено на підвищення вірогідності розпізнавання цими користувачами фішингових атак..
2) *За допомогою програмних засобів.* Цей підхід спрямований на більш точну класифікацію фішингових і справжніх сайтів та зменшення розриву за рахунок відсутності «людської похибки». Це важливий недолік що потребує уваги, так як навчання користувачів повинно проводитися періодично і не завжди дає необхідний результат. Окрім цього, не завжди є необхідні фінанси, змога та спеціалісти для повного навчання користувача. Приклад неможливо навчання користувача - сайт компанії PayPal, оскільки остання має надзвичайно велику базу активних користувачів.

Точність розпізнавання може бути поліпшена в процесі навчання класифікатора (будь то людина або ПЗ). У користувацького розпізнавання

якість покращується шляхом накопичення досвіду кінцевим користувачем. Завдяки своїм знанням, користувач після навчання зможем сам розрізняти сайт підробку та реальний ресурс

У випадку програмної класифікації, прогрес у розпізнанні може бути досягнуто в процесі тренування класифікатора побудованого на алгоритмах машинного навчання, чи шляхом поліпшенням правил виявлення в системі, побудованої на правилах.

Розпізнавання фішинговою атаки є першим кроком у протидії цій загрозі. На рисунку 2.1 за ображено схему підходів до розпізнавання фішингових атак.

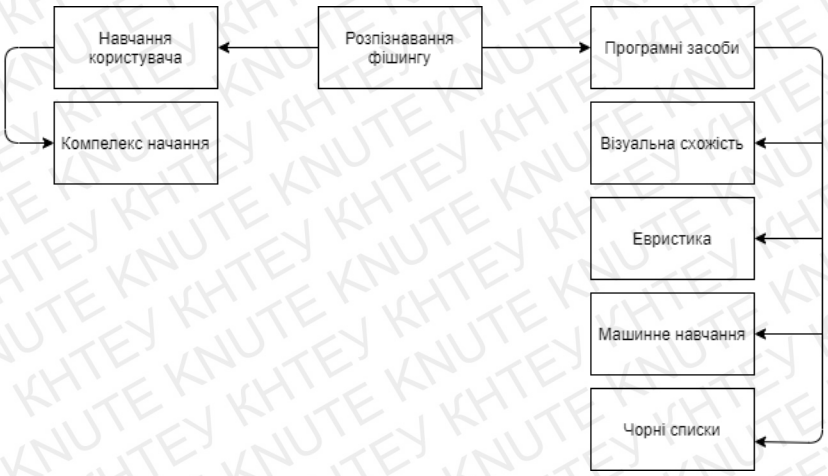

Рис. 2.1. Підходи до розпізнавання фішингової атаки

В рамках цієї роботи акцент ставиться на програмний підхід до виявлення, а саме на методах машинного навчання.

2.2. Програмний підхід до розпізнавання

Чорні списки представляють собою регулярно оновлюванні бази даних, що містять інформацію про раніше виявлені фішингові URL, IP-адреси. Недоліком таких списків є їх неспроможність захистити користувача від щойно створених фішингових сайтів – потребується певний час на виявлення і

занесення їх до списків. В роботі [6] чорні списки були визнані неефективними в виявленні нових сайтів і виявляли тільки 20% з них. Дослідження показало, що від 47% до 83% фішингових URL потрапили в чорні списки протягом 12 годин. Ця затримка є серйозною проблемою, так як значний відсоток фішингових кампаній згортається протягом першої доби з моменту створення.

Візуальна схожість. Метод розпізнавання ґрунтується на зовнішньому вигляді фішингових сайтів, а не аналізі вихідного коду та інформації мережевого рівня. Фішингові сайти імітують їх оригінальні, щоб ввести в оману користувача. При даному підході для прийняття рішення про тип сайту використовується набір характеристик текстового контенту, формат тексту, HTML теги, CSS, зображення і інші данні, які відповідають за візуальну складову сторінки.

Підходи, засновані на візуальній подібності, можуть швидко виявляти такі вбудовані об'єкти, присутні на фішинговою веб-сторінці. Так само методи, засновані на візуальній схожості, використовують підпис для ідентифікації фішингових веб-сторінок.

Евристика. Фішингова евристика являє собою набір характеристик, які присутні в реальних фішингових атаках, але не завжди гарантовано мають місце в кожному окремому випадку атак. Якщо визначено набір загальних евристичних тестів, то він може бути застосований для виявлення щойно створених сайтів. Недолік методу - існування ризику некоректної класифікації справжніх сайтів.

Сучасні веб-браузери та поштові клієнти побудовані з механізмами захисту від фішингу, такими як евристичні тести з метою виявлення фішингових атак. Евристики виявлення фішингу можуть бути включені до антивірусів.

Машинне навчання. Методи машинного навчання розглядають виявлення фішингових атак як проблему класифікації документів або кластеризації, де моделі будуються з використанням переваг алгоритмів машинного навчання і класифікації, таких як k-Nearest Neighbors (kNN), J48, Машини опорних векторів (SVM), нейронні мережі, Наївний Байєсівський метод і Логістична регресія.

2.3 Методи розпізнавання з машинним навчанням

Під цим підходом мається на увазі використання автоматичних класифікаторів, побудованих на алгоритмах машинного навчання та інтелектуального аналізу даних. Дані класифікатори працюють на стороні сервера і визначають клас переданої сторінки за набором характеристик. У таблиці 2.1. представлена порівняльна характеристику автоматизованих методів розпізнавання, основана на дослідженні [7].

Таблиця 2.1.

Порівняльна характеристика класифікаторів

Алгоритм	Основний принцип	Потреба навчання	Переваги	Недоліки
Naïve Bayes	•Кожен параметр класифікованих даних розглядається незалежно від інших параметрів класу. •Побудовано на теоремі Байеса. •Дозволяє передбачити клас використовуючи ймовірність.	Цей метод потребує навчання, оскільки алгоритм використовує розмічений набір даних для побудови таблиці.	•Алгоритм складається з простої арифметики (множення і ділення) •Швидке обчислення •Добре працює з великими розмірами	Покладається на припущення незалежності і буде погано працювати, якщо це припущення не буде виконано

Продовження таблиці 2.1.

LR	•Використовує лінійне рівняння з незалежними показниками для передбачення значення. •У центрі аналізу перебуває задача оцінки шансів на подію.	Цей метод потребує навчання	•Легко інтерпретовані результати •Модель є функціональною при прогнозуванні двійкових даних	•Потрібно більше статистичних припущень перед застосуванням•Більш функціонально зі змінними, які мають лінійну залежність, ніж комплексну •Точність прогнозування чутлива до вхідних даних
J48	•Алгоритм будує класифікатор в формі дерева рішень. •У кожній точці блок-схеми ставиться питання про значимість тієї чи іншої ознаки, в залежності від озна екземпляри потрапляють в клас. •Реалізує алгоритм Quinlan C4.5 •Будує дерева рішень на основі маркованих навчальних даних.	Цей метод потребує навчання, тут тренувальний набір даних розмічується класами.	•Проста інтерпретація •Висока швидкість роботи •Вихідні дані легко розуміються людиною	•Схильність до перенавчання •Можливі проблеми з діагональними межами рішення
Neural Network	•Організовано розташуванням взаємопов'язаних нерозпізнаних одиниць (нейронів) •З'єднання використовуються для відправки сигналів від одного нейрона до іншого.	Цей метод потребує навчання	•Дуже гнучкий, може бути використаний для задач регресії і класифікації •Можуть бути треновані з будь-якою кількістю вхідних даних•Після навчання прогнози швидкі	•Вимагають великих обчислювальних ресурсів •Збільшена тривалість роботи •Природа «чорного ящика»

Продовження таблиці 2.1.

kNN	•Контролйований метод машинного навчання •Для класифікації екземпляру використовуються клас k найближчих сусідніх екземплярів	Цей метод потребує навчання, оскільки kNN необхідний розмічений набір даних	•Легкий в розумінні •Легко реалізується •Залежно від вибору дистанційної метрики, kNN може показувати досить точні результати	•Може бути дуже ресурсозатратним \•Зашумлені дані можуть «зіпсуваи» kNN-класифікацію •Характеристики з великою кількістю значень можуть впливати на дистанційну метрику, по відношенню до характеристик з меншою кількістю значень

2.4. Огляд існуючих рішень

В роботі [8] автор використовував інформацію домену про посилання для розпізнавання фішингових сторінок. Одною з переваг їхнього методу було використання інформації про домен для виявлення фішингових сайтів, оскільки ця важлива характеристика не враховувалася в інших анти-фішингових методах. У цьому методі витягувалися всі прямі і непрямі посилання на обраний сайт. Далі, використовуючи вихідний код сторінки, всі домени пов'язані з прямими посиланнями витягувалися і поміщалися в набір S1. Потім непрямі посилання витягувалися і поміщалися в набір S2, після цього обидва набори об'єднувалися і витягувалися тільки загальні домени, з яких згодом, за допомогою DNS пошуку, витягувалися IP-адреси цих доменів. Результат дослідження показує, що accuracy, FP, FN, TN і TP запропонованого методу 99.54, 0.53, 0.5, 99.5 і 99.54%. Недоліком методу є залежність від різних зовнішніх методів типу DNS-пошуку і пошукових систем може вплинути на його ефективність.

У роботі [9] автор досліджує різні методи відбору характеристик з метою вибрати набір характеристик, що дозволить побудувати ефективні класифікатори з використанням інтелектуального аналізу даних. Так само автор досліджує чи може малий набір характеристик бути не менш ефективним при побудові добре прогнозованого класифікатора. Для цього він застосовує 2 алгоритми машинного навчання: введення правил IREP та дерево рішень C4.5.

