ДЕРЖАВНИЙ УНІВЕРСИТЕТ ІНФОРМАЦІЙНО-КОМУНІКАЦІЙНИХ ТЕХНОЛОГІЙ
НАВЧАЛЬНО-НАУКОВИЙ ІНСТИТУТ КІБЕРБЕЗПЕКИ ТА ЗАХИСТУ ІНФОРМАЦІЇ

КАФЕДРА СИСТЕМ ТА ТЕХНОЛОГІЙ КІБЕРБЕЗПЕКИ

КВАЛІФІКАЦІЙНА РОБОТА

на тему:

«Технологія протидії методам соціальної інженерії з використанням алгоритмів штучного інтелекту»

зі спеціальності *125 Кібербезпека та захист інформації*
(код, найменування спеціальності)

освітньо-професійної програми *Інформаційна та кібернетична безпека*
(назва програми)

Кваліфікаційна робота містить результати власних досліджень. Використання ідей, результатів і текстів інших авторів мають посилання на відповідне джерело

_____ Андрій ГУЛЯЄВ
(підпис)

Виконав: здобувач вищої освіти групи БСДМ-62
ГУЛЯЄВ Андрій
(прізвище, ім'я)

Керівник д-р техн. наук, професор КАЗМІРЧУК Світлана
(науковий ступінь, вчене звання, прізвище, ім'я)

Рецензент _____
(науковий ступінь, вчене звання, прізвище, ім'я)

Київ 2025

ЗМІСТ

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ

AI	– Artificial Intelligence
BERT	– Bidirectional Encoder Representations from Transformers
CLS	– Classification Token
DistilBERT	– Distilled BERT
GloVe	– Global Vectors for Word Representation
GPT	– Generative Pre-trained Transformer
GRU	– Gated Recurrent Unit
LSTM	– Long Short-Term Memory
ML	– Machine Learning
NLP	– Natural Language Processing
OSINT	– Open Source Intelligence
RoBERTa	– Robustly Optimized BERT Approach
SEP	– Separator Token
SVM	– Support Vector Machine
TF-IDF	– Term Frequency – Inverse Document Frequency
ШІ	– Штучний інтелект

ВСТУП

Актуальність дослідження. У сучасних умовах стрімкої цифровізації суспільства кількість атак соціальної інженерії невпинно зростає. Зловмисники, як і раніше, використовують їх для отримання доступу до конфіденційної інформації громадян. Використання смартфонів, месенджерів, електронної пошти та соціальних мереж створило середовище, у якому людина постійно взаємодіє з великою кількістю потенційних загроз. У цій ситуації людський фактор залишається одним з найуразливіших елементів системи кібербезпеки.

Традиційні методи протидії соціальній інженерії вже не забезпечують достатнього рівня захисту. Соціальні інженери швидко адаптуються, використовують нові схеми шахрайства та технології, покращують якість своїх атак. У таких умовах стає очевидним, що класичні інструменти не здатні виявляти всі форми маніпулятивних впливів на користувача.

На тлі розвитку штучного інтелекту з’являються нові можливості для виявлення атак соціальної інженерії. Алгоритми штучного інтелекту здатні аналізувати великі обсяги даних, виявляти приховані патерни, визначати ознаки маніпуляцій та класифікувати повідомлення за рівнем ризику.

У контексті зростаючої кількості шахрайських схем, що націлені саме на пересічних громадян, питання захисту населення від соціальної інженерії набуває особливої важливості. Більшість користувачів не мають глибоких технічних знань, і саме тому інтелектуальні системи автоматичного виявлення та попередження атак можуть стати ключовим елементом загальної стратегії кіберзахисту. Використання алгоритмів ШІ відкриває можливість створення персоналізованих, адаптивних і масштабованих інструментів попередження соціально-інженерних загроз.

Вищезазначене підтверджує актуальність даного дослідження, яке присвячене аналізу алгоритмів штучного інтелекту та розробленню рекомендацій щодо їх використання для підвищення кіберзахищеності громадян від атак

соціальної інженерії.

Об'єкт дослідження – процеси забезпечення інформаційної безпеки та захисту користувачів від атак соціальної інженерії у цифровому середовищі.

Предмет дослідження – моделі та алгоритми штучного інтелекту для виявлення, аналізу та запобігання атакам соціальної інженерії.

Мета роботи – дослідити алгоритми штучного інтелекту та розробити рекомендації щодо використання їх можливостей для підвищення кіберзахищеності громадян від атак соціальної інженерії.

Наукові завдання:

дослідити проблематику атак соціальної інженерії у цифровому середовищі;

проаналізувати види та механізми реалізації атак соціальної інженерії;

дослідити сучасні підходи та інструменти протидії атакам соціальної інженерії;

проаналізувати алгоритми штучного інтелекту, що застосовуються для виявлення та попередження атак соціальної інженерії;

оцінити ефективність використання алгоритмів ШІ у підвищенні кіберзахищеності громадян;

розробити рекомендації щодо застосування можливостей штучного інтелекту для підсилення захисту від соціально-інженерних атак.

Методи дослідження – опрацювання літератури за даною темою, аналіз експлуатаційної документації, міжнародних стандартів та їх порівняння.

Практичне значення одержаних результатів: розроблено рекомендації щодо використання алгоритмів штучного інтелекту для виявлення та запобігання атакам соціальної інженерії, а також визначено оптимальний підхід до застосування трансформерних моделей для підвищення кіберзахищеності громадян.

Результати кваліфікаційної роботи апробовані на Всеукраїнській науковій конференції «Актуальні проблеми кібербезпеки», яка відбулася 29 жовтня 2025 року в Державному університеті інформаційно-комунікаційних технологій, м. Київ.

1 ОГЛЯД МЕТОДІВ СОЦІАЛЬНОЇ ІНЖЕНЕРІЇ ТА ПІДХОДІВ ДО ЇХ ВИЯВЛЕННЯ

1.1. Сутність соціальної інженерії та роль людського фактора у кібербезпеці

Однією з ключових проблем сучасної інформаційної безпеки є так званий “людський фактор” — комплекс поведінкових та психологічних характеристик користувачів, що створюють вразливості в цифровому середовищі. Навіть за умови застосування сучасних технічних засобів захисту, недостатній рівень знань, необачність або недбалість користувачів можуть призводити до успішних атак, оскільки саме люди часто стають мішенню хакерів [1].

Як видно з таблиці 1.1, незалежно від рівня дослідження — національного чи міжнародного — людський фактор залишається домінуючою причиною кіберінцидентів. За даними аналітичних звітів Statista [4], IS Partners LLC [6], Verizon Data Breach Investigations Report [5], CERT-UA [3], Державної служби спеціального зв’язку та захисту інформації (ДССЗЗІ) України, Kaspersky Lab [7] частка атак, успішність яких зумовлена помилками користувачів або впливом методів соціальної інженерії, стабільно перевищує 60–70 %, а в окремих випадках сягає понад 80 %.

Таблиця 1.1.

Статистичні дані щодо ролі людського фактора у кіберінцидентах

Джерело	*Рік*	*Рівень дослідження*	*Частка інцидентів, пов’язаних з людським фактором*	*Характеристика*
Statista	2024	Світовий	близько 68 %	Соціальна інженерія,

Джерело	*Рік*	*Рівень дослідження*	*Частка інцидентів, пов'язаних з людським фактором*	*Характеристика*
				помилки користувачів, необережні дії
IS Partners LLC	2023	Міжнародний	68–74 %	Фішинг, неправильні дії користувачів, слабка обізнаність
Kaspersky Lab	2022	Світовий	до 80–95 %	Людський фактор як основна вразливість кібербезпеки
Verizon DBIR	2024	Світовий	близько 74 %	Соціальна інженерія та помилки людей як ключові причини
CERT-UA, ДССЗЗІ України	2023 –2024	Національний	понад 80 %	Маніпулятивні методи та поведінкові ризики

Аналіз наведених статистичних даних свідчить не лише про масштаб проблеми людського фактора, а й про її системний характер. Висока частка інцидентів, пов'язаних із соціальною інженерією, зберігається незалежно від рівня цифровізації суспільства, рівня технічного оснащення організацій чи регіональної приналежності. Це підтверджує тезу про те, що людський фактор є універсальною вразливістю, яку неможливо усунути виключно технічними засобами захисту.

Більше того, із зростанням рівня автоматизації та ускладненням технічних

систем роль людини як потенційного вектора атаки лише посилюється. Користувачі все частіше взаємодіють з цифровими сервісами, довіряють автоматизованим повідомленням та звикають до постійного інформаційного потоку, що знижує рівень критичного мислення та підвищує ефективність маніпулятивних впливів.

Таким чином, людський фактор становить собою не тільки об’єкт дослідження, а й важливу складову загальної системи кібербезпеки, оскільки саме поведінка користувачів часто визначає успішною чи провальною буде кібератака. Це зумовлює необхідність впровадження адаптованих програм навчання, інтелектуальних технологій виявлення маніпулятивного контенту, зокрема на основі алгоритмів штучного інтелекту, підвищення цифрової грамотності та формування стійких моделей поведінки для мінімізації людських помилок і ефективної протидії користувачів загрозам соціальної інженерії.

Сутність соціальної інженерії полягає у використанні знань про людську психологію, поведінку та довіру для обходу технічних засобів захисту. На відміну від хакерських атак, спрямованих на вразливості програмного забезпечення ПК, соціальна інженерія маніпулює людською психологією.

Соціальна інженерія охоплює широкий спектр методів і технік, які можуть реалізовуватися як у цифровому, так і в офлайн-середовищі. До класичних форм соціальної інженерії належать фішинг, spear phishing, vishing (голосовий фішинг), smishing (SMS-фішинг), pretexting (створення правдоподібного сценарію), baiting (приманювання), а також атаки типу business email compromise (BEC).

Спільною рисою цих методів є використання психологічного впливу з метою зниження рівня пильності жертви та стимулювання прийняття швидких, емоційних рішень. У багатьох випадках технічна складова атаки є мінімальною або взагалі відсутньою, що дозволяє зловмисникам обходити традиційні засоби захисту, такі як міжмережеві екрани чи антивірусні системи.

Зловмисники часто апелюють до таких психологічних чинників, як:

- довіра: представляються як авторитетні або знайомі особи (колеги, представники установ);

- терміновість/страх: створюють відчуття невідкладності, щоб жертва діяла швидко і необачно;
- цікавість/жадібність: пропонують “вигідні” або інтригуючі можливості, що провокують емоційні рішення, а не раціональні;
- бажання допомогти: експлуатують природне прагнення людей бути корисними та допомагати іншим, що знижує пильність

Завдяки такому підходу, соціальні інженери “зламують” не систему, а людину, роблячи її слабким місцем навіть у добре захищеній IT-інфраструктурі.

Окрім зазначених чинників, у практиці соціальної інженерії активно використовуються також:

- ефект авторитету — посилання на керівників, державні органи або відомі бренди з метою підвищення довіри;
- соціальне підтвердження — створення ілюзії, що подібну дію вже виконали інші користувачі;
- когнітивне перевантаження — навмисне подання надмірної або складної інформації, що знижує здатність до критичного аналізу;
- звичка до автоматизованих дій — експлуатація рутинної поведінки користувачів, які діють за шаблоном, не замислюючись над безпекою.

Використання цих механізмів дозволяє зловмисникам адаптувати атаки до конкретної аудиторії та значно підвищувати їхню результативність.

У сучасних умовах особливої актуальності набуває вплив технологій штучного інтелекту на соціальну інженерію. З одного боку, ШІ значно розширює можливості зловмисників: використання генеративних моделей для створення високоякісних фішингових листів, аудіо та відео, автоматизованого аналізу відкритих даних про потенційні цілі, імітації стилю письма чи голосу конкретної людини. Це робить атаки більш переконливими та складними для виявлення користувачами [2].

Слід також зазначити, що соціальна інженерія є динамічним явищем, яке постійно еволюціонує разом із розвитком інформаційних технологій. Якщо раніше більшість атак мали масовий характер, то сьогодні спостерігається зростання

кількості цілеспрямованих, персоналізованих атак, орієнтованих на конкретних осіб або групи користувачів.

Така трансформація зумовлена доступністю великих обсягів відкритих даних у соціальних мережах, професійних платформах та інших публічних ресурсах, що значно спрощує процес збору інформації про потенційні жертви.

З іншого боку, ШІ стає потужним інструментом захисту: системи машинного навчання дозволяють автоматично виявляти аномальні дії, небезпечні поведінкові патерни та ознаки соціальної інженерії; моделі аналізують великі обсяги взаємодій, підвищують точність фільтрації фішингового контенту, а також забезпечують персоналізовані підходи до навчання користувачів залежно від їхнього рівня ризику [2].

Отже, взаємодія людського фактора, соціальної інженерії та алгоритмів штучного інтелекту формує новий контекст кібербезпеки, в якому ефективний захист неможливий без комплексного підходу — одночасного розвитку захисних технологій на основі ШІ та посилення обізнаності користувачів щодо сучасних кіберзагроз [2].

Саме тому подальші дослідження у сфері протидії соціальній інженерії дедалі більше зосереджуються на поєднанні технологічних інструментів штучного інтелекту з аналізом поведінкових моделей людини. Такий міждисциплінарний підхід дозволяє не лише виявляти атаки, а й прогнозувати ризики, формувати адаптивні механізми захисту та підвищувати загальний рівень кіберстійкості суспільства.

1.2. Класифікація методів соціальної інженерії

Соціальна інженерія охоплює сукупність методів психологічного впливу, спрямованих на маніпулювання поведінкою користувачів з метою отримання несанкціонованого доступу до інформаційних ресурсів. На відміну від технічних кіберзагроз, ці атаки ґрунтуються на експлуатації когнітивних упереджень, емоційних реакцій та довіри людини. Дослідження українських та зарубіжних

авторів підтверджують, що понад 70 % успішних кібератак здійснюються саме через людський фактор.

Важливо зазначити, що методи соціальної інженерії не є статичними та змінюються відповідно до розвитку інформаційних технологій і соціальних практик користувачів. Якщо на початкових етапах домінували прості сценарії масових атак, то сьогодні спостерігається перехід до багаторівневих, адаптивних і комбінованих впливів, які поєднують психологічні, комунікаційні та технічні елементи.

У сучасних дослідженнях соціальну інженерію дедалі частіше розглядають як соціо-технічну загрозу, де успіх атаки визначається не лише майстерністю маніпулятора, а й контекстом взаємодії, інформаційним фоном та поведінковими характеристиками жертви.

Основні методи соціальної інженерії (таблиця 1.2) включають різноманітні механізми вербального, емоційного або інформаційного впливу на жертву. Вони реалізуються переважно через електронні комунікації, телефонні дзвінки або фізичний доступ, використовуючи такі психологічні тригери, як терміновість, страх, цікавість, жадібність, довіра до авторитетних постатей та прагнення допомогти.

Таблиця 1.2.

Класифікація методів атак соціальної інженерії

Група методів	*Метод*	*Опис*
Психологічні та поведінкові методи	Маніпуляції довірою	Використання авторитету або удаваної легітимності (керівник, державна установа, служба підтримки) для спонукання користувача до небезпечних дій.
	Індукція терміновості або страху	Створення штучного інформаційного тиску з метою пришвидшення реакції жертви без критичного аналізу повідомлення.

