

Кваліфікаційна робота:

**Розробка системи сентиментального
аналізу для виявлення шкідливого
контенту в персональних
електронних комунікаціях**

Виконав:

студент гр. АМСЗІм-23-2

Шедін Д.А

Керівник роботи:

ст. викладач кафедри ІКІ ім. В.В. Поповського

Андрушко Д.В.

Об'єкт, предмет, мета та методи дослідження

Об'єкт дослідження – процес забезпечення інформаційної безпеки електронної комунікації.

Предмет дослідження – методи й засоби виявлення шкідливого контенту в електронних повідомленнях з використанням сентиментального аналізу та технологій штучного інтелекту.

Мета роботи – пошук шляхів розширення функціональності системи «Adressant» для аналізу домену відправника та заголовків листа шляхом впровадження моделі «MaliciousContentDetector» на основі сентиментального аналізу з метою виявлення шкідливого контенту в персональних електронних комунікаціях.

Методи дослідження – аналіз, моделювання, самооцінка, статистична обробка результатів.

Актуальність роботи

1) Кількість користувачів електронної пошти щорічно зростає. У 2023 році близько 45,6 % електронних листів містили шкідливий контент.

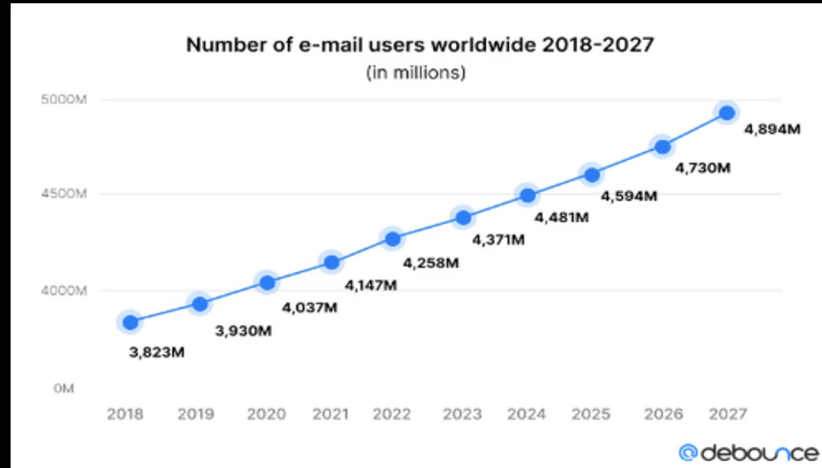


Рисунок 1 – Кількість користувачів електронної пошти (дані Debounce.io)

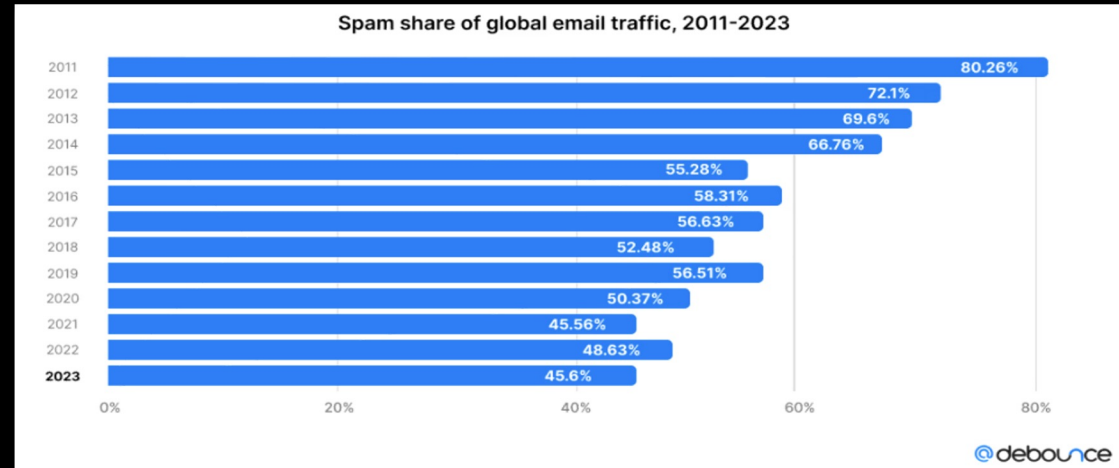


Рисунок 2 – Відсоток спаму серед загальної кількості електронних листів (дані Debounce.io)

2) Обмеження «Adressant 1.0» (аналіз лише доменів і заголовків листів, обмежена підтримка браузерів та поштових клієнтів, відсутність аналізу вмісту та адаптації до нових загроз, орієнтація виключно на електронну пошту).