В роботі [10] автори провели дослідження точності і чутливості 4 алгоритмів класифікації: NNet, Bayesian Network, SVM та Decision Tree. Їх завданням було виявити серед них найефективніший для розпізнавання фішингових сайтів. Результат - кращим алгоритмом став NNet.

Схожу роботу провів автор [11], оцінивши алгоритми Naïve Bayes, Support Vector Machine (SVM), Neural Net (NN), Random Forest (RF), IBK lazy classifier and Decision Tree (J48). За результатами найточнішими виявилися RF та kNN з показником Accuracy вище 97%.

В дослідженні [12] для відбору ознак використовувався метод Wrapper-based feature selection (WBFS). Цей метод використовує індуктивний класифікатор для оцінки підмножини характеристик. WBFS зазвичай дає найбільш ефективні характеристики, встановлені для цього конкретного виду класифікатора. Тому в цій роботі WBFS використовувався для вибору найбільш впливових характеристик, які можна використовувати для відмінності фішингу від справжніх веб-сайтів.

У дослідженні [13] була запропонована модель визначення, є сайт фішинговим чи ні. Використовувалося 6 різних алгоритмів класифікації (всі засновані на машинному навчанні): Naive Bayes, J48, SVM, Logistic Regression, Neural Network і ледачий класифікатор IBK з 92,7846%, 95,11%, 96,57%, 96,3%, 93,85%, 93,4039% точності класифікації відповідно. Співвідношення тренувального і тестового набору даних становить 70-30. Мета моєї роботи –

розширити їх дослідження для підвищення точності класифікації за допомогою описаних алгоритмів.

2.5. Алгоритм розпізнавання фішингового сайту

Огляд рішень з готових робіт по різним підходам до розпізнавання фішингових сайтів дозволив визначити ефективний підхід до вивчення цього питання за допомогою технік інтелектуального аналізу даних і методів машинного навчання. Спочатку досліджується точність розпізнавання фішингових сайтів з повним набором характеристик. Але оскільки витяг характеристик досить довга робота і вимагає багато часу, має сенс скоротити кількість характеристик (якщо це не значно позначиться на точності прогнозів) що і є предметом дослідження в даній роботі. Також підвищити точність прогнозування можливо за рахунок вдалої комбінації декількох класифікаторів, це буде представлено в останній частині дослідження. На рисунку 2.2. представлена поетапна схема реалізації процесу.

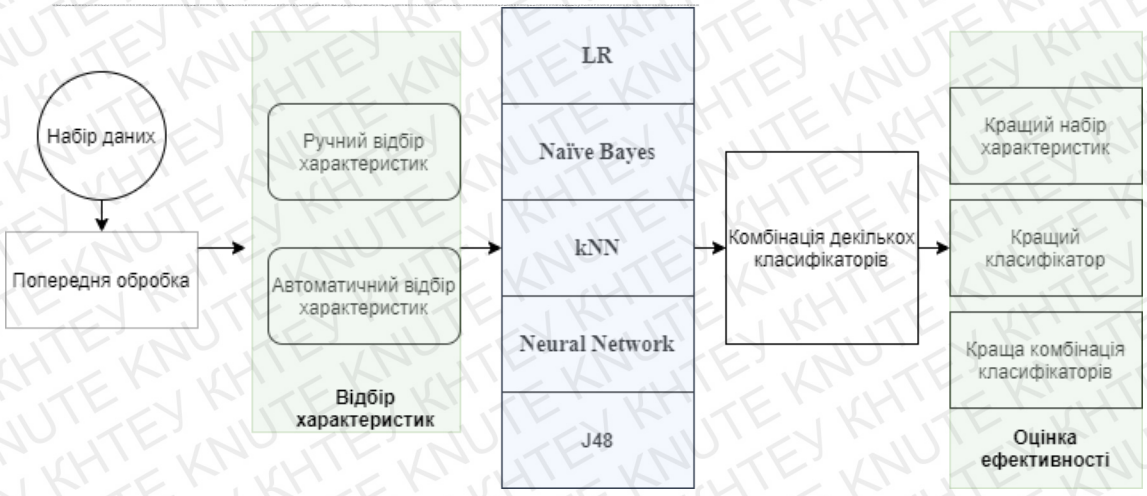

Рис. 2.2. Процес розпізнавання фішингового сайту

2.6. Попередня обробка

На етапі попередньої обробки виконується збір даних. Для спрощення роботи використовується набір даних з уже витягнутими характеристиками. Вони представлені у вигляді матриці, з колонками, що відповідають за кожну з 30 характеристик, а рядки представляють наявність тієї або іншої характеристики в дослідженому екземплярі набору.

Значення характеристики може бути представленим трьома числовими значеннями: 1 – надійний, 0 – підозрілий, -1 – фішинг. Остання колонка набору даних відповідає за кінцевий розподіл екземпляру на фішинг чи ні, представлена тільки двома значеннями: 1 і -1. Список характеристик представлений 4 групами розділеними за належністю характеристики до тієї чи іншої частини веб-сторінки. Перша група складається з характеристик адресного рядка і знаходиться в таблиці 2.2. Друга описує характеристики домену і представлена у таблиці 2.3. Третя група містить характеристики HTML і JavaScript, і продемонстрована в таблиці 2.4. Четверта, складається з аномальних характеристик і представлена в таблиці 2.5. Характеристики взяті з роботи [14], також у цій роботі доведена їх ефективність та впливовість в прогнозуванні фішингових та справжніх веб-сторінок.

Таблиця 2.2.

Характеристики адресного рядка

Назва характеристики	**Опис**
Using the IP Address	IP адреса використовується замість доменного імені, іноді трансформується у шістнадцятковий вигляд
Long URL	Фішери можуть використовувати довгу URL адресу щоб приховати підозрілу частину

Продовження таблиці 2.2.

Using URL Shortening Services “TinyURL”	Скорочення URL-адреси - це метод у "Всесвітній павутині", в якому URL може бути значно меншим за довжиною і привести до необхідної веб-сторінки. Це досягається за допомогою "HTTP Redirect" на короткому доменному імені, яке посилається на веб-сторінку з довгим URL
URL’s having “@” Symbol	Використання @ призводить до того що браузер ігнорує все що було до @, а справжня адреса йде за цим символом
Redirecting using “//”	Наявність у URL "//" означає що користувач буде перенаправлений на інший сайт
Adding Prefix or Suffix Separated by (-) to the Domain	Дефіс рідко використовується у перевірених адресах
Sub Domain and Multi Sub Domains	URL може мати не більше 2 суб-доменів, це домен першого рівня(країни) та другого рівня
HTTPS (Hyper Text Transfer Protocol with Secure Sockets Layer)	Наявність SSL сертифікату, його довіреність для вік
Domain Registration Length	Засновано на тому що фішинговий сайт живе короткий період часу

Таблиця 2.3.

Характеристики домену

Назва характеристики	**Опис**
Age of Domain	Співставляється з базою WHOIS, якщо вік менше 6 місяців - фішинг
DNS Record	Немає запису DNS для домену
Website Traffic	Вимірюється популярність сайту через визначення кількості відвідувачів і кількості сторінок що вони відвідали
PageRank	Рейтинг від 0 до 1, вказує наскільки важлива сторінка в інтернеті

Продовження таблиці 2.3.

Google Index	Чи індексується сторінка у google
Number of Links Pointing to Page	Чим більше сторінок посилається на дану тим більша імовірність що сторінка перевірена
Statistical-Reports Based Feature	Чи немає сайту у списку фішингових які збирає спеціальний сервіс, наприклад PhishTank

Таблиця 2.4.

Характеристики HTML та JavaScript

Назва характеристики	**Опис**
Website Forwarding	Скільки разів веб-сайт був перенаправлений
Status Bar Customization	Фахівці можуть використовувати JavaScript для показу підробленої URL-адреси в рядку стану для користувачів. Щоб витягти цю функцію, ми повинні викопати вихідний код веб-сторінки, зокрема подію "onMouseOver", і перевірити, чи вона вносить будь-які зміни у рядок стану.
Disabling Right Click	Фішери використовують JavaScript, щоб вимкнути функцію правої кнопки миші, так що користувачі не можуть переглядати та зберігати вихідний код веб-сторінки.
Using Pop-up Window	Просить користувачів надати персональну інформацію у спливаючому вікні
IFrame Redirection	Використання Iframe вказує на фішинг

Таблиця 2.5.

Аномальні характеристики

Назва характеристики	**Опис**
Submitting Information to Email	Веб-форма дозволяє користувачеві подавати свою особисту інформацію, яка спрямовується на сервер для обробки. Фішер може перенаправити інформацію користувача на його особисту електронну пошту. З цією метою можна використовувати мову скриптів на сервері. Ще однією функцією на стороні клієнта, яка може бути використана для цієї мети, є функція "mailto:"
URL of Anchor	Якір - це елемент, визначений тегом <a>. Ця характеристика рахується як попередня
Links in <Meta>, <Script> and <Link> tags	Звичайні веб-сайти використовують теги <Meta>, щоб запропонувати метадані про документ HTML; <Script> теги для створення клієнтського сценарію; і теги <Link> для отримання інших веб-ресурсів. Очікується, що ці теги пов'язані з тим доменом що і веб-сторінка
Server Form Handler (SFH)	SFH, які містять порожній рядок або "about: blank", вважаються сумнівними. Крім цього, якщо доменне ім'я в SFH відрізняється від доменного імені веб-сторінки, це показує, що веб-сторінка підозріла.
Request URL	Перевіряє чи завантажуються медіа ресурси з того ж домену
Abnormal URL	Перевіряється за допомогою WHOIS бази даних

2.7. Відбір характеристик

Для відбору характеристик використовується два різних шляхи, це ручний відбір і автоматизований. Алгоритм відбору представлений на рисунку 2.3.