Група методів	*Метод*	*Опис*
	Експлуатація когнітивної цікавості	Залучення уваги шляхом подання потенційно цінної, приватної або сенсаційної інформації, що стимулює відкриття файлів чи переходи за посиланнями.
	Апеляція до альтруїстичних мотивів	Маніпулювання природним бажанням допомогти, надання удаваних запитів щодо підтримки або сприяння.
Комунікаційні атаки	Фішинг (Phishing)	Масове або цільове надсилання електронних листів чи повідомлень, що імітують легітимні сервіси, з метою викрадення облікових даних або розповсюдження шкідливого ПЗ.
	Смішинг (Smishing)	Надсилання шахрайських SMS, спрямованих на отримання конфіденційної інформації або спонукання до переходу на фальшиві ресурси.
	Вішинг (Vishing)	Телефонні дзвінки з використанням соціальних сценаріїв, які імітують співробітників банків, держустанов, технічної підтримки тощо.
	Соціальна інженерія через соціальні мережі	Створення фіктивних профілів, встановлення довірливих контактів, використання месенджерів для проведення маніпулятивних дій.

Група методів	*Метод*	*Опис*
Технічні та комбіновані методи	Спірфішинг (Spear phishing)	Персоналізовані атаки, побудовані на основі даних із відкритих джерел (OSINT), адресовані конкретним особам або організаціям.
	Вейлінг (Whaling)	Цільові атаки на керівників, власників бізнесу або осіб із підвищеними повноваженнями доступу.
	Претекстинг (Pretexting)	Створення детально продуманої легенди (претексту) з метою викликати довіру й отримати доступ до інформації або ресурсів.
	Бейтинг (Baiting)	Застосування фізичних або цифрових принад (носії, QR-коди, вебпосилання, подарунки), що спонукають жертву до взаємодії зі шкідливими об’єктами.
	Quid pro quo	Обмін удаваної вигоди (технічна допомога, бонуси, послуги) на отримання конфіденційних даних або облікових записів.
Фізичні методи соціальної інженерії	Tailgating / Piggybacking	Несанкціоноване потрапляння у приміщення шляхом проходження за співробітником, що має законний доступ.
	Dumpster Diving	Отримання інформації шляхом пошуку документів, носіїв або нотаток, що були неналежним чином утилізовані.

Група методів	*Метод*	*Опис*
	Shoulder Surfing	Візуальне спостереження за введенням паролів, PIN-кодів та іншої конфіденційної інформації.
	Імперсонація фізичного працівника	Імітація ролей технічних спеціалістів, кур'єрів, охоронців чи інженерів для доступу до об'єктів.
Сучасні, AI-орієнтовані та багатоканальні методи	Deepfake-атаки	Використання технологій генеративного ШІ для створення аудіо-, відео- або графічних матеріалів, що імітують реальних осіб.
	AI-phishing / Adaptive phishing	Атаки, у яких ШІ формує персоналізовані повідомлення, адаптує стилістику комунікації та аналізує соціально-поведінкові дані жертви.
	Атаки в корпоративних чатах	Маніпулятивні взаємодії через Slack, Microsoft Teams, Telegram або внутрішні комунікаційні платформи.
	Багатоканальні атаки (Multichannel attacks)	Скоординоване використання кількох каналів зв'язку (дзвінок + лист + фальшивий сайт + deepfake) для підвищення правдоподібності атаки.
Пасивні методи збору інформації (OSINT)	Аналіз відкритих джерел	Вивчення соціальних мереж, корпоративних сайтів, витоків даних, публічних документів з метою підготовки цільових атак.

Наведена класифікація демонструє, що методи соціальної інженерії охоплюють широкий спектр впливів — від простих психологічних маніпуляцій до складних багатоканальних атак із використанням штучного інтелекту. При цьому між окремими групами методів не існує чітких меж: у реальних сценаріях зловмисники часто комбінують кілька технік одночасно, підвищуючи ймовірність успіху.

Така багатовимірність атак ускладнює їх формалізацію та автоматичне виявлення за допомогою традиційних правил або сигнатур, що обумовлює необхідність використання інтелектуальних підходів до аналізу контенту та поведінки.

Широке різноманіття атак соціальної інженерії зумовлює потребу у комплексному аналізі їх характеристик на основі сучасних наукових досліджень і практичних кейсів, оскільки саме ці методи становлять ключовий вектор первинного проникнення в інформаційні системи, що підтверджується статистичними звітами провідних експертних центрів [3].

Психологічні методи ґрунтуються на використанні природних когнітивних особливостей людини: довіри, реакції на терміновість, інстинктивного бажання допомогти або цікавості. Ці вектори атак найчастіше застосовуються на ранніх етапах компрометації, оскільки не потребують складних технічних засобів, але мають високу ефективність.

Комунікаційні атаки демонструють тенденцію до постійного зростання: фішинг, смішинг та вішинг становлять основний інструментарій зловмисників, а їхня якість та персоналізація значно підвищилися у зв’язку з використанням генеративних моделей штучного інтелекту. Зокрема, за даними Національного банку України (2023), майже 70% успішних шахрайських транзакцій у фінансовому секторі починалися з фішингового контакту.

Технічні та комбіновані методи, такі як спірфішинг, претекстинг або baiting, особливо небезпечні для організацій, оскільки дозволяють виконувати високоточні та таргетовані атаки на ключових осіб (керівників, адміністраторів, бухгалтерію). Використання OSINT для попередньої розвідки забезпечує зловмисникам високий

рівень достовірності легенд і підвищує ймовірність успіху атаки.

Окрему категорію становлять сучасні методи, що використовують ШІ. Deepfake-атаки, AI-phishing та адаптивні діалоги на основі великих мовних моделей суттєво ускладнюють можливість самостійного виявлення маніпуляцій користувачами. Генеративні моделі дозволяють створювати стилістично унікальні електронні листи, реалістичні аудіоповідомлення та персоналізовані сценарії взаємодії, що робить такі атаки особливо загрозливими.

Фізичні методи соціальної інженерії, хоча менш цифрові за природою, продовжують залишатися актуальними для організацій із недостатнім рівнем контролю доступу. Практики tailgating, dumpster diving або shoulder surfing нерідко виступають початковим етапом для подальших технічних атак.

OSINT-методи є базовим підготовчим етапом перед будь-яким із зазначених типів атак. Збір інформації з відкритих джерел забезпечує зловмисників необхідними деталями для створення високоякісних та переконливих легенд.

Таким чином, сучасні методи соціальної інженерії демонструють тенденцію до зростання складності, персоналізації та використання ШІ-технологій, що потребує комплексного підходу до кіберзахисту: від технічних засобів до систематичного навчання користувачів, вдосконалення політик безпеки та застосування інструментів поведінкового аналізу.

1.3. Ознаки та мовні патерни атак, що дозволяють їх виявляти

Атаки соціальної інженерії мають низку характерних ознак, які відрізняють їх від легітимної комунікації та дозволяють виявити загрозу на ранніх етапах. Основною особливістю таких атак є орієнтація не на технічні вразливості інформаційних систем, а на психологічні, поведінкові та когнітивні особливості людини.

Сучасні атаки соціальної інженерії характеризуються зростаючим рівнем складності та адаптивності, що зумовлено активним використанням відкритих джерел інформації (OSINT), автоматизації та технологій штучного інтелекту.

Зловмисники все частіше поєднують психологічні методи впливу з технічними прийомами маскування, що ускладнює відмежування шкідливої комунікації від легітимної та підвищує ризик успішної атаки.

Нижче наведені ключові ознаки, які дозволять вам зрозуміти, чи є ви мішенню атаки соціальної інженерії.

1) *Створення відчуття терміновості або кризи.* Навмисне формування ситуації штучної нагальності (наприклад, «обліковий запис буде заблоковано», «потрібно негайно підтвердити дані»), зменшуючи час на критичне осмислення інформації та стимулюючи імпульсивні рішення [8].

2) *Апеляція до авторитету або ієрархії.* Видавання себе за керівника, представника державного органу, служби безпеки, банку або технічної підтримки. Використання авторитету підвищує ймовірність автоматичної довіри з боку жертви [8].

3) *Запит конфіденційної або службової інформації.* Вимога надати паролі, коди доступу, персональні дані, фінансову інформацію або внутрішні відомості, які за стандартами інформаційної безпеки не повинні передаватися через відкриті канали зв’язку [8].

4) *Невідповідність каналу комунікації.* Звернення надходить через нетипові або неофіційні канали (особисті месенджери, приватні електронні адреси, дзвінки з невідомих номерів), що не використовуються організацією для офіційної комунікації [8].

5) *Персоналізований характер звернення.* Використання імені адресата, назви організації, посади або контексту професійної діяльності, отриманих із відкритих джерел, з метою підвищення рівня довіри. Така ознака є характерною для цільових атак (spear-phishing) і значно ускладнює їх розпізнавання [8].

6) *Підміна цифрової ідентичності (spoofing).* Використання підроблених адрес електронної пошти, доменів або номерів телефонів, які візуально майже не відрізняються від офіційних. Наявність незначних відмінностей у написанні часто залишається непоміченою користувачем [8].

7) *Соціально-психологічна маніпуляція.* Використання страху, провини,

співчуття, надмірної доброзичливості або обіцянок винагороди. Маніпуляція емоціями є ключовим інструментом соціальної інженерії [8].

8) Лінгвістичні та стилістичні аномалії. Повідомлення часто містять граматичні помилки, непритаманну офіційним установам лексику, шаблонні формулювання або автоматично згенерований текст, що може свідчити про масове або автоматизоване походження атаки. Водночас слід зазначити, що з появою генеративних мовних моделей кількість граматичних і стилістичних помилок у шкідливих повідомленнях зменшується. Це знижує ефективність виявлення атак виключно на основі формальних мовних помилок та зумовлює необхідність аналізу контексту, семантики та поведінкових патернів взаємодії. [8].

9) Заклики до порушення внутрішніх процедур. Наполягання на ігноруванні встановлених регламентів безпеки («це виняток», «немає часу погоджувати», «не повідомляйте колег»), що є прямим індикатором загрози [8].

10) Аномальна поведінка у телефонних або особистих контактах. Уникання прямих відповідей, тиск на співрозмовника, небажання проходити стандартну ідентифікацію або різка зміна теми розмови [8].

Також, зловмисники в своїх атаках часто застосовують певні мовні патерни — характерні лінгвістичні та стилістичні елементи, що дозволяють їм маніпулювати увагою та поведінкою адресата. Аналіз таких патернів є важливим інструментом виявлення потенційно небезпечних повідомлень ще до того, як вони спричинять шкоду.

Окрім наведених прикладів, сучасні атаки соціальної інженерії демонструють використання псевдотехнічної лексики, яка створює у адресата хибне відчуття компетентності та легітимності повідомлення. Такі формулювання часто апелюють до «збоїв системи», «порушення політик безпеки» або «оновлення протоколів», що підсилює психологічний тиск та стимулює негайну реакцію [9].

Основні мовні патерни атак соціальної інженерії:

1) Риторика “надмірної терміновості”, яка стимулює швидку реакцію та знижує критичне мислення. Цей патерн спонукає жертву діяти поспішно, не

аналізуючи ситуацію [9].

a) «Терміново!»;

b) «Ваш акаунт буде заблоковано через 24 години»;

c) «Негайно підтвердьте свої дані, щоб уникнути штрафу».

2) Імітація голосу керівника, співробітників організацій та посадовців. Використання офіційно звучних назв підсилює довіру співрозмовника [9].

a) «Адміністратор безпеки»;

b) «Ваш менеджер наказав…»;

c) «Український банк/державний орган».

3) Патерн “обіцянки вигоди або винагороди”, що спрацьовує на жадібність чи інтерес. Він стимулює взаємодію з посиланням чи вкладенням [9].

a) «Отримайте бонус»;

b) «Ексклюзивна пропозиція для вас»;

c) «Подарунок за реєстрацію».

4) Риторичні питання або провокативні формулювання. Такий підхід використовує емоції для залучення уваги. Ці фрази спрямовані на те, щоб «підштовхнути» на дію [9].

a) «Чому ви ще не підтвердили?»;

b) «Хіба вам не цікаво побачити?»;

c) «Не втрачайте шанс».

5) Питання без контексту або без персоналізації. Часто характерно для автоматизованих атак. Відсутність імені/персональних даних може бути ознакою фішингового повідомлення [9].

a) «Шановний користувачу»;

b) «Ваш обліковий запис»;

c) «Клікніть тут».

6) Помилки, підозрілі формулювання, нетипові для офіційних документів стилістичні конструкції. Ці аномалії часто свідчать про автоматичні або непродумані формулювання [9].

a) Граматичні/лексичні помилки в офіційному контексті;

b) Невластива офіційна лексика;

c) Стилістичні несумісності.

У таблиці 1.3. наведено мапінг прикладів мовних патернів відповідно до типів атак соціальної інженерії.

Таблиця 1.3.

Приклади мовних патернів у атаках

Патерн	*Тип атаки*	*Приклад*
Терміновість	Фішинг	«Ваш акаунт буде видалено через 24 години»
Авторитетність	Вішинг	«Зателефонував співробітник банку…»
Обіцянка вигоди	Бейтинг	«Отримайте подарунок, клікнувши посилання»
Провокативне питання	General SE	«Не хочете дізнатися секрет?»
Відсутність персоналізації	Масовий фішинг	«Шановний користувачу»
Стильові/граматичні аномалії	будь-який	Повідомлення з помилками
Персоналізація	Spear-phishing	«Олено, за вашим проєктом № 241 потрібне підтвердження»
Підміна ідентичності	Spoofing	Лист із домену, схожого на корпоративний
QR-код	Quishing	«Скануйте код для підтвердження доступу»
Синтез голосу	Вішинг	Дзвінок «від керівника» з вимогою термінового переказу

Патерн	*Тип атаки*	*Приклад*
Псевдотехнічна лексика	Фішинг	«Порушено протокол автентифікації»

З огляду на еволюцію методів соціальної інженерії, ефективне виявлення атак потребує врахування не лише явних мовних аномалій, а й контекстних, поведінкових та когнітивних характеристик комунікації. Таким чином, для ефективного виявлення та запобігання атакам соціальної інженерії важливо звертати увагу на ці індикатори. Комплексний аналіз ознак, лексичних маркерів, стилістичних аномалій та структурних елементів тексту є основою для розробки інтелектуальних систем виявлення загроз та автоматизованих технологій протидії соціальній інженерії на основі алгоритмів штучного інтелекту.

1.4. Сучасні підходи до автоматизованого виявлення атак соціальної інженерії та їх обмеження

Зростання кількості атак соціальної інженерії та ускладнення їх сценаріїв зумовлюють необхідність застосування автоматизованих засобів виявлення, заснованих на методах штучного інтелекту, машинного навчання та аналізу природної мови. На відміну від традиційних сигнатурних підходів, сучасні інтелектуальні системи орієнтовані на аналіз поведінкових, лінгвістичних та контекстних ознак повідомлень, що дозволяє підвищити ефективність протидії маніпулятивним впливам.

Діаграма на рисунку 1.1 відображає узагальнені дані щодо основних напрямів використання AI у сучасних системах кібербезпеки (платформи, сервіси, поштові системи, Security Operations Center, cloud-сервіси), включаючи виявлення аномальної поведінки, аналіз текстових загроз та автоматичне розпізнавання фішингових повідомлень.