3) Відсутність готових рішень спрямованих на перевірку безпечності саме україномовних текстів.

4) Поширення інформаційних загроз не тільки електронною поштою, а й іншими каналами комунікацій в Інтернеті (соціальними мережами, месенджерами, форумами тощо).

Методи сентиментального аналізу тексту

Лексичні

Кожному слову або фразі присвоюється певний емоційний бал – позитивний чи негативний. Під час аналізу тексту відбувається їх пошук та обчислення загального емоційного балу для всього тексту:

$$\text{Сентиментальний бал} = \sum_{w \in \text{text}} \text{score}(w), \quad (1)$$

де $\text{score}(w)$ – сентиментальний бал слова w , отриманий із лексикону.

```
urgency_terms = ['терміново', 'негайно', 'важливо', 'поспішайте', 'обмежено', 'швидко',
                 'urgent', 'immediately', 'important', 'hurry', 'limited', 'quickly']
urgency_count = sum(1 for term in urgency_terms if term.lower() in text.lower())

cta_terms = ['натисніть', 'перейдіть', 'заповніть', 'реєструйтесь', 'введіть', 'поділіться',
             'click', 'follow', 'fill', 'register', 'enter', 'share']
cta_count = sum(1 for term in cta_terms if term.lower() in text.lower())

reward_terms = ['виграш', 'приз', 'бонус', 'безкоштовно', 'шанс', 'ексклюзивно',
                'win', 'prize', 'bonus', 'free', 'chance', 'exclusive']
threat_terms = ['заблоковано', 'втрата', 'ризик', 'проблема', 'загроза', 'небезпека',
                'blocked', 'loss', 'risk', 'problem', 'threat', 'danger']

reward_count = sum(1 for term in reward_terms if term.lower() in text.lower())
threat_count = sum(1 for term in threat_terms if term.lower() in text.lower())
```

Рисунок 3 – Зняток екрану програмної реалізації коду для встановлення лексичних списків

Методи сентиментального аналізу тексту

З використанням TF-IDF ознак та випадкового лісу

TF-IDF (Term Frequency - Inverse Document Frequency) використовується для числового представлення важливості слів у тексті. Більшу вагу TF-IDF здобудуть слова з високою частотою виникнення в межах документа та низькою частотою вживання в інших документах колекції.

Формула обчислення:

$$\text{TF-IDF} = \text{TF} \cdot \text{IDF} = \frac{n_i}{\sum_k n_k} \cdot \log \frac{|D|}{|(d_i \supset t_i)|}, \quad (2)$$

де n_i – кількість входжень слова в документ;

$\sum_k n_k$ – загальна кількість слів в документі;

$|D|$ – кількість документів колекції;

$(d_i \supset t_i)$ – кількість документів, де зустрічається слово t_i (при $n_i \neq 0$).

Random forester поєднує результати кількох дерев рішень для отримання єдиного результату. Фінальне рішення щодо класифікації тексту як «шкідливого» (1) або «безпечного» (0) приймається шляхом голосування дерев рішень.

```
X_features = self.extract_features(all_texts)
self.tfidf = TfidfVectorizer(max_features=1000, ngram_range=(1, 2))
X_tfidf = self.tfidf.fit_transform(all_texts)
X_tfidf_df = pd.DataFrame(X_tfidf.toarray(), columns=self.tfidf.get_feature_names_out())
X = pd.concat([X_features, X_tfidf_df], axis=1)
self.feature_names = X.columns.tolist()
X_train, X_test, y_train, y_test = (
    train_test_split(*arrays: X, all_labels, test_size=0.2, random_state=42))

self.clf = RandomForestClassifier(n_estimators=100, random_state=42)
self.clf.fit(X_train, y_train)
y_pred = self.clf.predict(X_test)
print(classification_report(y_test, y_pred))
print(confusion_matrix(y_test, y_pred))
return self.clf
```

Рисунок 4 – Зняток екрану програмної реалізації коду для створення TF-IDF ознак та класифікатора випадкового лісу в моделі «MaliciousContentDetector»