Рис. 2.3. Відбір характеристик

Ручний відбір характеристик – це відбір відповідно до чотирьох груп описаних вище, що ґрунтується на структурі веб-сторінки. Таким чином згенеровано 9 піднаборів характеристик і надалі досліджувалася їх працездатність. Дані піднабори представлені в таблиці 2.6.

Таблиця 2.6.

Групи ручного відбору характеристик

Група характеристик	**Кількість характеристик**
Всі характеристики	30
Характеристики адресного рядка	12
Аномальні характеристики	6
Характеристики HTML & JS	5

Продовження таблиці 2.6.

Група характеристик	Кількість характеристик
Характеристики домену	7
Всі характеристики крім хар. адресного рядка	18
Всі характеристики крім аномальних	24
Всі характеристики крім HTML & JS	25
Всі характеристики крім доменних	23

Автоматизований відбір характеристик проводиться з використанням трьох алгоритмів відбору характеристик та відповідно генерує три групи ознак. Використовувалися наступні алгоритми:

- Consistency subset evaluator, який оцінює значимість піднабору характеристик за рівнем узгодженості значень класу, коли навчальні екземпляри проектуються на підмножину характеристик. Узгодженість будь-якої підмножини ніколи не може бути нижче, ніж у повного набору характеристик, тому звичайною практикою є використання цього оцінювача піднаборів в поєднанні з випадковим або вичерпним пошуком, який шукає найменшу підмножина з узгодженістю, рівній узгодженості повного набору ознак. [15]
- Correlation-based feature subset evaluator оцінює значимість підмножини характеристик, з огляду на індивідуальну прогностичну здатність кожної функції, а також ступінь надмірності між ними. [16]

- Wrapper subset evaluator оцінює набори характеристик, використовуючи схему навчання. Перехресна перевірка використовується для оцінки точності схеми навчання для набору атрибутів. [17]

Таблиця 2.7. показує яку техніку відбору і які алгоритми пошуку використовувалися, а також кількість відібраних ознак.

Таблиця 2.7.

Групи автоматичного відбору характеристик

Метод відбору	**Метод пошуку**	**Кількість характеристик**
Consistency subset evaluation	Greedy Stepwise	23
Wrapper subset evaluation	Best First	19
Correlation-based feature subset evaluation	Greedy Stepwise	9

2.8 Оцінка классифікаторів

Були обрані 5 найпоширеніших алгоритмів класифікації для тренування і тестування точності прогнозування фішингових сайтів на обраних групах характеристик. Основою вибору цих алгоритмів стала їх різноманітність стратегій тренування моделі класифікації і побудови правил та різні механізми навчання і тестування. Кожен з алгоритмів є представником одного з класів машинних методів навчання і класифікації: C4.8(Дерево прийняття рішень), Neural Network, Naïve Bayes, Logistic Regression(LR) та K-nearest neighbors(kNN).

Для оцінки класифікаторів машинного навчання, які використовуються для прогнозування фішингових веб-сайтів, використовувалася 10-кратна перехресна перевірка. При 10-кратної перехресній перевірці набір даних ділиться на 10 непересічних наборів даних однакового розміру. Кожен набір даних потім використовується в якості тестового набору даних, а решта 9 наборів даних об'єднуються і використовуються в якості навчального набору даних для тренування класифікатора. Потім цей процес запускається 10 разів. Точність розраховується для кожного прогону. Таким чином, остаточна точність навчання із цього набору даних є середнім значенням 10 точності для всіх прогонів.

2.9. Запропонована модель класифікації

Алгоритм прогнозування збудований на комбінації декількох класифікаторів полягає в наступному: вибираються три класифікатора з найвищою точністю прогнозування та мінімальні часом витраченим на побудову моделі, потім будується схема наведена на малюнку 2.4., в якій 2 з алгоритмів виносять свій прогноз по заданому екземпляру - є сайт фішинговим або справжнім. Якщо прогноз збігається, то екземпляру присвоюється відповідний результат, якщо прогноз відрізняється, то в справу вступає третій класифікатор і вже він остаточно визначає до якого класу належить екземпляр.

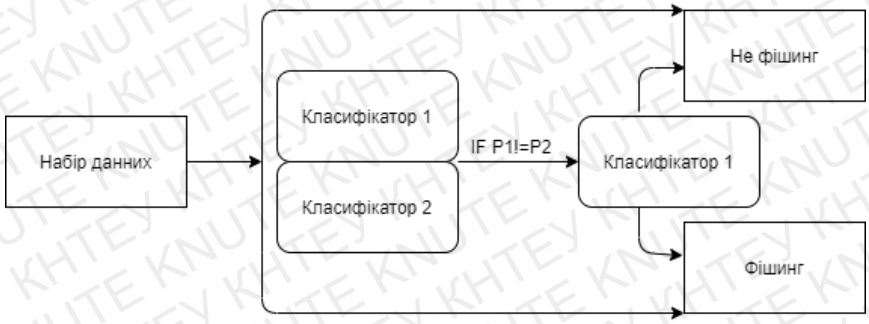

Рис. 2.4. Модель класифікатора

2.10. Висновки до розділу 2

У цьому розділі розглянуто методи розпізнавання фішингових сайтів, які полягають у пошуку візуальної схожості між справжнім та фальшивим сайтом.

До розглянутих машинних методів відносяться Logistic regression, J48(Decision tree), Naïve Bayes, Neural Network, та k-nearest neighbors. Кожен з методів має свої переваги і недоліки, а також мають різний принцип дії, це дозволить відшукати групу методів, які найбільш ефективні в даній конкретній ситуації.

Також у розділі 2 запропоновано підхід до знаходження ефективної моделі класифікатора, що розпізнає фішингові сайти з покращеною точністю. Підхід має декілька основних етапів. На першому відбулася попередня обробка даних, що складається зі збору даних, тобто формування датасету, а також витягнення характеристик и представлення їх у матричному вигляді. Другий етап, відбір характеристик, полягає у пошуку піднабору характеристик за яких точність розпізнавання майже не зміниться. Далі відбулася оцінка класифікаторів, для пошуку найефективнішого, і на завершення побудовано модель з 3 найкращих класифікаторів з поліпшеною точністю розпізнавання.

РОЗДІЛ 3.
АНАЛІЗ РЕЗУЛЬТАТІВ ДОСЛІДЖЕННЯ

3.1. Інструментарій та набір даних

Для роботи був використаний набір даних, що містить 10265 екземплярів, з яких 6045 надійних і 4220 фішингових. Дані отримані з сервісу UCI Machine Learning Repository. Витяг характеристик вбудовано в даний датасету, в якому кожен сайт представлений як вектор характеристик, а колонки показують приналежність сайту до одного з двох класів. Приклад частини датасету продемонстрований на рисунку 3.1.

ıg_IP_Ad	JRL_Lengt	rtining_Ser	ng_At_Syr	_slash_red	refix_Suffi	g_Sub_Do	Lfinal_Sta	registeratio	Favicon	port	TTPS_toke
-1	1	1	1	-1	-1	-1	-1	-1	1	1	-1
1	1	1	1	1	-1	0	1	-1	1	1	-1
1	1	0	1	1	-1	-1	-1	-1	1	1	-1
1	1	0	1	1	-1	-1	-1	1	1	1	-1
1	-1	0	1	1	-1	1	1	-1	1	1	1
-1	-1	0	1	-1	-1	1	1	-1	1	1	-1
1	-1	0	1	1	-1	-1	-1	1	1	1	1
1	1	0	1	1	-1	-1	-1	1	1	1	-1
1	-1	0	1	1	-1	1	1	-1	1	1	-1
1	-1	1	1	1	-1	-1	1	-1	1	1	1

Рис. 3.1. Частина датасету

Файл з даними був конвертований в формат ARFF для зручного використання в програмі WEKA. WEKA – це набір алгоритмів машинного навчання використовуються для глибокого аналізу даних, так само дана програма містить інструментарій для задач початкової обробки даних, задач регресії, класифікації, кластеризації та візуалізації. Алгоритми можуть бути застосовані безпосередньо, тобто з наявного набору, або отримані з коду на мові Java. Так само WEKA підходить для розробки нових схем машинного навчання.

3.2. Критерії оцінки

Кілька експериментів були приведені в різних сценаріях, експеримент і результат оцінювалися з використанням декількох вимірювань, порівнювалися результати декількох експериментів і були виведені результати.

Першим критерієм є accuracy (3.1). Це показник кількості вірних прогнозів, що були отримані в результаті запуску моделі на тестовому наборі даних і наступному порівнянні результатів з фактичними, що знаходяться в датасеті. Але у цього критерію є свої мінуси. Якщо набір даних незбалансований, то в рамках одного класу точність може бути високою, а інший визначається досить погано.