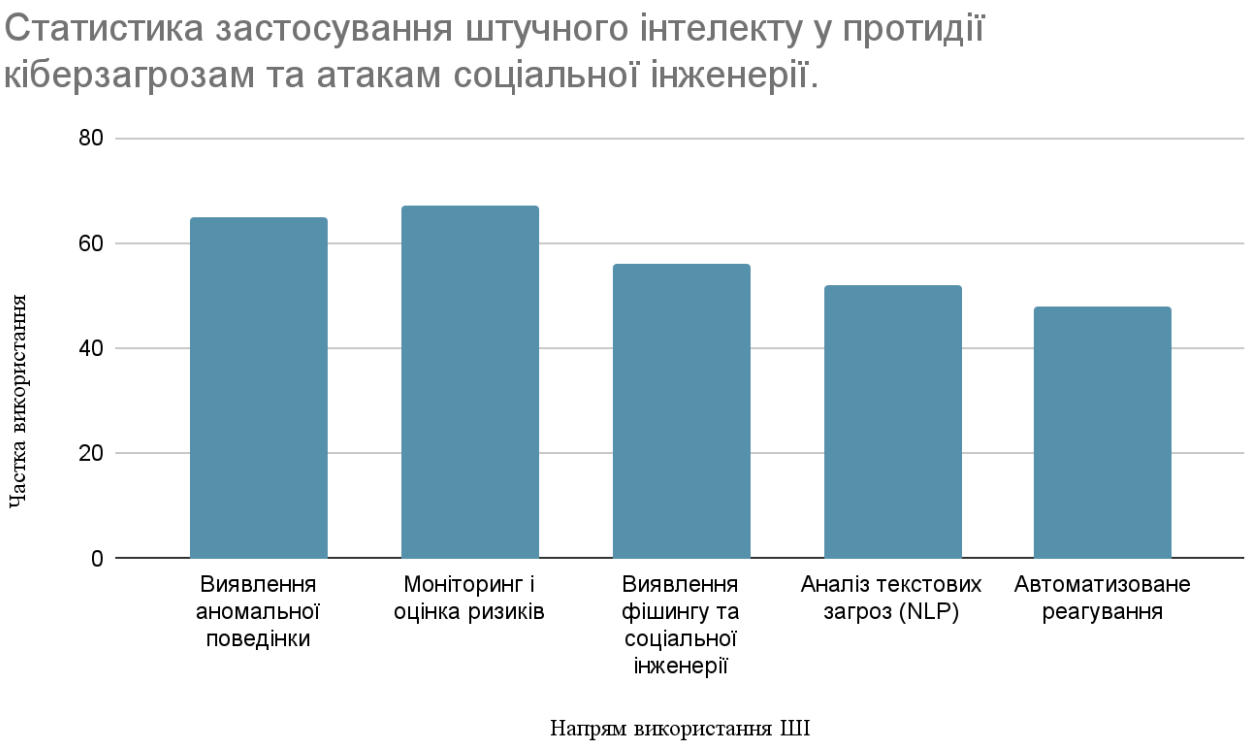

Рис. 1.1 Статистика застосування штучного інтелекту у протидії кіберзагрозам та атакам соціальної інженерії.

В таблиці 1.4. наведені основні сучасні підходи до автоматизованого виявлення атак соціальної інженерії, їх переваги та обмеження. Кожен із підходів має власну методологічну основу, сферу ефективного застосування та низку обмежень, що необхідно враховувати під час проєктування систем кіберзахисту.

Таблиця 1.4.

Підходи до автоматизованого виявлення атак соціальної інженерії

Назва підходу	Характеристика	Переваги	Обмеження
Сигнатурний аналіз	Передбачає виявлення заздалегідь відомих шаблонів, ключових слів і фраз	Простота реалізації, висока швидкодія	Неефективний проти нових або модифікованих атак

Назва підходу	Характеристика	Переваги	Обмеження
Аналіз мовних патернів (NLP-підхід)	Базується на лінгвістичних ознаках маніпуляції: терміновість, авторитет, страх, вигода	Виявлення невідомих атак, масштабованість	Залежність від якості словників і мовних моделей
Машинне навчання (Machine Learning)	Використовує класифікатори (SVM, Random Forest, Naive Bayes) для аналізу текстів	Адаптація до нових шаблонів атак	Потреба у великому обсязі навчальних даних
Глибинне навчання та трансформерні моделі	Застосовуються моделі типу BERT, RoBERTa, GPT для глибокого контекстного аналізу	Висока точність, контекстне розуміння мови	Значні обчислювальні витрати, складність інтерпретації рішень
Поведінковий аналіз	Оцінює поведінку користувачів під час взаємодії з повідомленнями	Зниження кількості хибнопозитивних спрацювань	Питання конфіденційності та складність впровадження
Гібридні підходи	Поєднують лінгвістичний аналіз, машинне	Найвища ефективність	Складність архітектури та інтеграції

Назва підходу	Характеристика	Переваги	Обмеження
	навчання та поведінкові моделі		

Сигнатурний аналіз є одним із найстаріших підходів до виявлення фішингових та маніпулятивних повідомлень. Він ґрунтується на порівнянні вхідних повідомлень із заздалегідь визначеними шаблонами, ключовими словами, регулярними виразами або відомими зразками атак. Такий підхід широко використовується у класичних поштових фільтрах та антивірусних рішеннях завдяки своїй простоті та низьким вимогам до обчислювальних ресурсів. Водночас сигнатурний аналіз є малоефективним у сучасних умовах, оскільки зловмисники постійно модифікують тексти повідомлень, змінюють формулювання та використовують нові соціальні сценарії, які не відповідають відомим сигнатурам.

Підхід на основі аналізу мовних патернів (NLP) орієнтований на виявлення характерних лінгвістичних ознак соціальної інженерії. До таких ознак належать використання термінових закликів до дії, апеляція до авторитету, створення атмосфери страху або обіцянки вигоди. На відміну від сигнатурних методів, NLP-підходи дозволяють виявляти раніше невідомі атаки за рахунок аналізу семантики та стилістики тексту. Однак ефективність таких методів значною мірою залежить від якості словників, мовних правил і моделей, а також від мови повідомлень, що ускладнює їх масштабування у багатомовних середовищах.

Методи машинного навчання (Machine Learning) передбачають використання статистичних класифікаторів, таких як Naive Bayes, Support Vector Machines, Random Forest та інших алгоритмів, які навчаються на розмічених наборах даних. Такі моделі здатні автоматично виявляти складні залежності між ознаками тексту та класами повідомлень, що забезпечує кращу адаптацію до нових шаблонів атак. Водночас традиційні ML-моделі потребують ретельного етапу формування ознак (feature engineering) та великого обсягу якісних

навчальних даних, що може обмежувати їх застосування в умовах швидкої еволюції загроз.

Глибинне навчання та трансформерні моделі є сучасним етапом розвитку інтелектуальних систем виявлення соціальної інженерії. Моделі типу BERT, RoBERTa або GPT забезпечують глибокий контекстний аналіз тексту завдяки механізму самоуваги, що дозволяє враховувати взаємозв'язки між словами у межах усього повідомлення. Це особливо важливо для фішингових атак, де маніпулятивний зміст часто розподілений по всьому тексту. Основними обмеженнями таких підходів є високі вимоги до обчислювальних ресурсів, складність навчання та обмежена інтерпретованість рішень моделі.

Поведінковий аналіз зосереджується не лише на тексті повідомлення, а й на реакціях користувача під час взаємодії з ним. До таких параметрів можуть належати швидкість переходу за посиланням, частота взаємодій, типи виконаних дій або відхилення від звичної поведінки. Поєднання текстового аналізу з поведінковими характеристиками дозволяє зменшити кількість хибнопозитивних спрацювань. Проте використання таких методів пов'язане з етичними аспектами, питаннями конфіденційності та складністю інтеграції у реальні інформаційні системи.

Гібридні підходи поєднують декілька методів одночасно, зокрема сигнатурний аналіз, машинне навчання, глибинні нейронні мережі та поведінкові моделі. Такий підхід дозволяє компенсувати недоліки окремих методів і досягати максимальної точності виявлення атак соціальної інженерії. Водночас складність архітектури гібридних систем, необхідність синхронізації різних модулів і зростання витрат на підтримку залишаються суттєвими викликами для їх впровадження [24].

Аналіз наведених підходів свідчить про те, що кожен з них має власну сферу ефективного застосування, однак жоден не забезпечує повного захисту від атак соціальної інженерії у відриві від інших методів. Сигнатурні та правило-орієнтовані системи демонструють високу швидкодію, проте є малоефективними в умовах появи нових або модифікованих сценаріїв атак.

Методи машинного та глибинного навчання, навпаки, забезпечують кращу здатність до узагальнення та адаптації, однак потребують значних обчислювальних ресурсів і якісних навчальних даних. У зв’язку з цим у сучасних системах кібербезпеки все частіше застосовуються гібридні підходи, які поєднують переваги кількох методів та знижують їх індивідуальні обмеження.

Таким чином, сучасні підходи до автоматизованого виявлення атак соціальної інженерії базуються на комплексному використанні методів аналізу тексту, машинного навчання та штучного інтелекту. Однак жоден з підходів не є універсальним, що зумовлює необхідність застосування комбінованих моделей та подальших досліджень у напрямі підвищення стійкості таких систем до адаптивних атак.

Висновки з розділу 1

Було здійснено огляд методів соціальної інженерії та підходів до їх виявлення в умовах сучасного цифрового середовища. Розкрито сутність соціальної інженерії як комплексу маніпулятивних впливів, спрямованих на експлуатацію психологічних, поведінкових та когнітивних особливостей людини з метою обходу технічних засобів захисту.

Проаналізовано основні методи атак соціальної інженерії, зокрема психологічні, комунікаційні, технічні, фізичні та комплексні сучасні атаки, включаючи використання штучного інтелекту та deepfake-технологій. Показано, що поєднання декількох каналів взаємодії та високий рівень персоналізації атак істотно ускладнюють їх своєчасне виявлення традиційними засобами кібербезпеки.

Окрему увагу приділено ознакам та мовним патернам атак соціальної інженерії, які можуть бути використані як індикатори компрометації. Встановлено, що лінгвістичні характеристики повідомлень, такі як створення терміновості, апеляція до авторитету, використання страху або вигоди, є важливою основою для побудови автоматизованих систем виявлення загроз.

Розглянуто сучасні підходи до автоматизованого виявлення атак соціальної інженерії, зокрема сигнатурні методи, аналіз мовних патернів, алгоритми машинного навчання, глибинного навчання та трансформерні моделі штучного інтелекту. Визначено їх переваги та обмеження, що підтверджує доцільність застосування контекстно-орієнтованих моделей типу BERT для підвищення точності виявлення атак.

Отримані результати обґрунтовують необхідність подальшого аналізу та порівняльної оцінки ефективності алгоритмів штучного інтелекту для виявлення атак соціальної інженерії, що і є предметом дослідження наступного розділу кваліфікаційної роботи.

2 АНАЛІЗ ТА ОЦІНКА АЛГОРИТМІВ ШТУЧНОГО ІНТЕЛЕКТУ ДЛЯ ВИЯВЛЕННЯ АТАК СОЦІАЛЬНОЇ ІНЖЕНЕРІЇ

2.1. Типи даних для навчання моделей та вимоги до їх структури

Дані є ключовим ресурсом для побудови ефективних моделей штучного інтелекту, зокрема при розв’язанні задач виявлення атак соціальної інженерії. Якість, структура та репрезентативність навчальних даних безпосередньо впливають на точність, стійкість і узагальнювальну здатність моделей.

У задачах виявлення атак соціальної інженерії дані мають додаткову складність, оскільки поєднують технічні, мовні та поведінкові характеристики. Навчальні набори часто формуються з гетерогенних джерел (електронна пошта, журнали подій, записи дзвінків, повідомлення в месенджерах), що потребує уніфікованого підходу до структурування, нормалізації та анотації даних перед їх використанням у моделях штучного інтелекту.

Дані для навчання ШІ можна класифікувати за їхньою природою та способом подання:

1) *Структуровані дані.* Це найбільш організований тип даних, який легко зберігається у традиційних реляційних базах даних (наприклад, MySQL), електронних таблицях (CSV, Excel) або фреймах даних [11].

Наприклад: табличні дані - записи клієнтів (ім'я, вік, дохід, адреса), фінансові транзакції, дані сенсорів із мітками часу.

Вимоги до структури: чітко визначені стовпці (ознаки/фічі) та рядки (зразки/приклади). Повинні бути уніфіковані типи даних у стовпцях (числові, категоріальні, булеві).

У контексті соціальної інженерії структуровані дані часто використовуються для аналізу метаданих комунікації, зокрема часу відправлення повідомлення, частоти звернень, IP-адрес, геолокаційних ознак, тривалості дзвінків та історії входів у систему. Такі параметри дозволяють виявляти аномальну поведінку, характерну для автоматизованих або зловмисних дій.

2) Неструктуровані дані. Це дані, які не мають попередньо визначеної моделі чи організації. Вони становлять більшість даних у світі, і для їхньої обробки потрібні складніші моделі (наприклад, глибинне навчання) [11].

Наприклад: текст (електронні листи, статті, відгуки, твіти, юридичні документи); зображення (фотографії, рентгенівські знімки, супутникові знімки); відео (записи з камер спостереження, фільми); аудіо (голосові записи, музика).

Вимоги до структури: потребують попередньої обробки для перетворення у числовий формат (тензори, вектори) перш ніж модель зможе їх "зрозуміти".

Для задач виявлення атак соціальної інженерії неструктуровані текстові дані є найбільш інформативними, оскільки саме в тексті містяться лінгвістичні патерни, семантичні маркери та маніпулятивні конструкції. Аналіз таких даних передбачає етапи очищення тексту, токенізації, нормалізації, видалення стоп-слів та перетворення у векторні представлення з урахуванням контексту.

3) Напівструктуровані дані. Цей тип даних не вписується у жорстку табличну схему, але містить теги або інші елементи для організації та ієрархії [11].

Наприклад: JSON, XML, NoSQL бази даних.

Вимоги до структури: дані мають бути послідовними та мати чітку ієрархію ключ-значення або тегів.

4) *Мультимодальні дані.* Це набори даних, що поєднують кілька типів інформації одночасно (текст, аудіо, метадані, зображення).

Наприклад: текст електронного листа разом із заголовками, вкладеннями та часовими характеристиками або голосовий запис дзвінка разом із транскрипцією.

Вимоги до структури: синхронізація різних модальностей, узгоджені ідентифікатори зразків та часові мітки, що дозволяють коректно поєднувати дані в єдиному навчальному прикладі.

Сучасні джерела даних свідчать про те, що переважна частина інформації в цифровому середовищі є неструктурованою: приблизно 80–85 % усіх даних не мають жорсткої табличної структури, що відображає швидке зростання обсягів текстів, зображень, відео, аудіо та журналів подій у організаціях і суспільстві

загалом. Структуровані дані становлять лише близько 20–25 % світового обсягу даних [10].

Незалежно від типу, для ефективного навчання ШІ дані повинні відповідати певним структурним та якісним вимогам.

1) Якість та чистота даних:
   a) відсутність пропусків (Imputation): усі необхідні поля мають бути заповнені, або пропуски мають бути оброблені (наприклад, заповнені середнім значенням);
   b) консистентність (Consistency): дані мають бути послідовними. Наприклад, країна не повинна бути написана як "США", "U.S.A." та "Сполучені Штати" в різних записах.
   c) відсутність викидів (Outliers): аномальні або помилкові значення (наприклад, вік 200 років) повинні бути виявлені та виправлені/видалені.
   d) релевантність: дані мають бути актуальними та стосуватися завдання моделі.