Методи сентиментального аналізу тексту

З використанням deep learning

У моделі «MaliciousContentDetector» для системи «Adressant 2.0» використовується попередньо навчена BERT-модель «nlptown/bert-base-multilingual-uncased-sentiment» для сентиментального аналізу.

```
class MaliciousContentDetector: 7 usages  ⤴ Shedin *  
    def __init__(self, model_name="sentiment-malicious-detector"):  ⤴ Shedin *  
        self.model_name = model_name  
        self.sentiment_tokenizer = AutoTokenizer.from_pretrained("nlptown/bert-base-multilingual-uncased-sentiment")  
        self.sentiment_model = AutoModelForSequenceClassification.from_pretrained(  
            "nlptown/bert-base-multilingual-uncased-sentiment")  
        self.clf = None  
        self.tfidf = None  
        self.feature_names = None
```

Рисунок 5 – Зняток екрану програмної реалізації коду для ініціалізації моделі «MaliciousContentDetector» з використанням попередньо навченої BERT-моделі

```
encoded_text = self.sentiment_tokenizer(text, return_tensors='pt', truncation=True, max_length=512)  
with torch.no_grad():  
    output = self.sentiment_model(**encoded_text)  
  
sentiment_scores = softmax(output.logits, dim=1).numpy()[0]
```

Рисунок 6 – Зняток екрану програмної реалізації коду для обробки тексту та отримання ймовірностей сентименту

Методи сентиментального аналізу тексту

Гібридний метод (комбінація лексичних методів, машинного навчання та глибинного навчання)

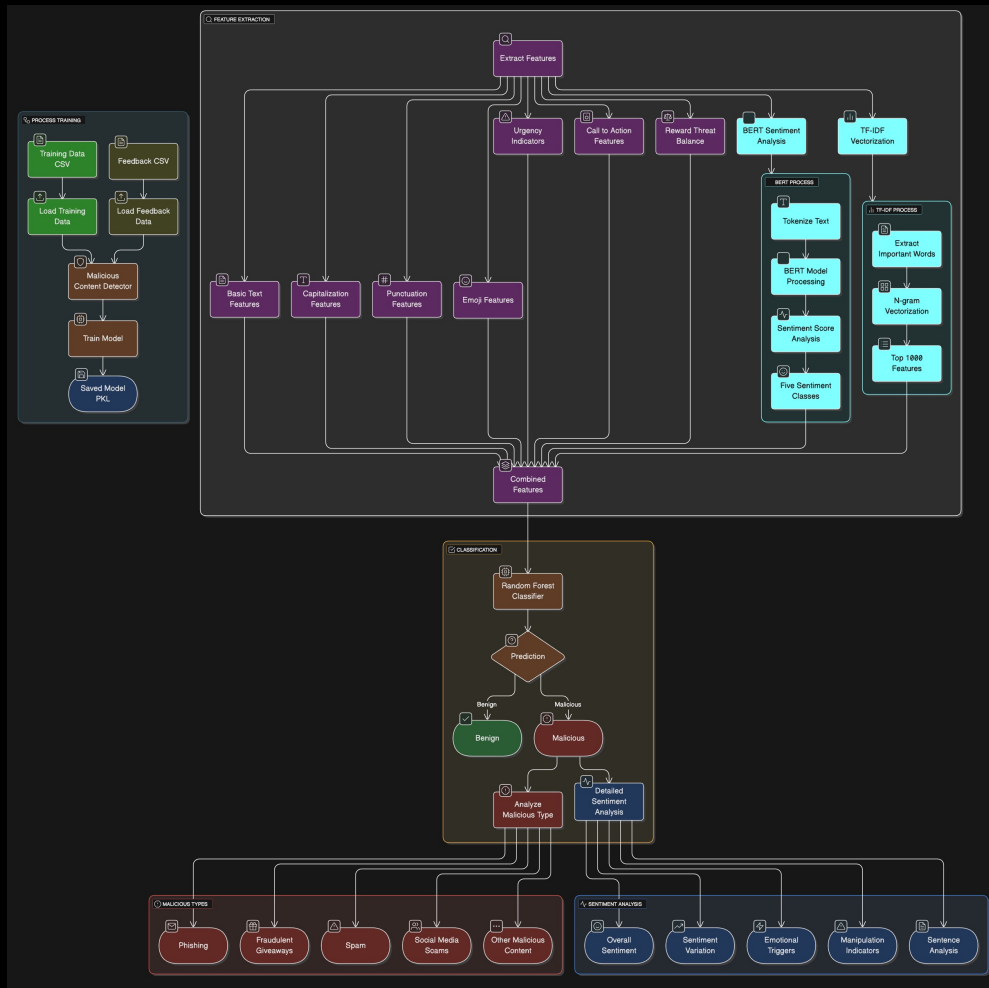


Рисунок 7 – Загальна архітектура моделі з використанням гібридного методу

У моделі «MaliciousContentDetector» гібридний підхід реалізовано через об'єднання лексичних, TF-IDF і сентиментальних ознак у єдиний вектор, який обробляється класифікатором «RandomForestClassifier».