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN} \tag{3.1}$$

Precision (точність) (3.2) і recall (повнота) (3.3) є критеріями, використовуваними для оцінки в більшій частині алгоритмів витягу інформації. Точність системи в межах класу – це частка екземплярів, дійсно належать даному класу щодо всіх документів, які система віднесла до цього класу. Повнота системи є часткою знайдених класифікатором екземплярів, що належать класу відносно всіх екземплярів цього класу в тестовій вибірці.

$$Precision = \frac{TP}{TP+FP} \tag{3.2}$$

$$Recall = \frac{TP}{TP+FN} \tag{3.3}$$

Значення TP, FP, TN, FN знаходяться в confusion matrix (матриця неточностей): True Positive(TP) є кількість фішингових сторінок ідентифікованих як фішинг. False Positive(FP): кількість надійних сторінок ідентифікованих як фішинг. False Negative(FN): кількість фішингових сторінок ідентифікованих як надійні. True Negative(TN): кількість

правомірних сторінок ідентифікованих як надійні. Представлення занчень матриці можна побачити у таблиці 3.1.

Таблиця 3.1.

Confusion matrix

Насправді надійний	**Класифіковано як фішинг**	**Класифіковано як надійний**
НІ	TP	FN
ТАК	FP	TN

3.3. Результати оцінки відбору характеристик

В ході експерименту розраховуються описані раніше метрики для кожного набору характеристик отриманого в результаті ручного та автоматичного відбору. Також оцінюється вплив кожної групи на кінцевий результат. У підсумку, отримані результати порівнюються з результатами отриманими при оцінці на наборі з усіх характеристик, щоб рекомендувати найкращий підхід до відбору ознак. Критерієм відбору в даному випадку є так само розмір набору ознак. Ррозрахунки були проведені для набору з усіх характеристик. Результати представлені в таблиці 3.2.

Таблиця 3.2.

Результати тесту для всіх характеристик

Алгоритм	Accuracy	Precision	Recall
Naïve Bayes	92,980	0,930	0,930
Neural Network	96,900	0,969	0,969
J48(c4.8)	95,870	0,959	0,959
kNN	97,180	0,972	0,972
Logistic Regression	93,990	0,940	0,940

Графічне представлення результатів – рисунок 3.2.

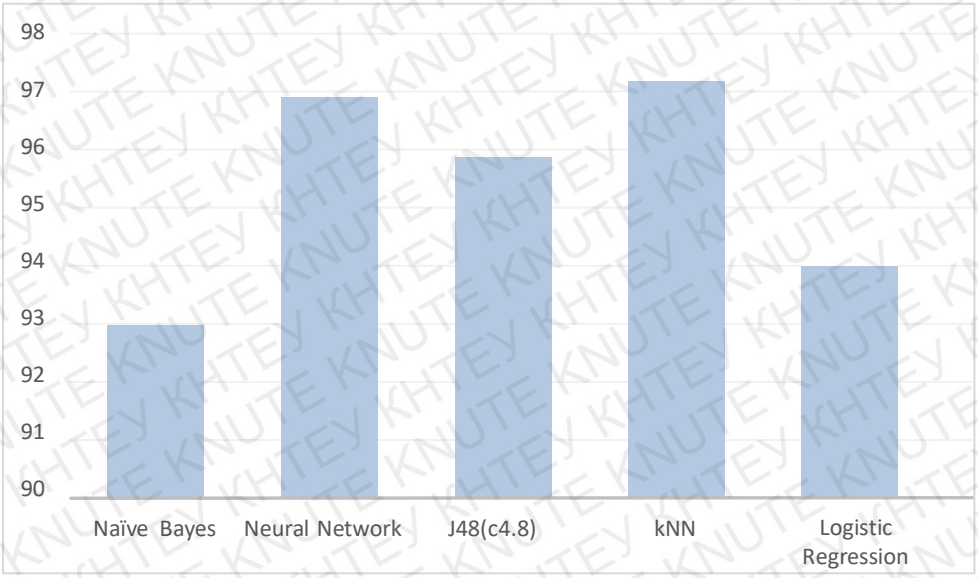

Рис. 3.2. Accuracy для всіх характеристик

Результати показують кращі оцінки за всіма трьома параметрами у алгоритмів Neural Network і kNN, які дорівнюють відповідно 96,9 і 97,18. Найгірший результат показав алгоритм Naïve Bayes, 92,98%.

Далі розглядаються тільки характеристики адресного рядка, результати представлені в таблиці 3.3.

Таблиця 3.3.

Результати тесту для характеристик адресного рядка

Алгоритм	Accuracy	Precision	Recall
Naïve Bayes	89,792	0,898	0,898
Neural Network	90,742	0,908	0,907
J48(c4.8)	90,385	0,904	0,904
kNN	90,837	0,909	0,908
Logistic Regression	89,765	0,898	0,898

Графічне відображення знаходиться на рисунку 3.3.

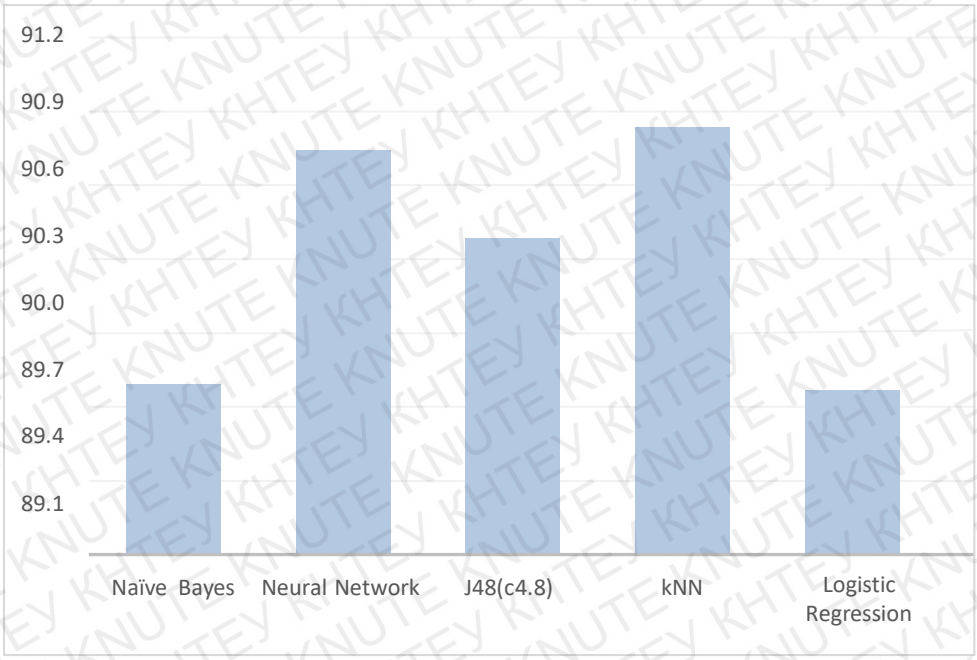

Рис. 3.3. Accuracy для характеристик адресного рядка

За результатами видно, що показники впали на 6,5% для алгоритму kNN і на 3,4% для Naïve Bayes, який і в цей раз показує найгірші результати. Найкращий результат у алгоритмів Neural Network та kNN.

Далі розглядаються тільки характеристики аномальні характеристики, результати представлені в таблиці 3.4 і на рисунку 3.4.

Таблиця 3.4.

Результати тесту для аномальних характеристик

Алгоритм	Accuracy	Precision	Recall
Naïve Bayes	87,024	0,879	0,870
Neural Network	87,445	0,880	0,874
J48(c4.8)	87,309	0,879	0,873
kNN	87,191	0,875	0,872
Logistic Regression	87,187	0,880	0,872

Графічне відображення результатів на рисунку 3.4.

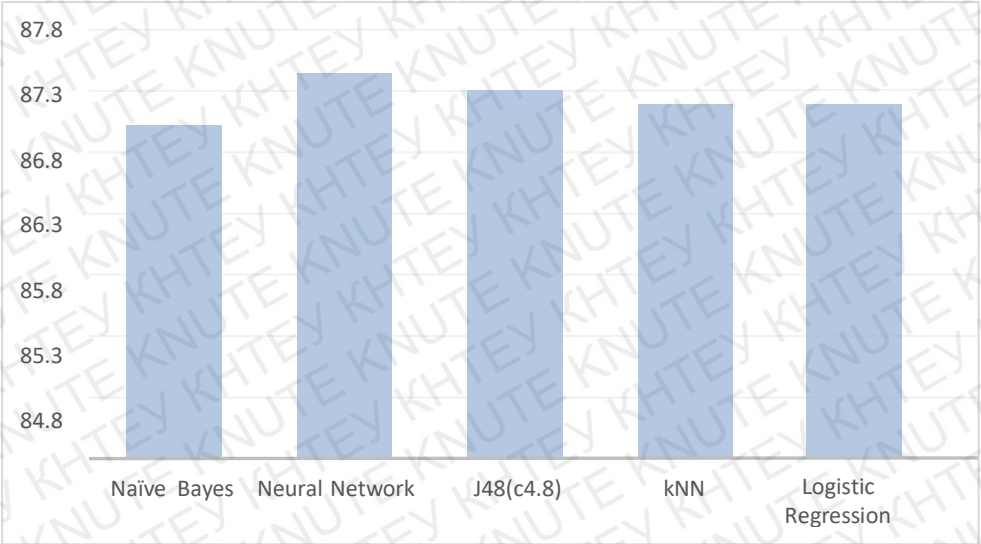

Рис. 3.4. Accuracy для аномальних характеристик

За отриманими даними видно що результати тільки погіршилися, а найменш точним алгоритмом виявився Naïve Bayes. Точність прогнозування впала на 10% для kNN та Neural Network. Можна зробити висновок, що цей набір характеристик не підходить для використання у подальшому.