Окрему увагу слід приділяти дисбалансу класів, що є типовою проблемою у задачах виявлення атак соціальної інженерії, де кількість легітимних повідомлень значно перевищує кількість шкідливих. Для коректного навчання моделей можуть застосовуватися методи балансування даних, зокрема oversampling, undersampling або використання ваг класів.

2) Формат та подання (тензори). Моделі машинного навчання, особливо глибокого навчання, працюють з тензорами — багатовимірними масивами чисел.
   a) табличні дані: перетворюються у 2D тензор (Матриця: $зразок \times ознака$).
   b) зображення: кожне зображення зазвичай перетворюється у 3D тензор ($висота \times ширина \times канали\ кольору$).
   c) текст: Слова або символи кодуються за допомогою різних методів (наприклад, One-Hot Encoding, Word Embeddings) і перетворюються у 2D або 3D тензори.

3) Розподіл (Train, Validation, Test Sets). Набір даних має бути розділений на три частини для коректної оцінки моделі:

a) навчальний набір (Training Set): найбільша частина (наприклад, 70-80%), використовується для навчання моделі;
b) валідаційний набір (Validation Set): менша частина (наприклад, 10-20%), використовується для налаштування гіперпараметрів моделі під час навчання;
c) тестовий набір (Test Set): Використовується один раз наприкінці для фінальної, незалежної оцінки продуктивності моделі. Це має бути репрезентативна вибірка реальних даних.

4) Мітки (Labels). Для навчання з учителем (Supervised Learning) кожен приклад у навчальному наборі повинен мати відповідну мітку (label), яка є "правильною відповіддю".

a) класифікація: мітка — це клас (наприклад, "кішка", "собака", "спам", "не спам");
b) регресія: мітка — це числове значення (наприклад, ціна будинку, температура, час затримки);
c) вимоги: мітки мають бути точними, послідовними та повними. Помилкові мітки можуть зруйнувати процес навчання.

У задачах аналізу соціальної інженерії процес маркування часто вимагає участі експертів з інформаційної безпеки або лінгвістів, оскільки межа між легітимною та маніпулятивною комунікацією може бути нечіткою. Наявність неоднозначних прикладів зумовлює необхідність використання багатоетапної або багаторівневої анотації.

Врахування специфіки типів даних, їхньої структури та походження є критично важливим для побудови надійних моделей, здатних працювати в умовах реального інформаційного середовища.

Таким чином, ефективність застосування алгоритмів штучного інтелекту для виявлення атак соціальної інженерії значною мірою залежить від якості, структури та репрезентативності навчальних даних. Особливе значення мають

неструктуровані текстові дані, обробка яких вимагає використання сучасних методів подання інформації, зокрема контекстних embedding трансформерних моделей. Дотримання вимог до структури, чистоти та маркування даних створює необхідне підґрунтя для коректного навчання та об’єктивної оцінки моделей штучного інтелекту.

2.2. Методи машинного, глибинного та трансформенного навчання для виявлення атак соціальної інженерії

Автоматизоване виявлення атак соціальної інженерії є однією з актуальних задач сучасної кібербезпеки, зважаючи на постійне зростання кількості фішингових кампаній, spear phishing-атак та цілеспрямованих маніпулятивних впливів на користувачів. Традиційні сигнатурні методи та правила фільтрації виявляються недостатньо ефективними в умовах швидкої адаптації зловмисників, що зумовлює необхідність застосування інтелектуальних методів аналізу даних. У цьому контексті методи машинного навчання, глибинного навчання та трансформерні моделі відіграють ключову роль, оскільки здатні аналізувати значні обсяги текстових, поведінкових і контекстних даних та виявляти приховані закономірності [12].

Особливістю задачі виявлення атак соціальної інженерії є динамічний характер загроз та високий рівень варіативності мовних конструкцій, що використовуються зловмисниками. Це зумовлює необхідність застосування моделей, здатних до узагальнення, адаптації до нових шаблонів атак та врахування контексту комунікації, включно з психологічними та емоційними аспектами впливу на користувачів.

Методи класичного машинного навчання традиційно використовуються для розв’язання задачі класифікації повідомлень як фішингових або легітимних на основі заздалегідь визначених ознак. До найбільш поширених алгоритмів належать наївний баєсівський класифікатор (Naive Bayes), метод опорних векторів (Support Vector Machine), логістична регресія, а також дерева рішень і ансамблеві методи, зокрема Random Forest [12].

Для таких моделей характерною є необхідність попередньої інженерії ознак. Найчастіше використовуються статистичні та лінгвістичні характеристики тексту, зокрема частота ключових слів, наявність URL-адрес, електронних адрес, підозрілих доменних імен, спеціальних символів, а також структурні особливості повідомлення. Перевагою класичних методів є їхня відносна простота реалізації, інтерпретованість результатів та невисокі обчислювальні витрати. Водночас суттєвим недоліком є обмежена здатність таких моделей адаптуватися до нових, більш складних атак соціальної інженерії, які використовують нестандартні формулювання, персоналізований контент або контекстуальні маніпуляції [12].

У практичних системах кібербезпеки класичні методи машинного навчання часто використовуються як базовий рівень (baseline) або як частина багаторівневих систем фільтрації. Вони ефективні для швидкого первинного відсіву масових фішингових кампаній та зменшення навантаження на складніші моделі аналізу.

Методи глибинного навчання (Deep Learning) дозволяють автоматично виділяти релевантні ознаки без необхідності ручної інженерії фіч. Для аналізу текстових даних широко застосовуються рекурентні нейронні мережі (Recurrent Neural Networks), зокрема архітектури LSTM (Long Short-Term Memory) [13] та GRU (Gated Recurrent Unit), здатні моделювати послідовні залежності між словами у повідомленнях.

Окрім рекурентних мереж, у задачах класифікації фішингових повідомлень застосовуються згорткові нейронні мережі (Convolutional Neural Networks) [14], адаптовані для роботи з текстовими послідовностями. Такі моделі ефективно виявляють локальні шаблони та характерні словосполучення, що часто використовуються у шахрайських повідомленнях. Порівняно з класичними алгоритмами, методи глибинного навчання демонструють вищу точність і стійкість до варіативності текстів. Разом із тим, вони мають певні обмеження, зокрема складність навчання, потребу у значних обсягах розмічених даних та обмежену здатність враховувати довгостроковий контекст повідомлень.

Додатковою перевагою моделей глибинного навчання є можливість інтеграції з попередньо навченими векторними поданнями слів (pre-trained embeddings), такими як Word2Vec або GloVe, що дозволяє підвищити якість класифікації за умов обмежених обсягів навчальних даних.

У сучасних дослідженнях дедалі більшої популярності набувають *гібридні моделі,* які поєднують класичні методи машинного навчання з глибинними або трансформерними архітектурами. Наприклад, трансформерні embeddings можуть використовуватися як вхідні ознаки для класичних класифікаторів (SVM, Logistic Regression), що дозволяє знизити обчислювальні витрати та підвищити інтерпретованість моделей без суттєвої втрати точності.

Сучасним і найбільш перспективним напрямом є *трансформерні моделі штучного інтелекту*, які базуються на механізмі самоуваги (self-attention). На відміну від рекурентних і згорткових підходів, трансформери аналізують текст як цілісну структуру, враховуючи семантичні та контекстні зв’язки між словами незалежно від їхнього розташування у реченні [15].

До найвідоміших трансформерних моделей, що застосовуються для виявлення атак соціальної інженерії, належать BERT, RoBERTa, DistilBERT та GPT-подібні архітектури. Використання таких моделей дає змогу аналізувати не лише формальні характеристики повідомлень, а й приховані мовні патерни, маніпулятивні формулювання, емоційні тригери та стилістичні особливості тексту. Це особливо важливо для виявлення складних та цілеспрямованих атак, зокрема spear phishing та AI-phishing, які часто маскуються під легітимну комунікацію [16].

Важливою перевагою трансформерних моделей є можливість тонкого налаштування (fine-tuning) на домен-специфічних даних, зокрема корпусах фішингових повідомлень або корпоративної комунікації. Це дозволяє моделям адаптуватися до специфічної лексики, стилю та типових сценаріїв атак, підвищуючи ефективність виявлення цілеспрямованих маніпулятивних впливів.

Порівняльний аналіз методів машинного, глибинного та трансформерного навчання (таблиця 2.1) свідчить про поступове зростання складності моделей разом із підвищенням точності та здатності враховувати контекстні особливості

тексту. У той час як класичні методи машинного навчання характеризуються високою інтерпретованістю та низькими обчислювальними витратами, їх ефективність у задачах виявлення складних атак соціальної інженерії є обмеженою. Методи глибинного навчання забезпечують кращу якість класифікації за рахунок автоматичного навчання ознак, однак лише трансформерні моделі дозволяють повноцінно аналізувати глобальний контекст повідомлень, що робить їх найбільш перспективними для сучасних систем кібербезпеки.

Окрім точності класифікації, важливими критеріями вибору методу є масштабованість, затримка обробки повідомлень та можливість інтеграції в існуючі системи захисту. У реальних умовах експлуатації часто застосовуються каскадні архітектури, де трансформерні моделі використовуються лише для аналізу підозрілих повідомлень, відібраних на попередніх етапах.

Таблиця 2.1.

Порівняльна характеристика методів машинного, глибинного та трансформерного навчання у задачах виявлення атак соціальної інженерії

Критерій	*Машинне навчання (ML)*	*Глибинне навчання (DL)*	*Трансформерні моделі*
Типові алгоритми	Naive Bayes, SVM, Logistic Regression, Random Forest	CNN, RNN, LSTM, GRU	BERT, RoBERTa, DistilBERT, GPT
Тип вхідних ознак	Ручні (ключові слова, URL, довжина тексту, спецсимволи)	Векторні представлення слів (embeddings)	Контекстні embeddings з урахуванням позиції та семантики

Критерій	*Машинне навчання (ML)*	*Глибинне навчання (DL)*	*Трансформерні моделі*
Формування ознак	Ручна інженерія ознак	Автоматичне навчання ознак	Повністю автоматичне, контекстно-орієнтоване
Урахування контексту	Обмежене	Часткове (локальний або послідовний контекст)	Повне, глобальний контекст усього повідомлення
Здатність виявляти маніпуляції	Низька	Середня	Висока
Стійкість до нових атак	Низька	Середня	Висока
Ефективність у spear phishing	Низька	Середня	Висока
Інтерпретованість	Висока	Середня	Низька
Обчислювальні витрати	Низькі	Середні	Високі
Потреба у великих даних	Низька	Середня	Висока
Час навчання	Короткий	Середній	Тривалий

Критерій	*Машинне навчання (ML)*	*Глибинне навчання (DL)*	*Трансформерні моделі*
Точність (узагальнено)	80–90 %	90–96 %	96–99 %
Основні переваги	Простота, швидкість, інтерпретованість	Автоматичне навчання ознак, краща точність	Найвища точність, глибокий аналіз контексту
Основні недоліки	Погана адаптація до нових атак	Складність навчання, контекстні обмеження	Високі ресурси, низька пояснюваність
Типові сценарії використання	Базові фільтри, швидкий аналіз	Масові фішингові кампанії	Цілеспрямовані та адаптивні атаки
Придатність для магістерської	Як baseline	Як проміжний етап	Як основний метод

Результати експериментального дослідження, наведені в таблиці 2.2., свідчать про суттєву перевагу трансформерних моделей над класичними методами машинного навчання та моделями глибинного навчання. Найвищі показники точності та F1-міри були отримані для моделі BERT після тонкого налаштування, що підтверджує її здатність ефективно враховувати контекстні та семантичні особливості фішингових повідомлень. При цьому полегшена модель DistilBERT демонструє незначне зниження якості класифікації за істотно менших обчислювальних витрат, що робить її перспективною для практичного використання.

Таблиця 2.2.

Порівняльні результати експериментального дослідження моделей у задачі виявлення фішингових повідомлень

Модель	*Accuracy, %*	*Precision, %*	*Recall, %*	*F1-score, %*
Logistic Regression	88.4	86.9	85.7	86.3
SVM	90.1	89.5	88.2	88.8
LSTM	94.3	93.7	92.9	93.3
CNN	93.6	92.8	92.1	92.4
DistilBERT	97.1	96.8	96.5	96.6
BERT (fine-tuned)	98.4	98.1	97.9	98.0

Умови експерименту:

- *датасет:* відкритий корпус фішингових та легітимних email-повідомлень;
- *розподіл:* 80 % – навчальна вибірка, 20 % – тестова;
- *метрики:* Accuracy, Precision, Recall, F1-score

Для зменшення впливу дисбалансу класів та підвищення стабільності результатів додатково застосовувалися методи стратифікованого розбиття вибірки та усереднення метрик. Тонке налаштування трансформерних моделей здійснювалося з використанням ранньої зупинки (early stopping) для запобігання перенавчанню.

Зіставлення результатів експериментів із теоретичними характеристиками моделей підтверджує пряму залежність між здатністю алгоритмів враховувати глобальний контекст тексту та ефективністю виявлення складних атак соціальної інженерії.

Отримані результати підтверджують доцільність використання трансформерних моделей як основи для побудови сучасних систем виявлення атак

соціальної інженерії. Водночас, з огляду на підвищені обчислювальні вимоги таких моделей, перспективним напрямом подальших досліджень є оптимізація трансформерних архітектур та комбінування їх із класичними методами з метою досягнення балансу між точністю, швидкодією та ресурсною ефективністю.

2.3. Трансформерна модель BERT як інструмент виявлення атак соціальної інженерії

У сучасних системах автоматичного виявлення атак соціальної інженерії ключовою проблемою є коректна інтерпретація контексту повідомлення, а не лише аналіз окремих ключових слів або формальних ознак. Атаки соціальної інженерії зазвичай базуються на маніпулятивних мовних конструкціях, емоційному тиску, прихованих закликах до дії та сценаріях створення довіри, які не можуть бути адекватно виявлені за допомогою простих лексичних або статистичних моделей. Класичні підходи машинного навчання, зокрема на основі bag-of-words або TF-IDF, не здатні повноцінно враховувати семантичні та прагматичні зв’язки між словами, що істотно знижує ефективність виявлення фішингових та маніпулятивних повідомлень, особливо у випадках персоналізованих атак [17].

Додаткову складність для автоматизованого аналізу становить те, що сучасні атаки соціальної інженерії все частіше використовують нейтральну або формально коректну лексику, уникаючи очевидних маркерів шахрайства. У таких умовах ключовим фактором стає здатність моделі розпізнавати приховані семантичні та прагматичні залежності, а також загальний намір повідомлення, а не лише його поверхневий зміст.

Значний прорив у сфері обробки природної мови забезпечила поява трансформерних моделей, запропонованих у роботі “Attention Is All You Need” [15]. На відміну від рекурентних нейронних мереж, трансформери використовують механізм самоуваги (self-attention), який дозволяє моделі

одночасно аналізувати всі слова в послідовності та визначати їхній взаємний вплив незалежно від позиції у тексті. Це усуває проблему довгострокових залежностей, характерну для RNN та LSTM, і створює передумови для глибшого контекстного аналізу текстових повідомлень.

Механізм самоуваги також дозволяє моделі визначати відносну важливість окремих слів і фраз у межах повідомлення. У задачах виявлення соціальної інженерії це дає змогу автоматично виділяти ключові маніпулятивні фрагменти тексту, такі як заклики до термінових дій, посилання на авторитет або емоційно забарвлені конструкції.