Лексичні ознаки включають підрахунок довжини тексту, кількості слів, середньої довжини слова, співвідношення великих літер, кількості знаків питання та оклику, кількості емодзі, а також тригерів.

Далі створюються TF-IDF ознаки за допомогою «TfidfVectorizer» із бібліотеки «scikit-learn», що дозволяє представляти текст у вигляді числових векторів на основі частоти слів та їхньої рідкості в корпусі.

Крім того, додається аналіз сентименту за допомогою попередньо натренованої моделі BERT, яка враховує контекст слів і забезпечує глибоке розуміння семантики тексту.

Класифікатор «RandomForestClassifier» інтегрує всі об'єднані ознаки для точного визначення шкідливого контенту.

Така комбінація забезпечує високу точність класифікації та дозволяє ефективно виявляти шкідливий контент у реальних умовах.

Оцінка точності моделі та порівняння методів

Для оцінки продуктивності моделі «MaliciousContentDetector» застосовуються стандартні метрики, які базуються на матриці помилок. Вона включає кількість true positives, true negatives, false positives, false negatives прогнозів. На основі цієї матриці помилок обчислюються чотири основні метрики:

$$\text{Precision} = \frac{TP}{TP+FP} \quad (3)$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad (4)$$

$$\text{F1-Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (5)$$

де TP – кількість правильно передбачених шкідливих текстів;

FP – кількість помилково передбачених шкідливих текстів;

FN – кількість шкідливих текстів, пропущених моделлю.

Метод	Прецизійність	Повнота	F1-міра	Точність
Лексичний	1	0,153846	0,266667	0,414449
Випадковий ліс (TF-IDF)	0,948718	0,973684	0,961039	0,943396
BERT	0,780749	0,802198	0,791328	0,707224
Гібридний	1	0,95	0,974359	0,962264

Таблиця 1 – Порівняння продуктивності різних методів
сентиментального аналізу

Навчання моделі

Дані для навчання моделі беруться з двох файлів:

- «training_data.csv» – створений на основі власної оцінки отриманих поштових листів, а також публікацій у популярних «Telegram» каналах України;
- «feedback_data.csv» – отримується на основі зворотного зв'язку від користувачів.

training_data.csv		feedback_data.csv				
c1	c2	c1	c2	c3	c4	c5
text	label	timestamp	analyzed_text	actual_prediction	is_valid	is_feedback_valid
ТЕРМІНОВО! Ваш банківський рахунок заблоковано через підозрілу активність!	1	2025-05-09T02:05:53.485276	Доброго дня! Надсилаємо скан-копію дипло...	False	True	True
ВАУ! ВІТАЄМО! Ви стали 1000-м відвідувачем нашого сайту і виграли НОВИЙ ТЕЛЕФОН!	1	2025-05-09T02:06:06.717491	Доброго дня! Повинен бути, але питанням ...	False	True	True
РЕВОЛЮЦІЙНИЙ АНТИВІРУС! ЗАХИСТІТЬ свій комп'ютер від UCIX атак СЬОГОДНІ!	1	2025-05-09T02:06:15.061051	Вітаємо. NBSPДляNBSPзаовленняNBSPдиплома п	False	True	True
Ексклюзивна інформація! Криптовалюта XYZ зросте на 500% за тиждень! Я вже заробив 50 00...	1	2025-05-09T02:06:33.847518	Шановні учасники виставки!	True	False	True
Увага! Ваша картка заблокована! Підтвердіть особу за посиланням або втратите гроші!	1	2025-05-09T02:06:55.482343	Добрий вечір Дмитро!	False	True	True
ШОК! Лікарі приховують правду! Цей засіб лікує всі хвороби за 3 дні!	1	2025-05-09T02:07:03.311216	Шановні учасники виставки!	True	True	False
Увага! Останній день розпродажу! Знижки до 90%! Кількість товарів обмежена!	1	2025-05-09T02:14:09.071849	Вітаю!	True	True	True
ВАЖЛИВО! Ваш обліковий запис буде видалено через 24 години! Підтвердіть дані зараз!	1	2025-05-09T02:17:01.612151	Ваше замовлення 109013 Очікує на вас в Y...	True	False	True
Секретна інформація! Приєднуйтеся до закритої групи інвесторів і заробляйте мільйони!	1	2025-05-09T02:17:49.726817	INTERTOP ШОПІНГ ФЕСТ: - 20% на весняні к...	True	True	True
Привіт! Як справи? Давно не бачились, може зустрінемося на вихідних?	0	2025-05-09T02:19:27.780953	Ви отримуєте це повідомлення, так як під...	False	False	True
Дякую за замовлення! Ваш товар буде доставлено протягом трьох робочих днів.	0	2025-05-09T02:28:23.222034	США закликають Україну та рф до 30-денно...	True	False	True
Нагадуємо про зустріч завтра о 10:00. Будь ласка, підготуйте звіт для обговорення.	0	2025-05-09T02:32:42.653008	Зеленський повідомив Трампа, що Україна ...	False	True	True