Далі розглядаються тільки характеристики HTML і JS, результати представ- лені в таблиці 3.5.

Таблиця 3.5.

Результати тесту для характеристик HTML и JS

Алгоритм	Accuracy	Precision	Recall
Naïve Bayes	56,101	0,573	0,561
Neural Network	55,789	0,629	0,558
J48(c4.8)	57,137	0,654	0,571
kNN	57,210	0,652	0,572
Logistic Regression	56,178	0,556	0,562

Графічне відображення результатів на рисунку 3.5.

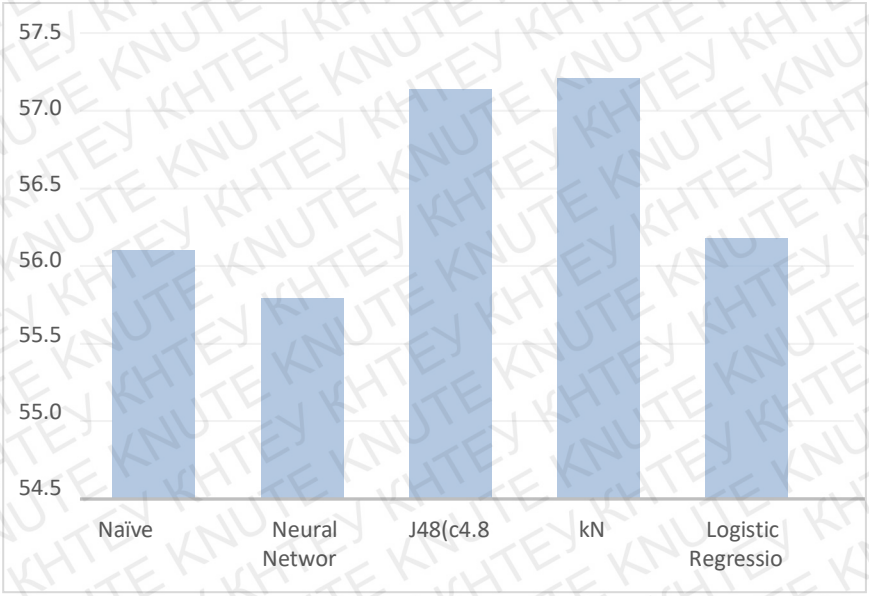

Рис. 3.5. Accuracy для характеристики HTML та JavaScript

Значно впали показники - це свідчить про повну непридатність даного набору ознак для використання в подальшому. Однак і в цьому наборі найкращі результати показали алгоритми kNN і J48.

Далі розглядаються тільки характеристики домену, результати представлені в таблиці 3.6.

Таблиця 3.6.

Результати тесту для характеристик домену

Алгоритм	Accuracy	Precision	Recall
Naïve Bayes	70,927	0,709	0,709
Neural Network	73,641	0,737	0,736
J48(c4.8)	74,170	0,741	0,742
kNN	74,545	0,745	0,745
Logistic Regression	71,646	0,716	0,716

Графічне відображення результатів на рисунку 3.6.

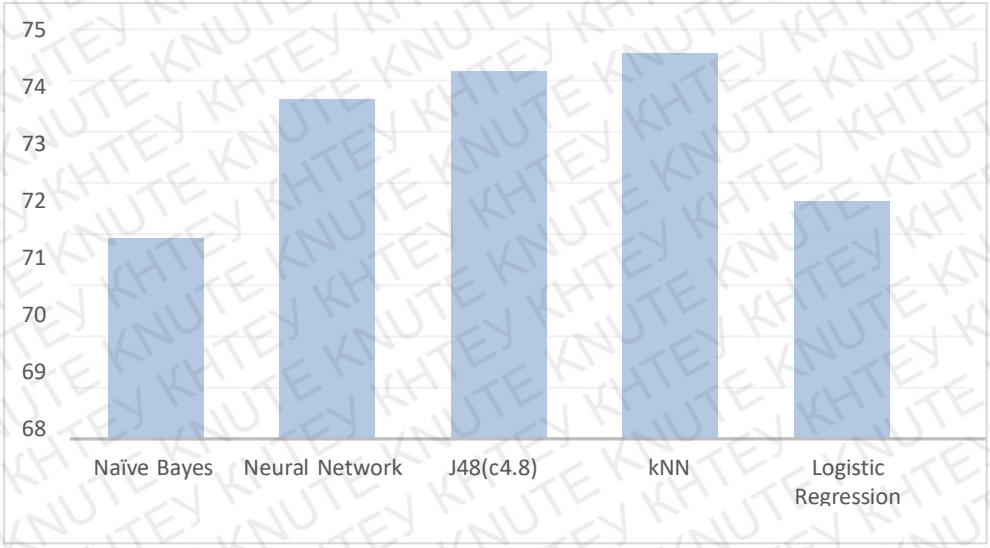

Рис. 3.6. Accuracy для характеристик домену

В даному випадку кращі показники всього лише в районі 74%, що не є підходящим варіантом для прогнозування фішингу. Алгоритм kNN все ще показує кращий результат.

Далі розглядаються всі характеристики за винятком характеристик адресного рядка, результати представлені в таблиці 3.7.

Таблиця 3.7.

Результати тесту для всіх характеристик крім характеристик адресного рядка

Алгоритм	Accuracy	Precision	Recall
Naïve Bayes	87,626	0,879	0,876
Neural Network	91,524	0,916	0,915
J48(c4.8)	91,361	0,916	0,914
kNN	92,583	0,926	0,926
Logistic Regression	88,621	0,889	0,886

Графічне відображення результатів на рисунку 3.7.

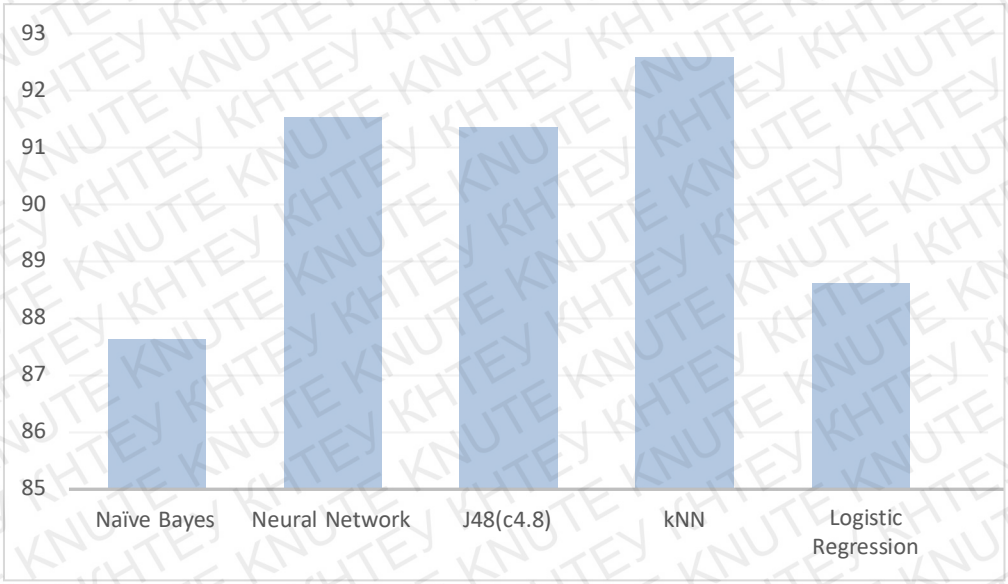

Рис. 3.7. Accuracy для всіх характеристик крім хар. адресного рядка

Результати показали що найкраща точність досягнута алгоритмом kNN 92,583, поряд Neural Network і J48 з різницею лише в 1%. Показники в загальному впали на 4,7%. Порівняно з вищерозглянутими наборами цей має поки найкращий результат.

Далі розглядаються всі характеристики за винятком аномальних характеристик, результати представлені в таблиці 3.8.

Таблиця 3.8.

Результати тесту для всіх характеристик крім аномальних характеристик

Алгоритм	Accuracy	Precision	Recall
Naïve Bayes	90,204	0,902	0,902
Neural Network	93,781	0,938	0,938
J48(c4.8)	93,152	0,932	0,932
kNN	94,568	0,946	0,946
Logistic Regression	91,28	0,913	0,913

Графічне відображення результатів на рисунку 3.8.

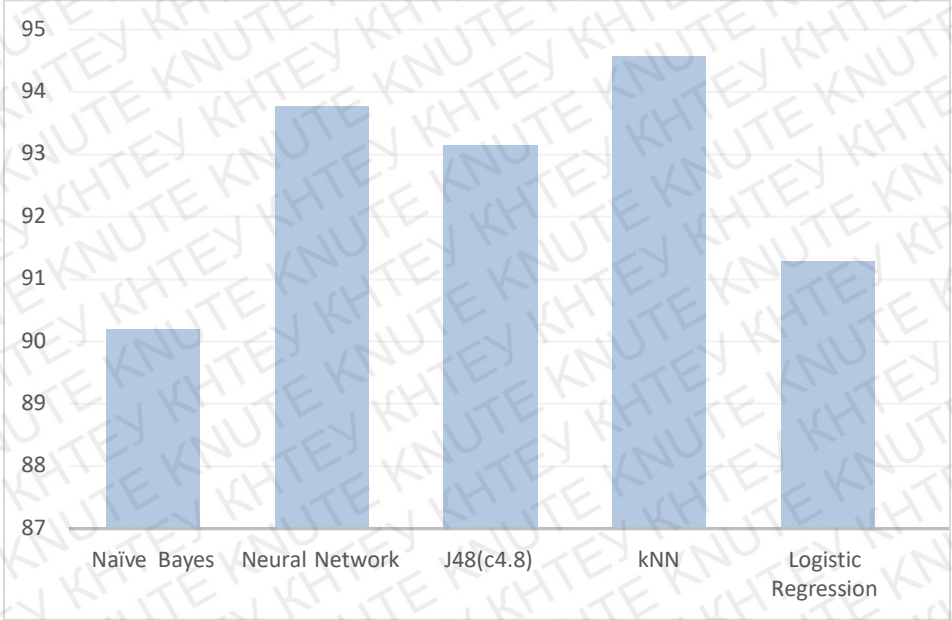

Рис. 3.8. Accuracy для всіх характеристик крім аномальних

Показники цього набору впали на 2,7%, це гарний результат, в лідерах залишаються алгоритми Neural Network, kNN і J48 (c4.8).