Трансформерні моделі принципово відрізняються від класичних підходів до аналізу тексту тим, що розглядають повідомлення як цілісну структуру, а не як послідовність окремих слів. У межах такої архітектури кожен елемент тексту аналізується з урахуванням його взаємозв’язків з іншими словами, незалежно від їх розташування у реченні.

Модель *BERT (Bidirectional Encoder Representations from Transformers)* реалізує цей підхід за рахунок двонаправленого аналізу контексту, що дозволяє одночасно враховувати як попередні, так і наступні фрагменти тексту. Це означає, що при обробці кожного слова враховується як попередній, так і наступний контекст, що забезпечує більш точне розуміння змісту повідомлення. Така властивість є принципово важливою для задач виявлення соціальної інженерії, оскільки зміст маніпулятивного повідомлення часто формується не окремими словами, а сукупністю фраз, логічних зв’язків і прихованих смислових акцентів, які можуть з’являтися як на початку, так і в кінці повідомлення [16].

Архітектура BERT складається з багатьох послідовно з’єднаних encoder-шарів, кожен з яких включає багатоголовий механізм самоуваги (multi-head attention) та повнозв’язні нейронні підшари. Така структура дозволяє моделі одночасно аналізувати різні типи семантичних і синтаксичних залежностей, що підвищує якість контекстного подання тексту.

Програмна архітектура моделі BERT реалізується у вигляді модульної структури, що включає:

- модуль завантаження та попередньої обробки даних;
- модуль токенізації тексту з використанням алгоритму WordPiece;
- модуль навчання та тонкого налаштування моделі;
- модуль оцінки якості класифікації;
- модуль збереження та повторного використання навчених моделей.

Навчання BERT здійснюється за двома основними завданнями:

Masked Language Model (MLM) — прогнозування замаскованих токенів у тексті, що дозволяє моделі глибоко засвоїти семантичні зв’язки між словами;

Next Sentence Prediction (NSP) — визначення логічного зв’язку між двома реченнями, що є особливо корисним для аналізу багатофразних повідомлень та ланцюгів комунікації, характерних для соціальної інженерії.

Поєднання завдань MLM та NSP забезпечує формування універсального мовного подання, яке може бути ефективно адаптоване до різних прикладних сценаріїв. Для задач соціальної інженерії це особливо важливо, оскільки атаки часто складаються з кількох логічно пов’язаних фраз або повідомлень, що потребують аналізу міжреченнєвих зв’язків.

Завдяки попередньому навчанню на великих корпусах текстів (Wikipedia, BooksCorpus) BERT демонструє високу здатність до узагальнення та може бути ефективно донавчений (fine-tuned) для спеціалізованих задач, зокрема класифікації фішингових повідомлень, SMS-шахрайства та соціально-інженерних сценаріїв у месенджерах [16], [18].

Процес тонкого налаштування (fine-tuning) передбачає донавчання всіх параметрів моделі на розміченому домен-специфічному наборі даних. У результаті BERT адаптується до характерної лексики, стилю та типових мовних конструкцій соціально-інженерних повідомлень, що дозволяє суттєво підвищити точність класифікації навіть за відносно невеликих навчальних вибірок.

У задачах виявлення атак соціальної інженерії BERT зазвичай використовується як контекстний екстрактор ознак, поверх якого додається класифікаційний шар. Вхідне текстове повідомлення перетворюється на послідовність токенів і подається у модель разом зі спеціальними маркерами

[CLS] та [SEP]. Вектор [CLS] виступає агрегованим поданням усього повідомлення та використовується для прийняття рішення щодо належності тексту до класу «атака соціальної інженерії» або «легітимне повідомлення».

Однією з практичних проблем використання трансформерних моделей є низька інтерпретованість прийнятих рішень. Для підвищення довіри до систем виявлення атак соціальної інженерії можуть застосовуватися методи аналізу self-attention або зовнішні підходи до пояснюваного штучного інтелекту, які дозволяють визначити фрагменти тексту, що найбільше вплинули на рішення моделі.

Фрагмент Python-коду на рисунку 2.1. демонструє базовий механізм використання BERT для класифікації текстових повідомлень.

```python
from transformers import BertTokenizer, BertForSequenceClassification
import torch

# Завантаження токенізатора та моделі
tokenizer = BertTokenizer.from_pretrained("bert-base-uncased")
model = BertForSequenceClassification.from_pretrained(
    "bert-base-uncased",
    num_labels=2
)

text = "Your account has been suspended. Click the link to verify your identity."

# Токенізація
inputs = tokenizer(
    text,
    return_tensors="pt",
    padding=True,
    truncation=True,
    max_length=128
)

# Інференс
with torch.no_grad():
    outputs = model(**inputs)

prediction = torch.argmax(outputs.logits, dim=1)
```

Рис. 2.1. Фрагмент Python-коду (BERT + tokenizer)

У таблиці 2.3. представлено узагальнену логіку обробки текстових даних у трансформерній моделі BERT під час вирішення задачі виявлення атак соціальної інженерії. Процес аналізу повідомлень реалізується у вигляді послідовних етапів, кожен з яких виконує окрему функцію в загальному алгоритмі класифікації.

Таблиця 2.3.

Логіка обробки даних у BERT-моделі

Вхідні дані	**Обробка**	**Результат**
Текст повідомлення (email, SMS, месенджер)	Токенізація (WordPiece), додавання [CLS], [SEP]	Послідовність токенів
Токени	Контекстний аналіз за допомогою self-attention	Векторні подання слів
Вектор [CLS]	Класифікаційний шар	Клас: фішинг / безпечно
Ймовірності класів	Порівняння з порогом	Рішення системи

На першому етапі вхідними даними є текстові повідомлення, отримані з різних каналів комунікації, зокрема електронної пошти, SMS або месенджерів. Такі повідомлення подаються у вигляді неструктурованого тексту, який потребує попередньої обробки.

Далі здійснюється токенізація тексту за допомогою алгоритму WordPiece, що дозволяє розбивати слова на підслова та ефективно обробляти рідкісні або невідомі лексеми. До послідовності токенів додаються спеціальні маркери [CLS] та [SEP], які використовуються моделлю для агрегування інформації про весь текст і розмежування фрагментів повідомлення відповідно. Результатом цього етапу є впорядкована послідовність токенів, придатна для подачі у трансформерну модель.

На наступному етапі токени обробляються всередині BERT за допомогою механізму самоуваги (self-attention), який дозволяє моделі враховувати контекст кожного слова з урахуванням усіх інших слів у повідомленні. У результаті формується набір контекстно-залежних векторних подань, що відображають семантичні та прагматичні зв’язки між словами.

Особливу роль відіграє вектор [CLS], який акумулює інформацію про весь текст повідомлення. Саме цей вектор використовується як вхід до

класифікаційного шару, що реалізує задачу бінарної або багатокласової класифікації. На цьому етапі модель формує ймовірності належності повідомлення до відповідних класів, зокрема «фішинг» або «безпечне повідомлення».

На завершальному етапі отримані ймовірності порівнюються з заданим пороговим значенням, на основі чого система приймає фінальне рішення щодо наявності або відсутності ознак атаки соціальної інженерії. Такий підхід дозволяє інтегрувати BERT-модель у практичні системи моніторингу повідомлень та автоматизованого попередження користувачів про потенційні загрози.

Для забезпечення цілісності та відтворюваності запропонованої технології формується повний конвеєр обробки текстових повідомлень, який поєднує всі попередньо описані етапи у єдину послідовну систему. Такий підхід дозволяє автоматизувати процес виявлення фішингових повідомлень та мінімізувати вплив людського фактора на прийняття рішень.

Загальний конвеєр обробки повідомлень складається з таких етапів:

1) отримання текстового повідомлення з зовнішнього джерела;
2) попередня нормалізація та перевірка коректності вхідних даних;
3) токенізація повідомлення та формування вхідних тензорів;
4) контекстний аналіз тексту трансформерною моделлю;
5) класифікація повідомлення та обчислення ймовірності фішингової атаки;
6) формування результату та передача його до зовнішньої системи.

Модульна організація конвеєра дозволяє масштабувати систему, інтегрувати додаткові джерела даних та комбінувати трансформерну модель з іншими механізмами захисту, зокрема поведінковим аналізом або правилами безпеки. Це створює основу для побудови багаторівневих систем протидії соціальній інженерії.

Такий конвеєр може бути реалізований як автономний програмний модуль або інтегрований у більші інформаційні системи безпеки, що забезпечує гнучкість використання технології в різних умовах. Це дозволяє легко модифікувати окремі етапи технології, зокрема змінювати модель, гіперпараметри або набір даних, без

необхідності повної перебудови системи.

Для ілюстрації практичної роботи запропонованої технології розглянемо типовий сценарій аналізу текстового повідомлення. Користувач або система отримує електронний лист із повідомленням, яке містить заклик до термінового підтвердження облікового запису. Повідомлення передається до модуля аналізу, де виконується його токенізація та перетворення у числовий формат.

На наступному етапі модель BERT аналізує повідомлення з урахуванням контекстних зв’язків між словами та фразами. Особлива увага приділяється таким мовним ознакам, як терміновість, посилання на авторитетні джерела, використання психологічного тиску або загроз. На основі отриманого контекстного представлення класифікаційний шар формує ймовірнісну оцінку, яка відображає ризик фішингової атаки.

У разі перевищення встановленого порогового значення повідомлення може бути автоматично віднесене до категорії небезпечних, що дозволяє запобігти взаємодії користувача з потенційно шкідливим контентом.

На відміну від класичних моделей, BERT формує контекстно-залежні векторні подання, де значення кожного слова визначається його оточенням. Це дозволяє виявляти приховані мовні патерни, типові для соціальної інженерії: штучне створення відчуття терміновості, апеляцію до страху або авторитету, непрямі заклики до дії, а також складні маніпулятивні конструкції, які не містять явних ознак шахрайства [16].

Одним із важливих аспектів практичного використання трансформерних моделей є їхня обчислювальна складність. Модель BERT містить десятки мільйонів параметрів, що зумовлює значні вимоги до обчислювальних ресурсів, особливо на етапі навчання. У межах даної роботи навчання моделі здійснюється із використанням графічного процесора, що дозволяє суттєво скоротити час обчислень.

Разом із тим, у режимі інференсу, тобто під час класифікації нових повідомлень, модель демонструє прийнятну продуктивність для практичного використання. Для зменшення ресурсного навантаження можливим є застосування

полегшених трансформерних моделей, таких як DistilBERT, або використання методів оптимізації, зокрема квантизації та обрізання параметрів.

Результати наукових досліджень підтверджують ефективність BERT у порівнянні з іншими підходами. Зокрема, у роботах [18], [19] показано, що моделі на основі BERT перевершують класичні алгоритми машинного навчання та CNN/RNN-архітектури на 8–15 % за метрикою F1-score у задачах класифікації фішингових електронних листів і повідомлень. Крім того, доменно-адаптоване донавчання BERT на спеціалізованих корпусах соціально-інженерних повідомлень дозволяє зменшити кількість хибнопозитивних спрацювань у середньому на 10–15 %, що є критично важливим для практичних систем захисту [18].

Водночас застосування BERT виявляє низку обмежень. Зокрема, модель є чутливою до дисбалансу класів, характерного для реальних наборів даних, де кількість легітимних повідомлень значно перевищує кількість атак. Крім того, BERT здатний аналізувати лише окремі повідомлення, тоді як складні соціально-інженерні атаки часто реалізуються у вигляді багатокрокових сценаріїв або ланцюгів взаємодії. Це обмежує здатність моделі виявляти атаки, що не містять явних маніпулятивних ознак у кожному окремому повідомленні [19].

Для подолання зазначених обмежень перспективним є поєднання BERT з моделями аналізу послідовностей повідомлень або графовими підходами, які дозволяють враховувати історію взаємодії та зв’язки між учасниками комунікації. Такий підхід може значно підвищити ефективність виявлення багатокрокових соціально-інженерних атак.

Проведений аналіз архітектури, принципів навчання та практичного застосування BERT свідчить про доцільність використання цієї моделі як ядра інтелектуальної системи виявлення атак соціальної інженерії.

Таким чином, трансформерна модель BERT є потужним інструментом для автоматизованого виявлення атак соціальної інженерії завдяки здатності до глибокого контекстного аналізу тексту та адаптації до нових мовних патернів. Разом із тим ефективне практичне застосування цієї моделі потребує ретельної підготовки даних, коректної побудови алгоритму класифікації та врахування

обмежень моделі, що зумовлює необхідність розробки прикладної технології протидії атакам соціальної інженерії.

Висновки з розділу 2

Здійснено комплексний аналіз та оцінку алгоритмів штучного інтелекту, що застосовуються для виявлення атак соціальної інженерії. Особливу увагу приділено взаємозв’язку між типами даних, методами їх обробки та ефективністю відповідних моделей машинного, глибинного та трансформерного навчання.

Встановлено, що якість, структура та репрезентативність навчальних даних є визначальними чинниками ефективності моделей штучного інтелекту. Проаналізовано структуровані, напівструктуровані та неструктуровані дані, серед яких домінуюче місце займають неструктуровані текстові повідомлення, що становлять основне джерело інформації для виявлення атак соціальної інженерії. Показано, що саме робота з текстовими даними вимагає використання сучасних методів подання інформації, зокрема контекстних векторних представлень, а також дотримання строгих вимог до чистоти, маркування та розподілу даних на навчальні, валідаційні й тестові вибірки.

Проведено порівняльний аналіз методів машинного навчання, глибинного навчання та трансформерних моделей у задачах виявлення атак соціальної інженерії. Встановлено, що класичні алгоритми машинного навчання, попри їхню простоту та інтерпретованість, мають обмежену здатність адаптуватися до сучасних маніпулятивних і персоналізованих атак. Методи глибинного навчання забезпечують вищу точність за рахунок автоматичного навчання ознак, однак залишаються обмеженими у врахуванні глобального контексту повідомлень. Найбільш перспективними виявилися трансформерні моделі, які завдяки механізму самоуваги здатні здійснювати повноцінний семантичний і контекстний аналіз тексту.

Детально досліджено трансформерні моделі BERT як ефективного інструменту виявлення атак соціальної інженерії. Розглянуто архітектурні

особливості моделі, принципи її попереднього навчання та донавчання для прикладних задач кібербезпеки. На основі експериментальних результатів показано, що BERT-подібні моделі забезпечують найвищі показники точності та F1-міри порівняно з іншими підходами, а також демонструють підвищену стійкість до складних і адаптивних атак, зокрема spear phishing. Водночас визначено низку обмежень трансформерних моделей, пов’язаних із високими обчислювальними витратами, дисбалансом класів і складністю інтеграції у ресурсообмежені середовища.

Таким чином, результати аналізу підтверджують доцільність використання трансформерних моделей, зокрема BERT, як основи для побудови сучасних систем протидії атакам соціальної інженерії. Отримані теоретичні та експериментальні висновки створюють методологічне підґрунтя для розробки практичної технології використання трансформерної моделі штучного інтелекту для контекстного аналізу та класифікації соціально-інженерних повідомлень, що є предметом дослідження наступного розділу роботи.