Рисунок 8 – Зміст файлів «training_data.csv» та «feedback_data.csv»

Розгортання моделі та її інтеграція до системи «Adressant»

Модель «MaliciousContentDetector» інтегровано до системи «Adressant 2.0» через мікросервісну архітектуру, розгорнуту в хмарі Microsoft Azure

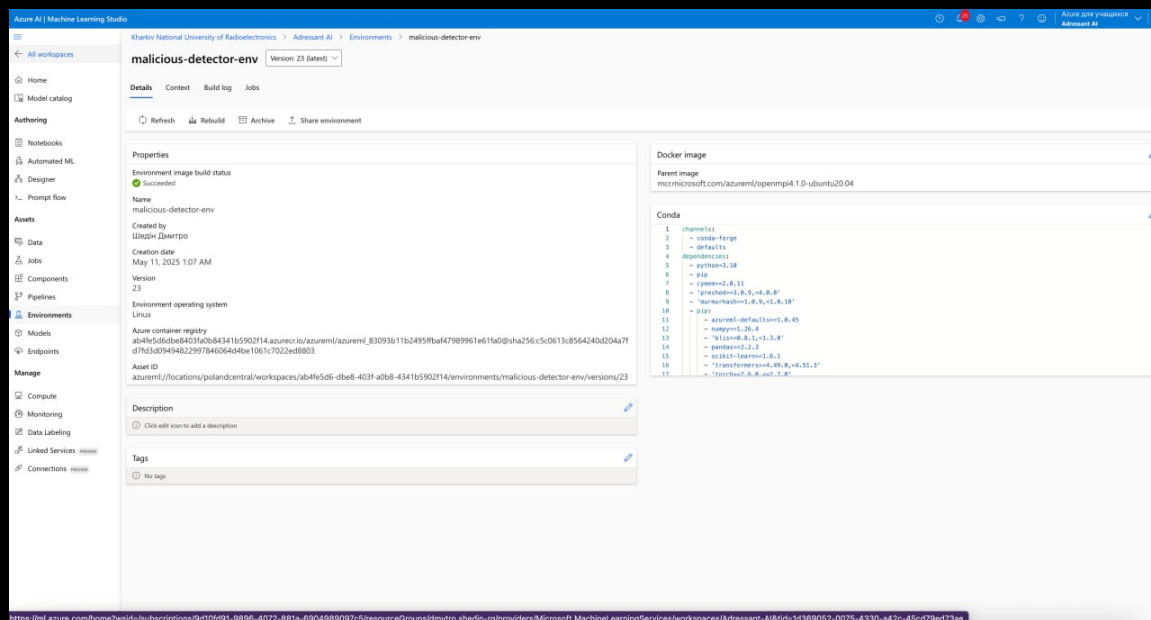


Рисунок 9 – Розгорнуте в «Azure AI | Machine Learning Studio» середовище моделі

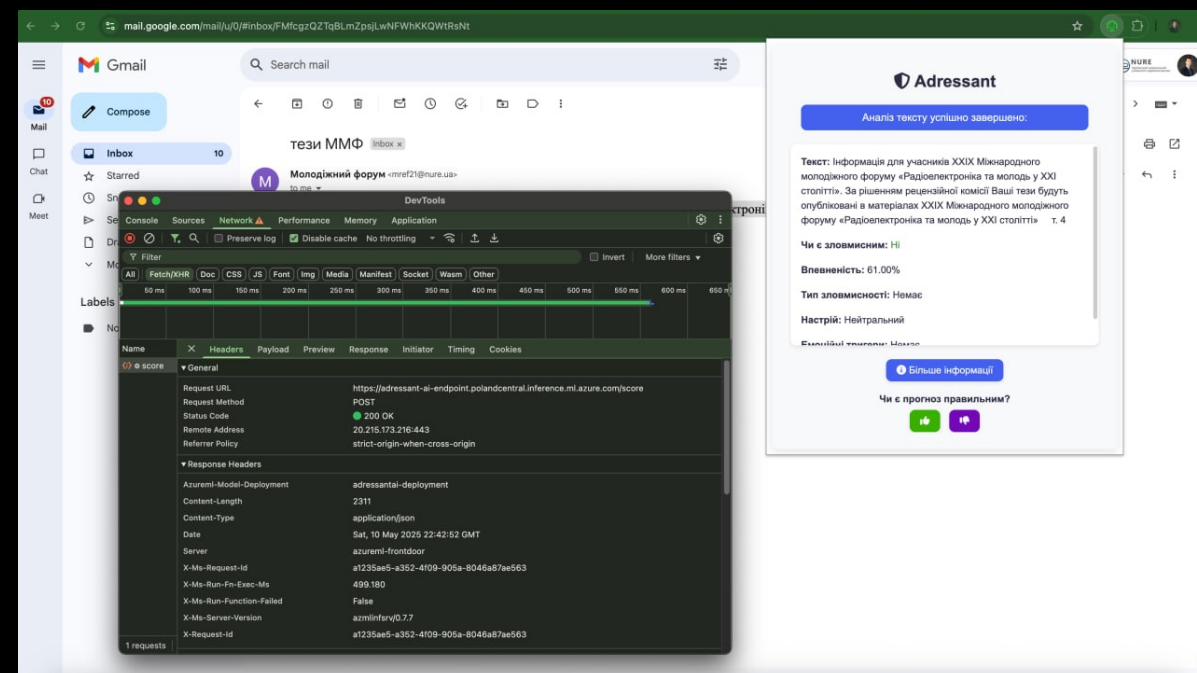


Рисунок 10 – Тестування моделі через розширення для браузера «Adressant»

Тестування моделі для електронної пошти

Функціонал «Adressant 1.0» обмежувався аналізом домену відправника та заголовків листа, що не дозволяло виявляти шкідливий контент у самому тексті повідомлення. Модель «MaliciousContentDetector» виправляє це.