Далі розглядаються всі характеристики за винятком характеристик HTML і JS, результати представлені в таблиці 3.9 і на рисунку 3.9.

Таблиця 3.9.

Результати тесту для всіх характеристик крім характеристик HTML і JS

Алгоритм	Accuracy	Precision	Recall
Naïve Bayes	92,962	0,93	0,93
Neural Network	96,662	0,967	0,967
J48(c4.8)	95,889	0,959	0,959
kNN	97,087	0,971	0,971
Logistic Regression	93,731	0,937	0,937

Графічне відображення результатів на рисунку 3.9.

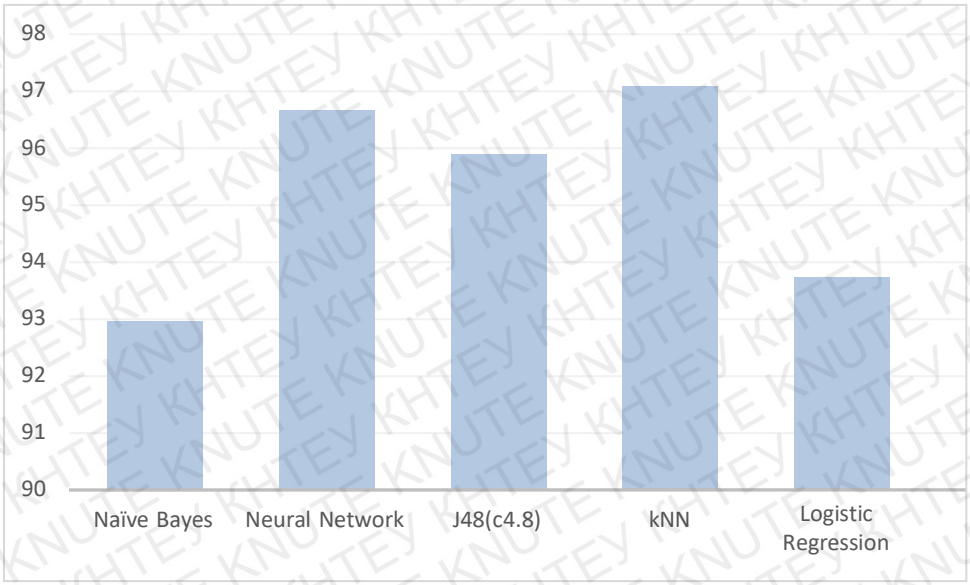

Рис. 3.9. Accuracy для всіх характеристик крім HTML & JS

Показники впали менш ніж на 1%, цей набір однозначно можна використовувати для прогнозування фішингових сайтів. Набір складається з 25 показників.

Далі розглядаються всі характеристики за винятком характеристик домену, результати представлені в таблиці 3.10.

Таблиця 3.10.

Результати тесту для всіх характеристик крім характеристик домену

Алгоритм	Accuracy	Precision	Recall
Naïve Bayes	92,741	0,928	0,927
Neural Network	94,627	0,946	0,946
J48(c4.8)	94,496	0,945	0,945
kNN	94,984	0,95	0,95
Logistic Regression	93,162	0,932	0,932

Графічне відображення результатів на рисунку 3.10.

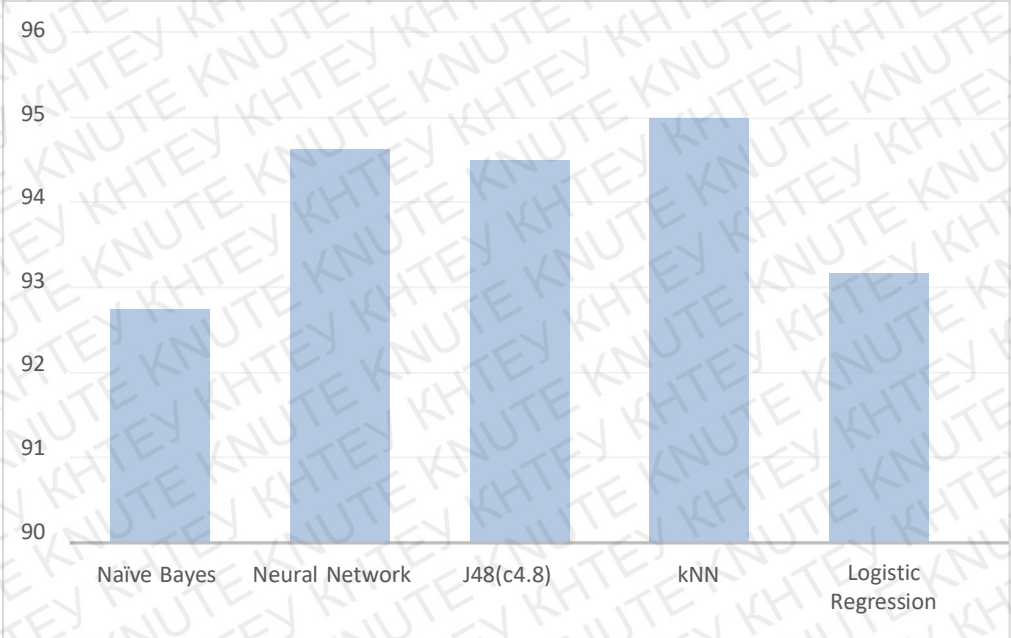

Рис. 3.10. Accuracy для всіх характеристик крім доменних

Точність результатів впала на 2-3%, цей набір є конкурентоспроможним у завданні розпізнання фішингових сайтів. Набір включає в себе 23 характеристики.

Підбиваючи підсумки в дослідженні вручну відібраних наборів кращі показники представили такі набори як:

- Всі характеристики крім HTML і JS (показники впали менш ніж на 1%)
- Всі характеристики крім доменних (показники впали менш ніж на 3%) Найточнішими алгоритмами класифікації виявилися kNN і Neural Network, J48 також показав хороші результати.
- Серед обраних найгіршим виявився Naïve Bayes.

Далі розглядаються набори автоматично відібраних ознак.

Результати дослідження всіх трьох груп представлені в таблиці 3.11 і на рисунку 3.11.

Таблиця 3.11.

Результати тесту accuracy для автоматично відібраних груп

Алгоритм	Група 1(cfs)	Група 2(consistency)	Група 3(wraper)
Naïve Bayes	92,668	92,899	92,474
Neural Network	94,491	96,662	95,884
J48(c4.8)	94,315	95,803	95,984
kNN	94,387	97,078	96,314
Logistic Regression	93,211	93,917	93,424

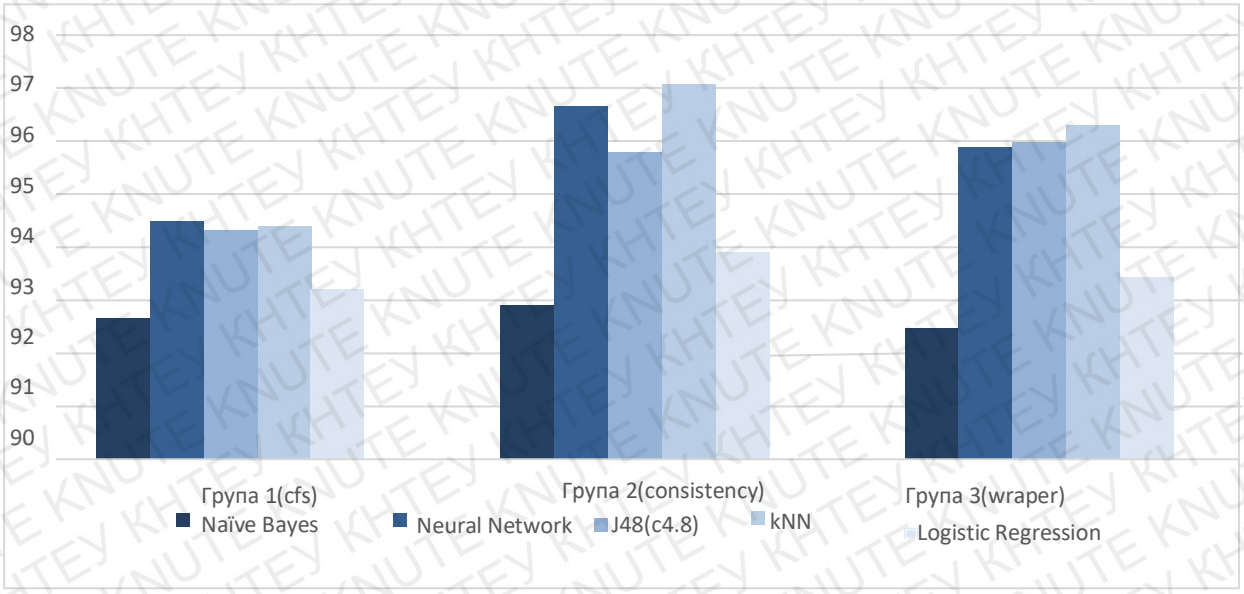

Рис. 3.11. Accuracy для автоматично відібраних груп

Згідно з одержаними результатами, розрахунки параметра accuracy показали, що найточнішим для виявлення фішингових сайтів є другий набір характеристик, він складається з 23 ознак і був отриманий методом Consistency subset evaluation, його найкраща точність дорівнює 97,078, а найгірша 92,956. Група складається з 23 ознак. Не далеко пішла третя група ознак, отриманих методом Wrapper subset evaluation і складається з 19 показників. Кращий результат в цьому піднаборі 96,314, а найгірший 92,374. Показники впали

менш ніж на 1%, це означає, що піднабір можна використовувати для передбачення фішингових сайтів. Група складається з 19 ознак. У першої групи ознак показники найгірші з усіх трьох, але обсяг групи всього 9 характеристик. Кращий результат в цій групі дорівнює 94,491, а найгірший 92,668, в порівнянні з початковим набором характеристик, точність прогнозування впала на менш ніж на 2% для кращого результату, а для гіршого майже не змінилася.