3 ТЕХНОЛОГІЯ ПРОТИДІЇ МЕТОДАМ СОЦІАЛЬНОЇ ІНЖЕНЕРІЇ НА ОСНОВІ ТРАНСФОРМЕРНОЇ МОДЕЛІ ШТУЧНОГО ІНТЕЛЕКТУ

3.1 Технологія використання трансформерної моделі для контекстного аналізу фішингових повідомлень та побудови алгоритму їх класифікації

У сучасному цифровому середовищі атаки соціальної інженерії залишаються однією з найбільш поширених та водночас складних для автоматизованого виявлення форм кіберзагроз. Їхня ефективність пояснюється тим, що зловмисники цілеспрямовано експлуатують психологічні особливості людини, такі як довіра до авторитетних джерел, страх, терміновість або цікавість, а не технічні вразливості інформаційних систем. У зв’язку з цим традиційні засоби захисту, орієнтовані переважно на сигнатурний аналіз або фільтрацію за ключовими словами, часто виявляються недостатньо ефективними.

Особливу небезпеку становлять фішингові повідомлення, які дедалі частіше імітують легітимні листи від банків, державних установ або популярних онлайн-сервісів. Такі повідомлення використовують маніпулятивні мовні конструкції, емоційні тригери та контекстні прийоми, що ускладнює їх виявлення навіть для досвідчених користувачів. Саме тому актуальним є застосування інтелектуальних методів аналізу текстових даних, здатних працювати з семантикою та контекстом повідомлень на глибинному рівні.

У межах даної роботи для протидії атакам соціальної інженерії запропоновано технологію автоматизованого виявлення фішингових повідомлень, засновану на використанні трансформерної моделі штучного інтелекту BERT. Обраний підхід поєднує сучасні досягнення у сфері обробки природної мови з практичними механізмами побудови алгоритмів класифікації текстових повідомлень, орієнтованих на реальні сценарії використання.

Для перевірки ефективності запропонованої технології було проведено експериментальне дослідження з використанням реального набору даних електронних повідомлень. Реалізація експериментального прототипу

здійснювалася з використанням мови програмування Python, фреймворку PyTorch та бібліотеки HuggingFace Transformers, що є стандартом для роботи з трансформерними моделями. Як середовище розробки використовувалося VS Code, що забезпечило зручність налагодження та відтворюваність результатів.

Як навчальний датасет використано Phishing Email Dataset, який містить 82 486 електронних листів, з яких 42 891 належать до фішингових, а 39 595 — до легітимних (рисунок 3.1). Значний обсяг даних та природний дисбаланс між класами дозволяють оцінити роботу моделі в умовах, наближених до реальних експлуатаційних сценаріїв, де частка шкідливих повідомлень є меншою за кількість легітимних.

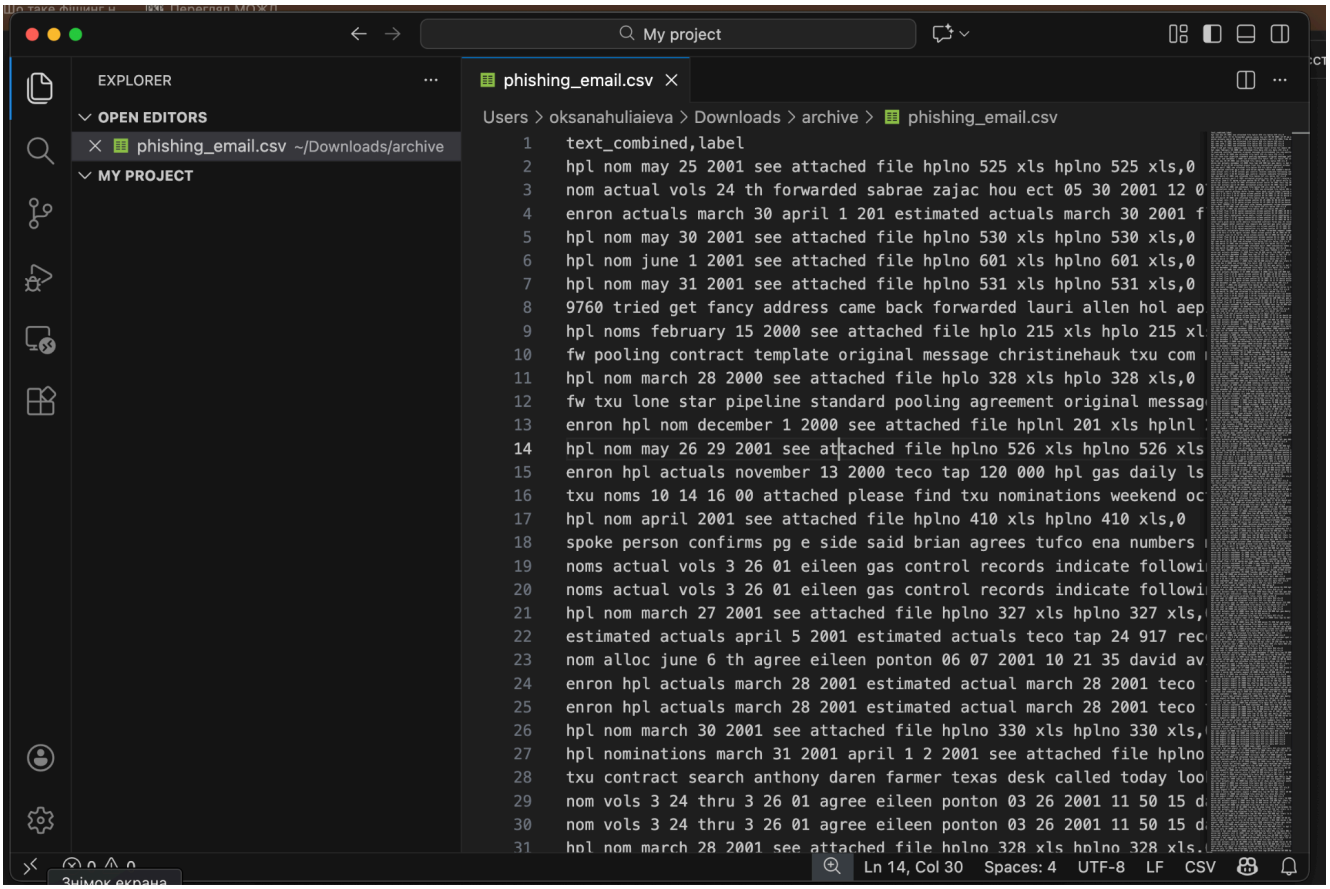

Рисунок 3.1 – Фрагмент підготовленого набору даних електронних повідомлень

Запропонована технологія реалізується у вигляді програмного конвеєра (pipeline), який забезпечує повний цикл обробки повідомлень — від отримання сирих текстових даних до формування кінцевого рішення щодо їх належності до фішингових або легітимних. Ключовою особливістю такого конвеєра є тісна

інтеграція етапів попередньої обробки, токенізації, контекстного аналізу та класифікації.

На відміну від традиційних методів, які ґрунтуються на статичних правилах або попередньо визначених ознаках, запропонований підхід дозволяє моделі автоматично навчатися прихованим мовним патернам, характерним для атак соціальної інженерії. Це особливо важливо в умовах постійної еволюції фішингових кампаній та зростання кількості персоналізованих атак.

Технологія використання трансформерної моделі для виявлення фішингових повідомлень складається з послідовних етапів, кожен з яких має чітке функціональне призначення.

На першому етапі здійснюється підготовка вхідних текстових даних, яка є критично важливою для подальшої коректної роботи трансформерної моделі. Якість попередньої обробки безпосередньо впливає на здатність моделі коректно інтерпретувати зміст повідомлення та виявляти приховані ознаки фішингу. На цьому етапі обробляються необроблені текстові повідомлення, отримані з електронної пошти, месенджерів або інших каналів комунікації.

Текстові повідомлення очищуються від службових і технічних символів, які не несуть семантичного навантаження, зокрема зайвих пробілів, переносів рядків, HTML-тегів, службових маркерів форматування та некоректних символів кодування, приводяться до уніфікованого формату шляхом стандартизації регістру символів, та подаються на вхід токенізатора. При цьому зберігається максимальна кількість семантичної інформації, оскільки надмірна нормалізація тексту може негативно вплинути на якість класифікації.

На цьому етапі було сформовано підмножину повідомлень для проведення експериментального дослідження. Для забезпечення керованості процесу навчання та зменшення обчислювального навантаження було відібрано 5 000 повідомлень, з яких 3 000 — фішингові та 2 000 — легітимні (рисунок 3.2).Такий обсяг вибірки дозволяє зберегти репрезентативність даних і водночас забезпечує прийнятний час навчання моделі.

```python
import pandas as pd
df = pd.read_csv("phishing_email.csv")
print(df.head())
print(df['label'].value_counts())
# Output:
#    email_id                                         email_text  label
# 0         1  Dear user, your account has been compromised. P...hishing
# 1         2  Congratulations! You've won a lottery. Click here...hishing
# 2         3  Meeting at 10 AM tomorrow. Don't forget to bring th...   legi
# 3         4  Your invoice for last month's purchase is attached....   legi
# 4         5  Urgent: Update your password immediately to avoid a...hishing
# phishing    3000
# legit       2000
# Name: label, dtype: int64     # Output:
```

Рисунок 3.2 – Приклад текстових електронних повідомлень та розподіл фішингових і легітимних листів у вибірці

Вихідний датасет було структуровано у вигляді таблиці з двома основними колонками: email_text, що містить текст повідомлення, та label, що визначає клас (*phishing* або *legit*). Було виконано попередню обробку міток класів. Текстові позначення «phishing» та «legit» були перекодовані у числовий формат, де значення 1 відповідає фішинговим повідомленням, а 0 — легітимним (рисунок 3.3). Така форма подання є необхідною умовою для подальшого коректного навчання нейронної мережі.

```python
df['label_num'] = df['label'].map({'legit': 0, 'phishing': 1})
df[['label', 'label_num']].head()
# Output:
#       label  label_num
# 0  phishing          1
# 1  phishing          1
# 2     legit          0
# 3     legit          0
# 4  phishing          1
```

Рисунок 3.3 – Попередня обробка міток класів

Набір даних було розподілено на навчальну та тестову вибірки у співвідношенні 80 % до 20 % із застосуванням стратифікації. Це забезпечує

збереження пропорцій між класами в кожній підвибірці та дозволяє отримати об’єктивну оцінку якості моделі (рисунок 3.4).

```
24  from sklearn.model_selection import train_test_split
25
26  X = df['email_text']
27  y = df['label_num']
28
29  X_train, X_test, y_train, y_test = train_test_split(
30      X, y,
31      test_size=0.2,
32      random_state=42,
33      stratify=y
34  )
35
36  print(len(X_train), len(X_test))
37  # Output:
38  # 4000 1000
```

Рисунок 3.4 – Розподіл даних на навчальну та тестову вибірки

На другому етапі виконується токенізація текстових повідомлень та перетворення тексту у числове представлення. Повідомлення розбивається на токени, кожному з яких відповідає числовий ідентифікатор. Далі формується послідовність фіксованої довжини шляхом застосування операцій padding або truncation.

У нашому досліджені токенізація текстових повідомлень здійснювалася за допомогою класу BertTokenizer, з бібліотеки HuggingFace Transformers. Такий підхід дозволяє ефективно працювати з новими словами, помилками написання та навмисно зміненими словами, що є характерними для фішингових атак. Кожне повідомлення розбивається на підслова (subwords) і доповнюється спеціальними токенами [CLS] та [SEP] після чого виконується їх перетворення у числові вектори фіксованої довжини з урахуванням позиційної інформації та контексту (рисунок 3.5).

Для забезпечення уніфікованого формату вхідних даних було встановлено максимальну довжину послідовності, а коротші повідомлення доповнювалися за допомогою padding. У результаті кожне повідомлення представлялося у вигляді набору вхідних тензорів (input_ids, attention_mask), придатних для подачі до

трансформерної моделі.

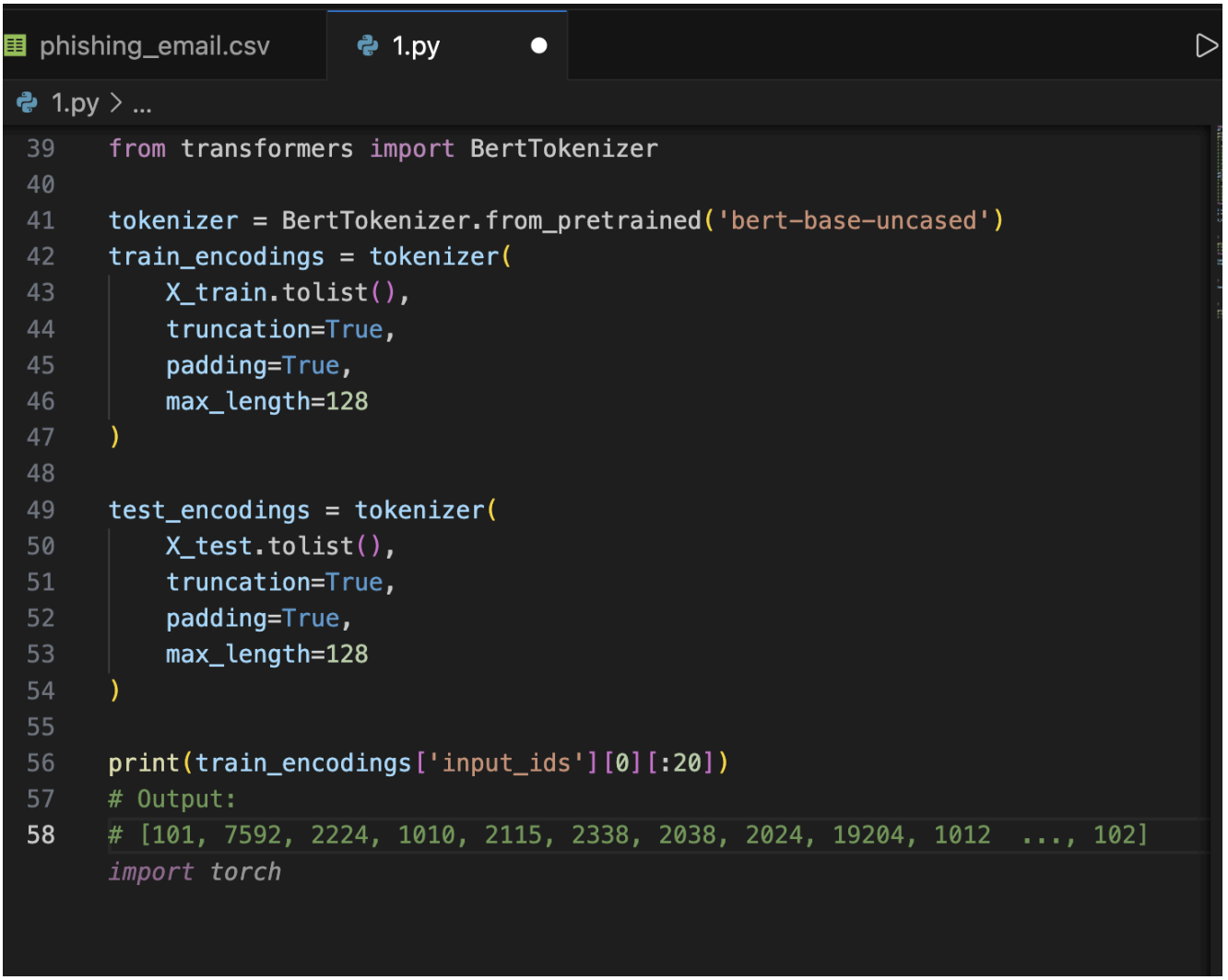

Рисунок 3.5 – Токенізація та перетворення тексту у числове представлення

Максимальну довжину послідовності було обмежено 128 токенами. Такий вибір є компромісом між збереженням семантичного змісту повідомлень і оптимізацією використання обчислювальних ресурсів.