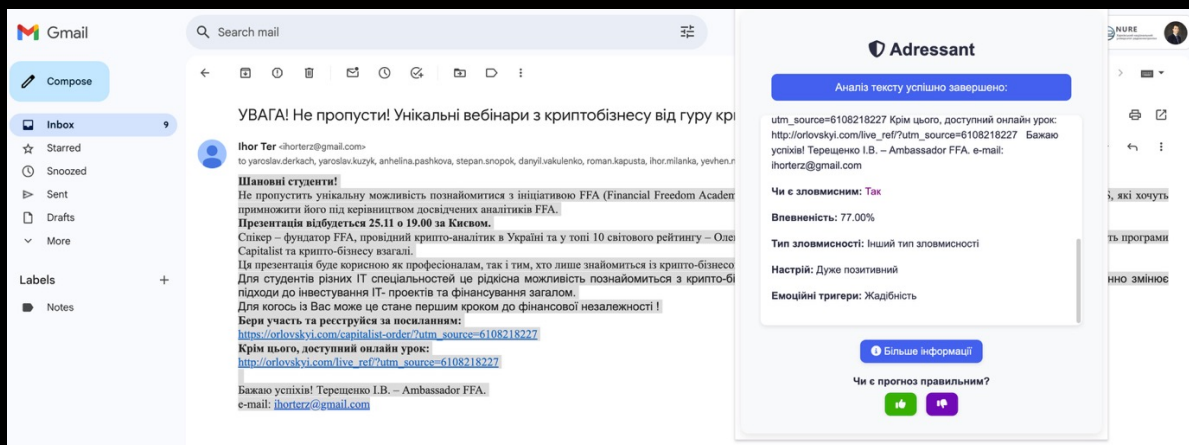


Рисунок 11 – Приклад повідомлення з легітимного домену, але зі шкідливим змістом

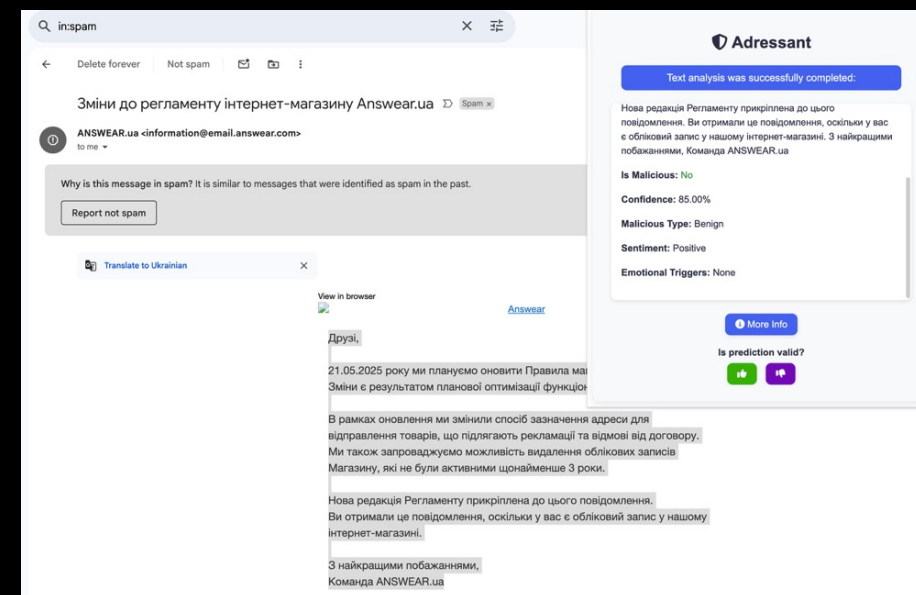


Рисунок 12 – Приклад повідомлення з потенційно небезпечного домену, але з безпечним змістом