У таблиці 3.12. представлені зведені результати кращих груп ручного відбору та груп автоматичного відбору.

Таблиця 3.12.

Результати тесту accuracy для кращих відібраних груп

Алгоритм	**Автоматичний відбір**			**Ручний відбір**		**Всі**
	Група 1(cfs)	**Група 2(consistency)**	**Група 3(wraper)**	**Всі крім доменних**	**Всі крім HTML & JS**	
Naïve Bayes	92,668	92,899	92,374	92,741	92,956	92,980
Neural Network	94,491	96,662	95,884	94,627	96,662	96,900
J48(c4.8)	94,315	95,803	95,984	94,496	95,889	95,870
kNN	94,387	97,078	96,314	94,984	97,087	97,180
Logistic Regression	93,211	93,917	93,424	93,162	93,731	93,990
Кількість характеристик	9	23	19	23	25	30

З таблиці 3.12 видно, що результати групи «1 автоматична» і «все крім доменних» майже рівні з відзнакою 0,1-0,5, але оскільки група 1 складається всього з 9 ознак, її, очевидно, використовувати краще. Також видно схожість результатів 2-ї групи і групи «все крім HTML & JS», вони відрізняються на 0,01-0,2.

Для наочності результатів демонструється графік на малюнку 3.12.

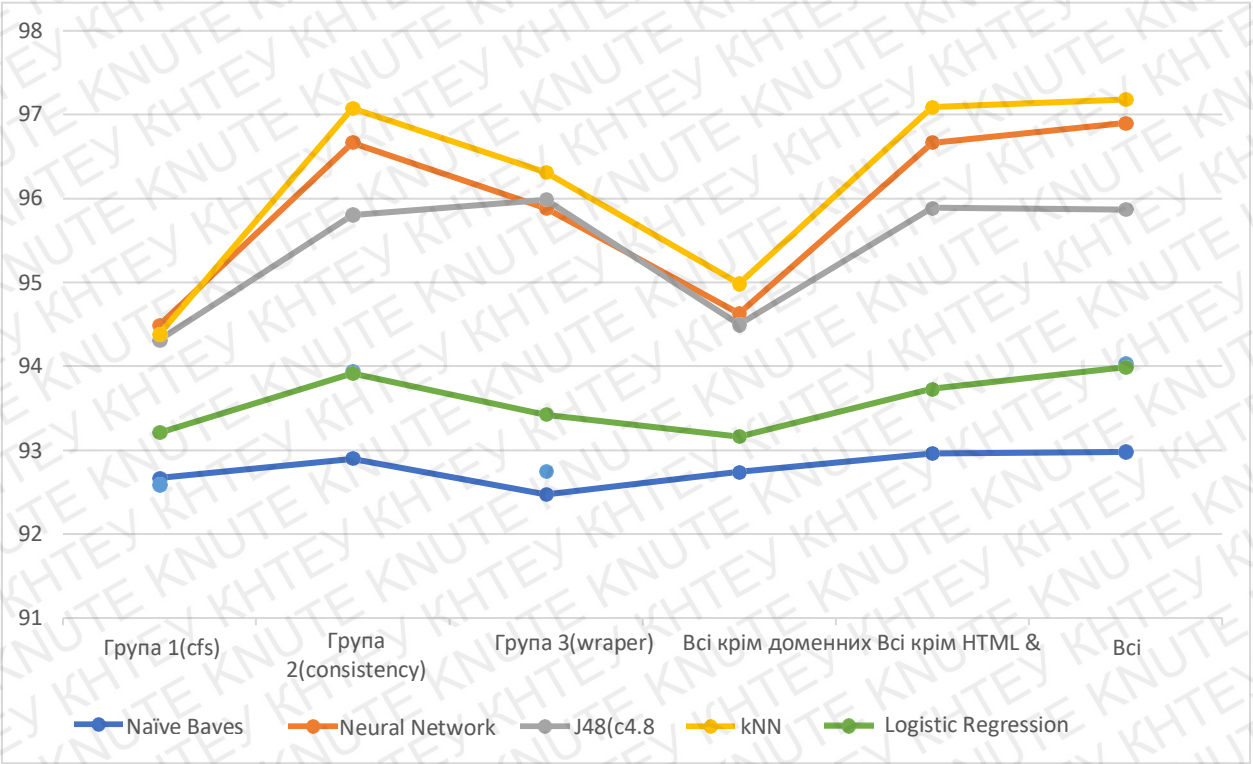

Рис. 3.12. Accuracy для кращих відібраних груп

3.4. Результати побудови комбінованої моделі

Метою дослідження так само була побудова комбінованої моделі класифікаторів, яка дозволила б з більшою точністю і акуратністю визначати фішингові сайти. Виходячи з вищенаведеного експерименту кращі результати показували kNN, Neural Network і Decision tree (J48), в порядку зменшення точності. Ще одним важливим критерієм відбору класифікаторів є його час роботи, так час витрачений на побудову моделі kNN одно 0,03 секунди, на побудову Neural Network 156,84 секунд та на побудову Decision tree (J48) - 0,24 секунда. Для прискорення роботи моделі замість Neural Network взяти наступний за ефективністю алгоритм - це Logistic Regression. Час побудови моделі, основаної на цьому алгоритмі - 1,11 секунди. Таким чином для

побудови комбінованої моделі класифікаторів обрано алгоритми kNN, LR і J48. Код алгоритму моделі знаходиться у додатку А.

Для оцінки роботи моделі було взято 5 груп характеристик, що мали найкращі результати у попередньому дослідженні, а також початковий набір. Результати оцінки accuracy представлені у таблиці 3.13.

Таблиця 3.13.

Результати тесту accuracy для комбінації класифікаторів

Група характеристик	Max accuracy до	Max accuracy після	% покращення
Група 1(cfs)	94,491	95,564	1,14
Група 2(consistency)	97,078	98,102	1,05
Група 3(wraper)	96,314	97,022	0,74
Всі крім доменних	94,984	95,78	0,84
Всі крім HTML & JS	97,087	98,34	1,29
Всі	97,18	98,766	1,63

В ході експерименту було встановлено, що завдяки новій схемі класифікації спостерігається покращення таких параметрів як accuracy, precision та recall. В таблиці 3.13 наведені результати тільки тесту accuracy, оскільки не має значної різниці між цим показником та двома іншими. Встановлене підвищення accuracy в межах 0,74 – 1,63%.

3.5. Висновки до розділу 3

В розділі 3 були проведені експерименти щодо пошуку найкращого класифікатора, найкращого піднабору характеристик та кращої комбінації класифікаторів. Результати оцінювались з точки зору трьох показників точності, це accuracy, precision та recall, а також враховувався час побудови моделі при відборі класифікаторів для подальшого використання у комбінованій моделі. За наведеними результатами можна зробити такі висновки:

1) Серед розглянутих груп характеристик ручного відбору найкращі результати продемонструвала група характеристик, де не враховувалися характеристики HTML&JavaScript, група містить 25 характеристик. Серед груп автоматичного відбору найкращі результати продемонструвала група відібрана методом Consistency subset evaluator, група складається з 23 характеристик. Різниця у результатах між цими двома наборами дуже незначна, складає 0,01% для найкращого алгоритму(kNN) та 0,06% для найгіршого(NB). Враховуючи кількість ознак у групі, що впливає на швидкість роботи, найкращою обрано групу 2(consistency). Слід зазначити, що група 1(cfs) має найбільшу швидкодію, оскільки у групі лише 9 ознак.
2) Найкращим класифікатором для цього завдання виявився kNN, з максимальною точністю 97,18%.
3) Щодо нової моделі класифікації, побудованої на простих алгоритмах класифікації kNN, LR, J48, точність розпізнавання виросла у середньому на 1,11%. Також було з'ясовано, що порядок класифікаторів у моделі не впливає на точність прогнозування. Найбільший показник має перевірка на всіх характеристиках зі значенням accuracy 98,766%.

ВИСНОВКИ ТА ПРОПОЗИЦІЇ

Фішингові сайти є постійною загрозою для користувачів Інтернет, загрозою, зо не втрачає свою актуальність з кожним новим роком. Серед існуючих технік розпізнавання фішингових сайтів, великий відсоток таких, що мають низьку точність: контент може бути таким самим як і на справжньому сайті, тому не розпізнається, через що рівень розпізнавання невисокий.