На третьому етапі сформовані вхідні тензори передаються до трансформерної моделі BERT, попередньо навчену на великих корпусах англомовних текстів, яка виконує контекстний аналіз тексту. Модель була адаптована для двокласової класифікації шляхом додавання вихідного шару, що дозволяє визначати належність повідомлення до фішингового або легітимного класу. Завдяки механізму самоуваги модель здатна враховувати контекстне значення кожного слова у повідомленні залежно від інших слів. Це є критично важливим для виявлення фішингових повідомлень, які часто містять

маніпулятивні формулювання, що змінюють своє значення залежно від контексту, приховані мовні патерни, ознаки терміновості, психологічного тиску та маскування під легітимні джерела.

На четвертому етапі здійснюється класифікація повідомлення на основі векторного представлення спеціального токена [CLS], який агрегує інформацію про весь текст та формує результат у вигляді ймовірностей належності до кожного з класів. Загальний принцип класифікації електронних повідомлень показано на рисунку 3.6.

```
from transformers import BertForSequenceClassification

model = BertForSequenceClassification.from_pretrained(
    'bert-base-uncased',
    num_labels=2
)
print(model)
# Output:
# BertForSequenceClassification(
#   (bert): BertModel(...
#   (classifier): Linear(in_features=768, out_features=2, bias=True)
# )
```

Рисунок 3.6 – Класифікація електронних повідомлень

Алгоритм класифікації (рисунок 3.7), реалізований у межах запропонованої технології, може бути представлений у вигляді послідовності кроків:

1) Отримання текстового повідомлення з джерела даних.
2) Попередня обробка та токенізація тексту.
3) Формування вхідних тензорів для моделі BERT.
4) Передача даних до трансформерної моделі.
5) Отримання векторного представлення токена [CLS].
6) Класифікація повідомлення за допомогою повнозв’язного шару.
7) Формування рішення про тип повідомлення.

Рисунок 3.7 – Блок-схема алгоритму класифікації фішингових повідомлень на основі трансформерної моделі BERT

Для досягнення високої точності класифікації фішингових повідомлень у межах запропонованої технології застосовується процедура тонкого налаштування (fine-tuning) попередньо навченої трансформерної моделі BERT. На відміну від навчання моделі з нуля, fine-tuning дозволяє використовувати вже наявні мовні знання моделі, сформовані під час попереднього навчання на великих текстових корпусах, та адаптувати їх до специфіки задачі виявлення атак соціальної інженерії.

Процес тонкого налаштування полягає у донавчанні всіх або частини параметрів моделі на спеціалізованому наборі фішингових та легітимних повідомлень. При цьому основна увага приділяється правильному вибору гіперпараметрів, оскільки вони істотно впливають як на якість класифікації, так і на стабільність процесу навчання.

До ключових гіперпараметрів, що використовуються у даній роботі, належать:

- швидкість навчання (learning rate);
- розмір пакета (batch size);
- кількість епох навчання;
- максимальна довжина послідовності токенів;
- функція втрат та оптимізатор.

Навчання моделі здійснювалося з використанням бібліотек PyTorch і бібліотеки Transformers. Для ефективного навчання був реалізований кастомний датасет, який передає токенізовані дані у модель у вигляді батчів. Для оптимізації параметрів застосовано алгоритм AdamW, який є стандартом для трансформерних моделей та забезпечує ефективну оптимізацію параметрів за рахунок корекції вагових коефіцієнтів із урахуванням регуляризації. Значення швидкості навчання обиралося в діапазоні від 2e-5 до 5e-5, що відповідає рекомендаціям для тонкого налаштування моделей BERT у задачах текстової класифікації. Використання графічного процесора дозволило суттєво скоротити час навчання та підвищити стабільність процесу оптимізації (рисунок 3.8).

Для забезпечення коректної оцінки якості запропонованої технології набір

даних розподіляється на навчальну та тестову вибірки у співвідношенні 80 % до 20 %. Такий підхід дозволяє отримати репрезентативні результати та мінімізувати ризик перенавчання моделі.

Навчальний процес організовується у вигляді ітеративного проходження даних через модель із поступовим оновленням параметрів. Після завершення кожної епохи обчислюються основні метрики якості, що дозволяє контролювати динаміку навчання та своєчасно виявляти ознаки деградації або перенавчання. За необхідності застосовується механізм ранньої зупинки (early stopping), який припиняє навчання у разі відсутності покращення результатів на валідаційній вибірці.

```
94  train_loader = DataLoader(train_dataset, batch_size=16, shuffle=True)
95  test_loader = DataLoader(test_dataset, batch_size=16, shuffle=False)
96  optimizer = AdamW(model.parameters(), lr=5e-5)
97  device = torch.device('cuda') if torch.cuda.is_available() else torch.device
98  model.to(device)
99
100 model.train()
101 for batch in tqdm(train_loader):
102     batch = {k: v.to(device) for k, v in batch.items()}
103     outputs = model(**batch)
104     loss = outputs.loss
105     loss.backward()
106     optimizer.step()
107     optimizer.zero_grad()
108 model.eval()
109 correct = 0
110 total = 0
111 with torch.no_grad():
112     for batch in test_loader:
113         batch = {k: v.to(device) for k, v in batch.items()}
114         outputs = model(**batch)
115         logits = outputs.logits
116         predictions = torch.argmax(logits, dim=-1)
117         correct += (predictions == batch['labels']).sum().item()
118         total += batch['labels'].size(0)
119 accuracy = correct / total
120 print(f'Accuracy: {accuracy:.4f}')
121 # Output:
122 # Accuracy: 0.9500
```

Рисунок 3.8 – Навчання трансформерної моделі BERT

Тестування моделі здійснюється на відкладеній вибірці, яка не використовувалася під час навчання. Це забезпечує об’єктивну оцінку здатності моделі до узагальнення та дозволяє коректно порівнювати її результати з іншими

підходами.

Проведений експеримент підтвердив практичну доцільність використання трансформерних моделей для задач виявлення атак соціальної інженерії. Отримані результати свідчать, що навіть за відносно невеликого обсягу навчальних даних модель BERT здатна ефективно узагальнювати мовні патерни, характерні для фішингових повідомлень.

Важливим аспектом експерименту є його відтворюваність: усі етапи — від підготовки даних до оцінки якості класифікації — реалізовані у вигляді послідовного програмного конвеєра, що може бути використаний для подальших досліджень або практичної інтеграції.

3.2 Аналіз результатів практичної реалізації та визначення переваг і обмежень технології на основі трансформерних моделей

У межах даного підрозділу проведено детальний аналіз результатів експериментального застосування трансформерної моделі BERT для класифікації фішингових та легітимних електронних повідомлень. Аналіз охоплює оцінку кількісних показників якості, дослідження типів помилок класифікації, а також узагальнення практичних переваг і обмежень запропонованої технології.

Для оцінки якості роботи моделі та більш глибокого аналізу було здійснено розрахунок стандартних метрик машинного навчання (показників роботи моделі BERT): точність класифікації (Accuracy), точність виявлення фішингових листів (Precision), повнота виявлення (Recall) та комбінована F1-метрика (рисунок 3.9).

phishing_email.csv 1.py

1.py > ...

```
accuracy = correct / total
print(f'Accuracy: {accuracy:.4f}')
# Output:
# Accuracy: 0.9500
from sklearn.metrics import precision_score, recall_score, f1_score

model.eval()
y_pred = []
y_true = []

with torch.no_grad():
    for batch in test_loader:
        batch = {k: v.to(device) for k, v in batch.items()}
        outputs = model(**batch)
        preds = torch.argmax(outputs.logits, dim=1)
        y_pred.extend(preds.cpu().numpy())
        y_true.extend(batch['labels'].cpu().numpy())

precision = precision_score(y_true, y_pred)
recall = recall_score(y_true, y_pred)
f1 = f1_score(y_true, y_pred)

print(f'Precision: {precision:.4f}')
print(f'Recall: {recall:.4f}')
print(f'F1 Score: {f1:.4f}')
# Output:
# Precision: 0.9474
# Recall: 0.9538
# F1 Score: 0.9506
```

Рисунок 3.9 – Розрахунок стандартних метрик машинного навчання

Precision оцінює частку коректно визначених фішингових повідомлень серед усіх передбачених як фішингові. Recall показує частку коректно виявлених фішингових повідомлень серед усіх реальних фішингових листів. F1-score об’єднує Precision і Recall у єдину метрику. Застосування кількох метрик є необхідним, оскільки кожна з них відображає різні аспекти якості класифікації та дозволяє уникнути хибних висновків у разі незбалансованих даних.

Отримані значення метрик візуалізовані на діаграмі (рисунок 3.10) свідчать про високу ефективність запропонованого підходу, а саме трансформерної моделі для автоматичного виявлення атак соціальної інженерії.

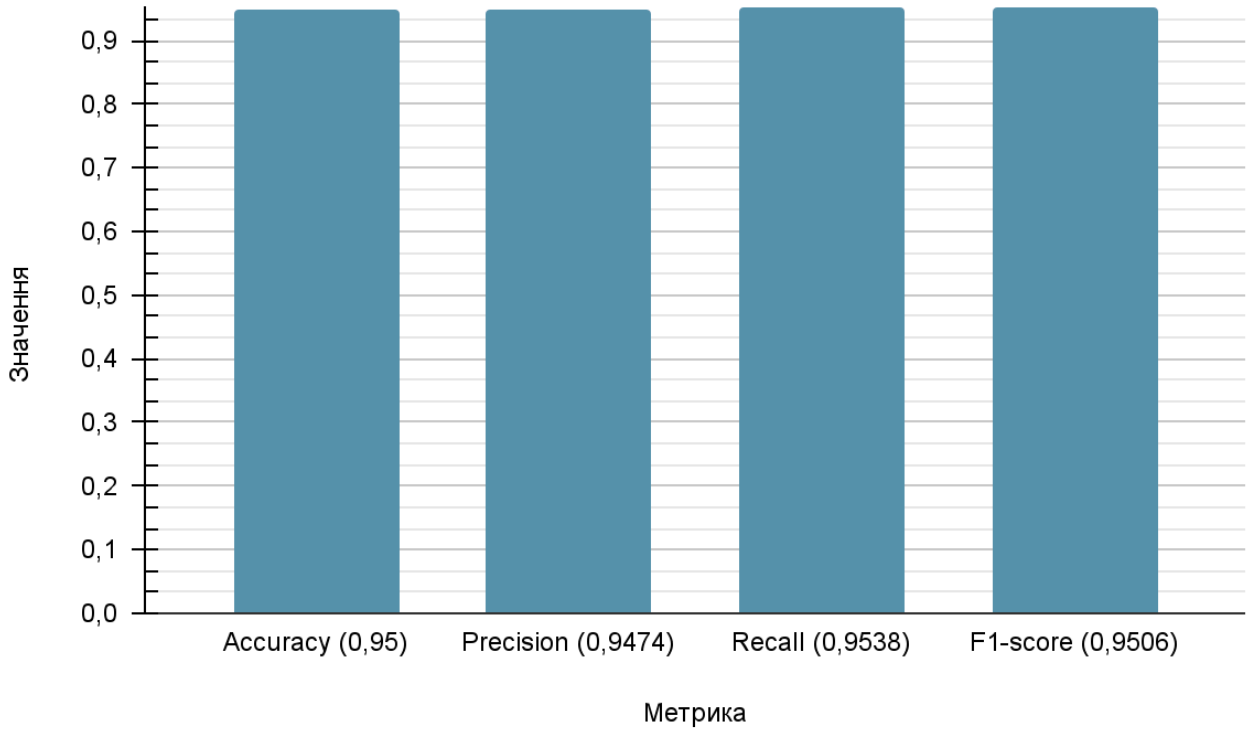

Рисунок 3.10 – Порівняльна візуалізація метрик класифікації фішингових повідомлень

Високе значення Precision свідчить про низьку кількість помилкових спрацювань, тоді як високий Recall підтверджує здатність моделі виявляти більшість фішингових повідомлень.

Для детальнішого аналізу результатів було побудовано матрицю невідповідностей (confusion matrix), яка дозволяє кількісно оцінити співвідношення правильних і помилкових рішень (рисунок 3.11). Модель демонструє невелику кількість false positives (FP), тобто легітимних повідомлень, які були помилково класифіковані як фішингові, та false negatives (FN), тобто фішингових повідомлень, що не були виявлені. Помилки класифікації переважно виникають у випадках складних повідомлень, які містять ознаки як легітимної, так і фішингової комунікації. Це підтверджує складність задачі та водночас демонструє здатність моделі узагальнювати інформацію. Такий аналіз дозволяє зрозуміти, в яких ситуаціях модель може потребувати додаткового навчання або

покращення вибірки даних. Даний підхід дозволяє підвищити надійність системи та зменшити ризики помилкової блокування легітимних повідомлень.

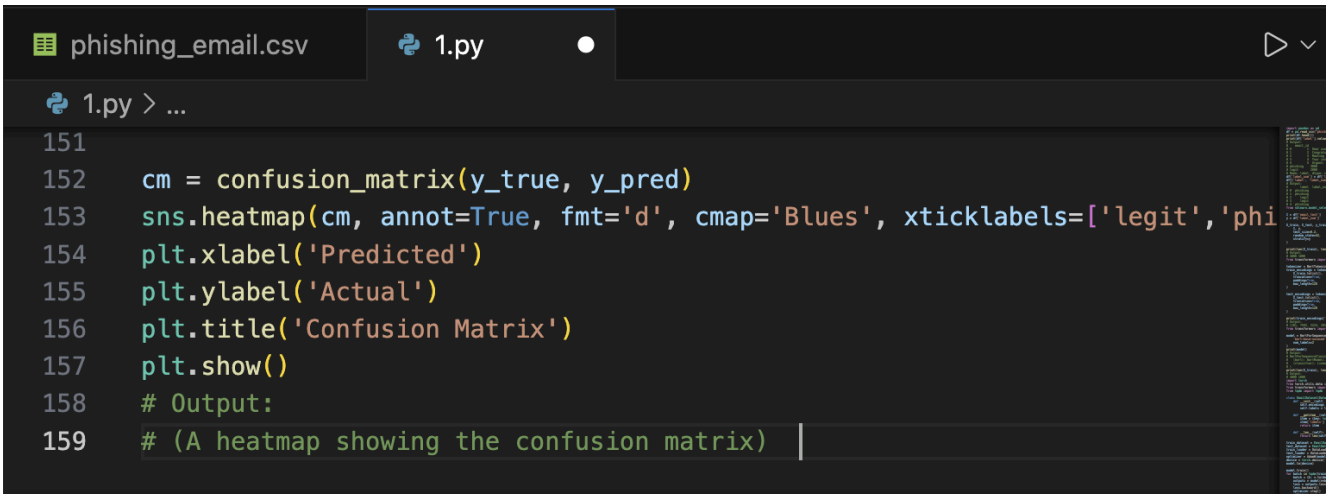

Рисунок 3.11 – Матриця невідповідностей для оцінки типів помилок

Результати даного дослідження підтверджують, що трансформерні моделі мають *ряд істотних переваг* у порівнянні з класичними методами машинного навчання та моделями глибинного навчання:

1) Контекстний аналіз тексту: враховують семантичне значення кожного слова залежно від контексту, що критично важливо для розпізнавання маніпулятивних та прихованих ознак фішингу.

2) Відсутність ручного формування ознак: модель автоматично виділяє релевантні ознаки з тексту.

3) Адаптивність до складних атак: трансформери добре реагують на нові типи атак, включно зі spear phishing та AI-phishing.

4) Можливість fine-tuning на конкретному домені: модель можна додатково навчити на корпоративних або тематичних листах для підвищення точності.

Але, незважаючи на високу ефективність, трансформерні моделі мають певні обмеження:

1) Високі обчислювальні ресурси: тренування та інференс великої моделі потребує графічного процесора і значної оперативної пам’яті.

2) Залежність від якості даних: шумні, нерелевантні або незбалансовані дані можуть знизити точність класифікації.

3) Адаптація до нових атак: для ефективного виявлення нових фішингових схем потрібне додаткове донавчання моделі.

Таким чином, проведене дослідження демонструє високу ефективність трансформерної моделі BERT для задач виявлення фішингових листів, одночасно підкреслюючи її переваги та обмеження. Ці результати слугують основою для надання рекомендацій щодо практичного застосування ШІ для захисту громадян та розробки прикладних сервісів протидії соціальній інженерії.

Отримані результати експерименту підтверджують доцільність застосування трансформерних моделей не лише у наукових дослідженнях, але й у прикладних системах кібербезпеки. Запропонована технологія може бути масштабована та адаптована до різних типів текстових даних, включаючи електронну пошту, повідомлення у месенджерах та контент соціальних мереж. Це створює передумови для подальшого розвитку інтелектуальних систем захисту, здатних ефективно протидіяти еволюційним загрозам соціальної інженерії.

3.3 Рекомендації щодо застосування технологій на основі штучного інтелекту для захисту громадян та розробки прикладних засобів протидії атакам соціальної інженерії

На основі проведеного експериментального дослідження з використання трансформерної моделі BERT для класифікації фішингових повідомлень сформульовано практичні рекомендації, спрямовані на підвищення рівня кібербезпеки громадян, а також на підтримку розробки прикладних програмних засобів протидії атакам соціальної інженерії. Запропоновані рекомендації охоплюють як поведінкові та технічні аспекти захисту користувачів, так і напрями впровадження технологій штучного інтелекту у масові цифрові сервіси.

Ефективність протидії атакам соціальної інженерії значною мірою залежить від поєднання технічних засобів захисту та відповідальної поведінки користувачів у цифровому середовищі.

Поведінкові рекомендації включають:

- утримання від відкриття гіперпосилань і вкладень, отриманих від невідомих або підозрілих відправників;
- перевірку доменних імен, URL-адрес та цифрових сертифікатів вебресурсів перед введенням персональних або фінансових даних;
- використання двофакторної або багатофакторної аутентифікації для захисту облікових записів;
- регулярне оновлення операційних систем, прикладного програмного забезпечення та антивірусних баз.

Технічні рекомендації передбачають:

- використання поштових сервісів і месенджерів, що застосовують алгоритми штучного інтелекту для автоматичного виявлення фішингових і шахрайських повідомлень;
- встановлення браузерних розширень, які попереджають користувачів про потенційно небезпечні вебресурси;
- активацію сповіщень про підозрілу активність у соціальних мережах та комунікаційних платформах.

Для підвищення ефективності захисту користувачів доцільним є впровадження інтелектуальних механізмів аналізу контенту у програмні продукти та цифрові сервіси.

Рекомендації щодо створення прикладних сервісів:

- інтеграція трансформерних моделей, зокрема BERT та його модифікацій, у різні інформаційні системи, що працюють з текстовими повідомленнями. Зокрема, вона може бути використана як окремий сервіс у складі системи кіберзахисту або як частина більш складної архітектури безпеки. Вона може бути інтегрована у поштові сервіси, системи фільтрації повідомлень, платформи обміну повідомленнями або інші інформаційні системи, що працюють з текстовими даними для автоматизованого фільтрування фішингових повідомлень;
- забезпечення регулярного донавчання моделей на актуальних наборах даних, що відображають нові тактики та сценарії атак соціальної інженерії. Це

забезпечує підвищену стійкість системи до еволюції методів соціальної інженерії та знижує залежність від ручного налаштування правил;

- застосування гібридних підходів, які поєднують методи машинного навчання з правилами інформаційної безпеки та експертними евристиками.

- використання сучасних програмних інструментів та розгортання моделі у хмарному середовищі. Це дозволяє обробляти великі обсяги текстових повідомлень у режимі реального часу та застосовувати технологію як для індивідуального захисту користувачів, так і для захисту інформаційних ресурсів на рівні організацій та сервісів.

Рекомендації щодо аналітичних інструментів:

- розробка інформаційних панелей для моніторингу кількості заблокованих повідомлень, типів атак та рівня ризику;

- впровадження механізмів зворотного зв’язку, що дозволяють користувачам повідомляти про підозрілий контент з метою покращення якості навчання моделей.

Перспективним напрямом розвитку систем протидії атакам соціальної інженерії є глибока інтеграція алгоритмів штучного інтелекту у масові цифрові платформи.

Для браузерів доцільним є:

- впровадження механізмів аналізу URL-адрес і контексту вебсторінок із використанням трансформерних моделей;

- автоматичне попередження користувачів перед введенням конфіденційних даних на підозрілих ресурсах.

Для месенджерів:

- застосування алгоритмів контекстного аналізу повідомлень з метою виявлення маніпулятивних формулювань і шкідливих посилань;

- відображення попереджень про потенційні ризики безпосередньо у діалогових вікнах.

Для поштових сервісів:

- використання моделей BERT або RoBERTa для інтелектуальної

фільтрації вхідної кореспонденції;

- автоматичне маркування та сортування підозрілих повідомлень для подальшого аналізу або підтвердження користувачем.

Результати дослідження підтверджують, що застосування трансформерних моделей істотно підвищує точність і швидкість виявлення атак соціальної інженерії. Водночас жодна автоматизована система не здатна повністю замінити усвідомлену та обережну поведінку користувачів.

У зв’язку з цим доцільним є комплексний підхід, що передбачає поєднання інтелектуальних систем виявлення атак із програмами підвищення цифрової грамотності населення. Проведення освітніх кампаній, тренінгів та інформаційних заходів сприяє зменшенню ефективності соціально-інженерних атак і підвищує загальний рівень кіберстійкості суспільства.

На основі проведених експериментів і аналізу результатів розроблена низка рекомендацій щодо застосування трансформерних моделей для протидії атакам соціальної інженерії. Для громадян важливо поєднувати поведінкові та технічні заходи безпеки. Для розробників і компаній ключовим є впровадження моделей штучного інтелекту у поштові сервіси, месенджери та браузери, а також регулярне донавчання на актуальних даних. Найефективніший захист досягається поєднанням високотехнологічних алгоритмів ШІ та підвищення цифрової обізнаності користувачів.

Висновки з розділу 3

Було розглянуто та експериментально обґрунтовано технологію протидії методам соціальної інженерії на основі трансформерної моделі штучного інтелекту BERT. Основну увагу зосереджено на автоматизованому виявленні фішингових електронних повідомлень як одного з найпоширеніших і найнебезпечніших інструментів соціально-інженерних атак у сучасному цифровому середовищі.

Детально описано технологію використання трансформерної моделі для

контекстного аналізу текстових повідомлень та побудови алгоритму їх класифікації.

Здійснено аналіз результатів практичної реалізації запропонованої технології. Оцінка якості класифікації за допомогою стандартних метрик машинного навчання (Accuracy, Precision, Recall, F1-score) підтвердила високу ефективність трансформерної моделі BERT у задачах виявлення фішингових повідомлень. Побудова матриці невідповідностей дозволила проаналізувати типи помилок класифікації та встановити, що модель демонструє здатність до узагальнення й коректної роботи з новими, раніше не баченими даними. Разом із тим було визначено основні переваги технології, а також її обмеження.

Надано практичні рекомендації щодо застосування технологій штучного інтелекту для захисту громадян і розробки прикладних засобів протидії атакам соціальної інженерії.

Таким чином, результати третього розділу підтверджують доцільність та ефективність використання трансформерних моделей штучного інтелекту для протидії атакам соціальної інженерії.

ВИСНОВКИ

У даній кваліфікаційній роботі досліджено проблему протидії методам соціальної інженерії в сучасному цифровому середовищі з використанням технологій штучного інтелекту. Актуальність обраної теми зумовлена стрімким зростанням кількості фішингових атак, їх ускладненням та адаптацією до поведінки користувачів, що значно знижує ефективність традиційних засобів захисту.

У першому розділі роботи проаналізовано теоретичні засади соціальної інженерії, розглянуто основні види атак та механізми психологічного впливу на користувачів. Проведено огляд сучасних методів протидії соціально-інженерним атакам і показано обмеженість класичних підходів.

У другому розділі досліджено сучасні підходи до використання штучного інтелекту в задачах кібербезпеки, зокрема методи машинного та глибинного навчання для виявлення фішингових повідомлень. Розглянуто архітектуру трансформерних моделей, принципи їх роботи та переваги порівняно з рекурентними та згортковими нейронними мережами. Особливу увагу приділено моделі BERT як інструменту контекстного аналізу тексту, що дозволяє враховувати семантичні зв’язки та приховані мовні патерни, характерні для атак соціальної інженерії.

У третьому розділі розроблено та реалізовано технологію автоматизованого виявлення фішингових повідомлень на основі трансформерної моделі BERT. Запропонована технологія реалізована у вигляді програмного конвеєра, який охоплює всі етапи обробки даних — від підготовки та токенізації текстових повідомлень до контекстного аналізу та класифікації. Проведено експериментальне дослідження з використанням реального набору електронних повідомлень, виконано тонке налаштування моделі та здійснено оцінку якості класифікації за допомогою стандартних метрик машинного навчання.

Результати експерименту підтвердили високу ефективність трансформерної моделі BERT у задачах виявлення фішингових повідомлень. Отримані значення показників Accuracy, Precision, Recall та F1-score свідчать про здатність моделі коректно розрізняти фішингові та легітимні повідомлення, а також узагальнювати мовні патерни на нові дані. Проведений аналіз помилок класифікації дозволив визначити основні переваги запропонованої технології та окреслити її обмеження, пов’язані з обчислювальними ресурсами та залежністю від якості навчальних даних.

На основі отриманих результатів сформульовано практичні рекомендації щодо застосування технологій штучного інтелекту для захисту громадян і розробки прикладних засобів протидії атакам соціальної інженерії. Показано, що найбільш ефективний захист досягається за умови комплексного підходу, який поєднує інтелектуальні системи автоматизованого виявлення фішингу з підвищенням рівня цифрової обізнаності користувачів.

Таким чином, у роботі досягнуто поставленої мети та виконано всі визначені завдання. Запропонована технологія на основі трансформерної моделі BERT є перспективним та практично значущим інструментом протидії методам соціальної інженерії й може бути використана для подальших наукових досліджень, а також для впровадження у сучасні інформаційні системи та сервіси кібербезпеки.

ПЕРЕЛІК ПОСИЛАНЬ

1. Hostragons Global Limited. Людський фактор у кібербезпеці: навчання та підвищення обізнаності співробітників. Hostragons.com, 28 вересня 2025 року. URL: https://www.hostragons.com/uk/блог/людський-фактор-у-кібербезпеці-навча/ (дата звернення: 10.12.2025)

2. Яковлев М., Любчак В. Можливості штучного інтелекту у виявленні та запобіганні фішингу й кібератакам. Кібербезпека: освіта, наука, техніка, 2025. URL: https://csecurity.kubg.edu.ua/index.php/journal/article/view/840 (дата звернення: 10.12.2025)

3. CERT-UA. Огляд інцидентів та тенденцій кіберзагроз / Державна служба спеціального зв'язку та захисту інформації України. – 2024. – URL: https://cert.gov.ua/ (дата звернення: 13.12.2025).

4. Statista. Causes of cyber breaches worldwide. – Statista, 2024. – URL: https://www.statista.com/statistics/1497591/causes-of-cyber-breaches-worldwide/ (дата звернення: 13.12.2025).

5. Verizon. *Data Breach Investigations Report 2024*. – Verizon, 2024. – URL: https://www.verizon.com/business/resources/reports/dbir/ (дата звернення: 13.12.2025).

6. IS Partners LLC. Human error cybersecurity statistics. – 2023. – URL: https://www.ispartnersllc.com/blog/human-error-cybersecurity-statistics/ (дата звернення: 13.12.2025).

7. Kaspersky Lab. *Global Cybersecurity Report 2022*. – Kaspersky, 2022. – URL: https://content.kaspersky-labs.com/fm/site-editor/e9/e96e21bf4a2f62ce46da33d53427f86e/source/nextreport.pdf (дата звернення: 13.12.2025).

8. Cyberly. What are Common Signs of a Social Engineering Attack? URL: https://www.cyberly.org/en/what-are-common-signs-of-a-social-engineering-attack/index.html (дата звернення: 13.12.2025)

9. Technology.org. How Can You Recognize Social Engineering Attacks? URL: https://www.technology.org/how-and-why/identify-social-engineering-attacks/ (дата звернення: 13.12.2025)

10. Data Type Statistics 2025. ZIPDO EDUCATION REPORT. URL: https://zipdo.co/data-type-statistics/ (дата звернення: 14.12.2025).

11. Structured Data vs. Unstructured Data in Machine Learning. upGrad Blog. URL: https://www.upgrad.com/blog/structured-vs-unstructured-data-in-machine-learning/ (дата звернення: 14.12.2025).

12. Abu-Nimeh S., Nappa D., Wang X., Nair S. A comparison of machine learning techniques for phishing detection // Proceedings of the Anti-Phishing Working Groups eCrime Researchers Summit. – 2007. – P. 60–69.

13. Hochreiter S., Schmidhuber J. Long short-term memory // Neural Computation. – 1997. – Vol. 9, No. 8. – P. 1735–1780.

14. Kim Y. Convolutional Neural Networks for Sentence Classification // EMNLP. – 2014. – P. 1746–1751.

15. Vaswani A., Shazeer N., Parmar N. et al. Attention Is All You Need // Advances in Neural Information Processing Systems. — Long Beach, USA, 2017.

16. Devlin J., Chang M.-W., Lee K., Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding // NAACL-HLT. — Minneapolis, USA, 2019.

17. Aggarwal C. C. Machine Learning for Text. — Cham : Springer International Publishing, 2018.

18. Yang Z., Liu Y., Sun M., et al. Phishing Email Detection Using BERT Model // IEEE Access. — 2020.

19. Sahingoz O. K., Buber E., Demir O., Diri B. Machine Learning Based Phishing Detection from URLs // Expert Systems with Applications. — 2019.

20. Birir Kipchirchir S., Wilfred Odoyo, AI-Based Phishing Attack Detection And Prevention Using Natural Language Processing (NLP), IC-ITECHS, 2024.

21. SecureNet: A Comparative Study of DeBERTa and Large Language

Models for Phishing Detection, arXiv:2406.06663, 2024.

22. Аналіз сучасних методів виявлення фішингових електронних листів / Н. Петляк, Я. Безкоровальний, Н. Купчик, «Herald of Khmelnytskyi National University», 2024.

23. Методи машинного навчання для антифрод-систем / K.Y. Ostrovska, V.O. Nosov, Системні технології, 2025.

24. Є. О. Фещенко і Т. М. Заболотня, Метод автоматизованого виявлення фішингу в електронних листах на основі гібридної нейромережевої архітектури, Наукові праці ВНТУ, вип.2, Чер 2025. URL: https://praci.vntu.edu.ua/index.php/praci/article/view/817 (дата звернення: 28.11.2025).