У цьому дослідженні було вивчено та проаналізовано різні підходи до виявлення фішингових сайтів та встановлено, що вони базуються на навчанні користувачів або використанні програмних засобів. Також було вивчено актуальні методи класифікації та відбору характеристик за допомогою алгоритмів машинного навчання.

Відібрано характеристики, які можуть бути ознаками фішингового сайту. Обрано за двома підходами, ручний відбір та автоматичний. Також було сформовано 11 груп для подальшого дослідження обраними алгоритмами класифікації. Визначено ступінь важливості та інформативності даних характеристик при виявленні фішингових сайтів. Окрім цього:

1) Проаналізовано алгоритми класифікації та обрано найбільш ридатні для розв’язання задачі розпізнавання фішингового сайту.
2) Розроблено програмне забезпечення та здійснено експериментальне дослідження.
3) Запропоновано нову модель класифікації даних, яка є комбінацією вже існуючих та розповсюджених класифікаторів, яка відрізняється покращеною точністю розпізнавання.

Практична цінність результатів полягає у можливості використання отриманого класифікатору для подальшого створення програмного забезпечення з метою розпізнавання фішингових сайтів. Останнє у комбінації з набором характеристик може бути впроваджений у антифішингові розширення для браузерів або використовуватися користувачем як інструмент для боротьби з фішингом.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Phishing Activity Trends Report 2nd Quarter 2020 [Електронний ресурс]. – 2020. – Режим доступу до ресурсу: https://docs.apwg.org/reports/apwg_trends_report_q2_2020.pdf
2. PhishTank [Електронний ресурс] – Режим доступу до ресурсу: https://www.phishtank.com/what_is_phishing.php.
3. How to Recognize and Avoid Phishing Scams [Електронний ресурс] // FEDERAL TRADE COMMISSION Consumer Information. – 2019. – Режим доступу до ресурсу: https://www.consumer.ftc.gov/articles/how-recognize-and-avoid-phishing-scams.OBD-II (Check Engine Light) Trouble Codes [Електронний ресурс]. - Режим доступу: https://www.obd-codes.com/trouble_codes/
4. Khonji M. hishing Detection: A Literature Survey [Електронний ресурс] / M. Khonji, Y. Iraqi, J. Andy // 2013 – Режим доступу до ресурсу: https://www.researchgate.net/publication/256841808_Phishing_Detection_A_Literature_Survey.
5. Augustine P. Study of Phishing Attacks and Preventions [Електронний ресурс] / P. Augustine, T. Dakpa // International Journal of Computer Applications (0975 – 8887). – 2017. – Режим доступу до ресурсу: https://www.ijcaonline.org/archives/volume163/number2/dakpa-2017-ijca-913461.pdf.
6. An Empirical Analysis of Phishing Blacklists [Електронний ресурс] / [S. Sheng, L. Cranor, B. Wardman та ін.] – Режим доступу до ресурсу: http://www.cs.cmu.edu/~jasonh/publications/ceas2009-testbed-final.pdf.

7. Top 10 algorithms in data mining [Електронний ресурс] / X. Wu, V. Kumar, R. Quinlan. – 2007. – Режим доступу до ресурсу: https://www.researchgate.net/publication/29467751_Top_10_algorithms_in_data_mining.
8. Gowtham R. Identification of phishing webpages and its target domains by analyzing the feign relationship [Електронний ресурс] / Ramesh Gowtham // Journal of Information Security and Applications 35:75-84. – 2017. – Режим доступу до ресурсу: https://www.researchgate.net/publication/317550308_Identification_of_phishing_webpages_and_its_target_domains_by_analyzing_the_feign_relationship.
9. Detection and Prediction of Phishing Websites using Classification Mining Techniques [Електронний ресурс] // International Journal of Computer Applications. – 2016. – Режим доступу до ресурсу: https://www.researchgate.net/publication/306124393_Detection_and_Prediction_of_Phishing_Websites_using_Classification_Mining_Techniques.
10. Abdeyazdan M. Detecting Internet Phishing Attacks Using Data Mining Methods [Електронний ресурс] / M. Abdeyazdan, A. Rayat Pisheh // 3rd International conference on Innovative Engineering Technologies. – 2016. – Режим доступу до ресурсу: https://pdfs.semanticscholar.org/4a43/b1278b3ecf07c0aaf7c20a8806c136814bae.pdf.
11. Thomas C. Detection of phishing URLs using machine learning techniques [Електронний ресурс] / Ciza Thomas // International Conference on Control Communication and Computing. – 2013. – Режим доступу до ресурсу: https://www.researchgate.net/publication/269032183_Detection_of_phishing_URLs_using_machine_learning_techniques.

12. Waleed A. Phishing Website Detection based on Supervised Machine Learning with Wrapper Features Selection [Електронний ресурс] / Ali Waleed // International Journal of Advanced Computer Science and Applications. – 2017. – Режим доступу до ресурсу: https://www.researchgate.net/publication/320131222_Phishing_Website_Detection_based_on_Supervised_Machine_Learning_with_Wrapper_Features_Selection.
13. МЕЖДУНАРОДНЫЙ ЖУРНАЛ ИНЖЕНЕРНЫХ ИССЛЕДОВАНИЙ И ТЕХНОЛОГИЙ (IJERT) [Електронний ресурс] – Режим доступу до ресурсу: https://www.ijert.org.
14. Rami M. A. An assessment of features related to phishing websites using an automated technique [Електронний ресурс] / M. A. Rami, F. Thabtah, T. Mccluskey // Conference: Internet Technology And Secured Transactions. – 2012. – Режим доступу до ресурсу: https://www.researchgate.net/publication/261081735_An_assessment_of_features_related_to_phishing_websites_using_an_automated_technique.
15. Class ConsistencySubsetEval [Електронний ресурс] – Режим доступу до ресурсу:https://weka.sourceforge.io/doc.stable/weka/attributeSelection/ConsistencySubsetEval.html
16. Class CfsSubsetEval [Електронний ресурс] – Режим доступу до ресурсу: http://weka.sourceforge.net/doc.stable/weka/attributeSelection/CfsSubsetEval.html
17 Class WrapperSubsetEval [Електронний ресурс] – Режим доступу до ресурсу:http://weka.sourceforge.net/doc.stable/weka/attributeSelection/WrapperSubsetEval.html

ТЕХНІЧНЕ ЗАВДАННЯ

Положення технічного завдання для розробки класифікатора фішингових web-сторінок.

1. Загальні відомості

Планові строки розробки: серпень 2020 – вересень 2020.

Потенційні користувачі: користувачі мобільних телефонів та ПК з будь-якою ОС, у якій передбачено користувацький інтерфейс та підтримка web-браузеру.

Призначення програмного рішення: захистити користувача від використання потенційно небезпечної web-сторінки, що містить ознаки фішингу.

Мета створення програмного рішення: опанування технологій побудови класифікаторів на основі машинного навчання.

2. Структура проекту

Класифікатор використовує датасет для аналізу у програмі WEKA. Аналіз відбувається на основі алгоритму, написаного на мові програмування Python. На етапі розробки користувацький інтерфейс не передбачено.

3. Функціональні вимоги

Класифікатор повинен з мінімальною похибкою вміти розрізнити фішингову web-сторінку та не фішингову.

					КНТЕУ 121 06з-16.МР		
Зм.	*Аркуш*	*№ докум.*	*Підпис*	*Дата*			
Зав. каф.		Криворучко О.В.		28.02.20	*Розробка аналітичної системи розпізнавання фішингових сайтів на основі штучного інтелекту*	*Стадія*	*Аркуш*	*Аркушів*
Керівник		Пашорін В.І.		28.02.20		*ТЗ*	*56*	*55*
Гарант		Криворучко О.В.		28.02.20		Факультет інформаційних технологій 2м курс, 6з група		
Розробив		Філяєв М.В..		28.02.20	*Технічне завдання*			

ДОДАТКИ

Додаток А

```
import weka.core.jvm as jvm
import weka.core.serialization as serialization
from weka.classifiers import Classifier
from weka.core.converters import Loader
from weka.core.classes import Random
from sklearn.metrics import accuracy_score
jvm.start()

loader = Loader(classname="weka.core.converters.ArffLoader")
data = loader.load_file("../all.arff")
data.class_is_last()
train, test = data.train_test_split(70.0, Random(1))

J48 = Classifier(jobject = serialization.read("../J48.model"))
IBk = Classifier(jobject = serialization.read("../IBk.model"))
Logistic = Classifier(jobject = serialization.read("../Logistic.model"))
pred_J48 = []
pred_IBk = []
pred_Logistic = []

actual = []
for index, inst in enumerate(test):
  pred = J48.classify_instance(inst)
  actual.append(inst.get_string_value(inst.class_index))
  pred_J48.append(inst.class_attribute.value(int(pred)))
for index, inst in enumerate(test):
 pred = IBk.classify_instance(inst)
 pred_IBk.append(inst.class_attribute.value(int(pred)))
     for index, inst in enumerate(test):
       pred = Logistic.classify_instance(inst)
       pred_Logistic.append(inst.class_attribute.value(int(pred)))

MultiClass =[]
  for i in range(0,len(test)):
    if pred_Logistic[i] == pred_IBk[i]:
    MultiClass.append(pred_Logistic[i])
  else:
    MultiClass.append(pred_J48[i])

print(accuracy_score(actual,MultiClass))
```