Тестування моделі в соціальній мережі

Для тестування моделі було обрано популярний український канал у «Telegram» та його повідомлення за 5 днів.

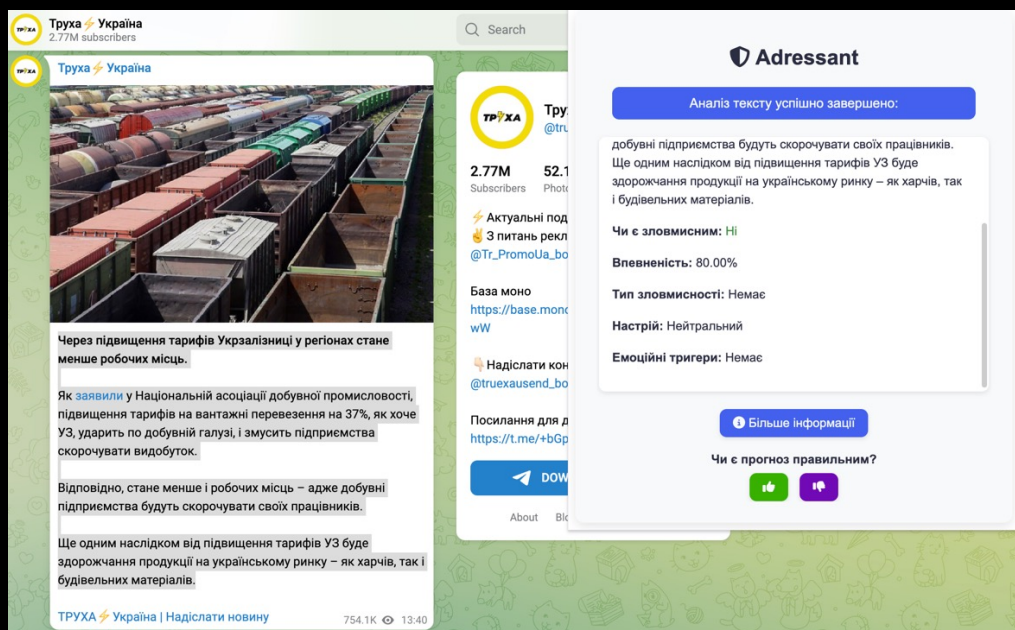


Рисунок 13 – Приклад виявленого безпечного повідомлення

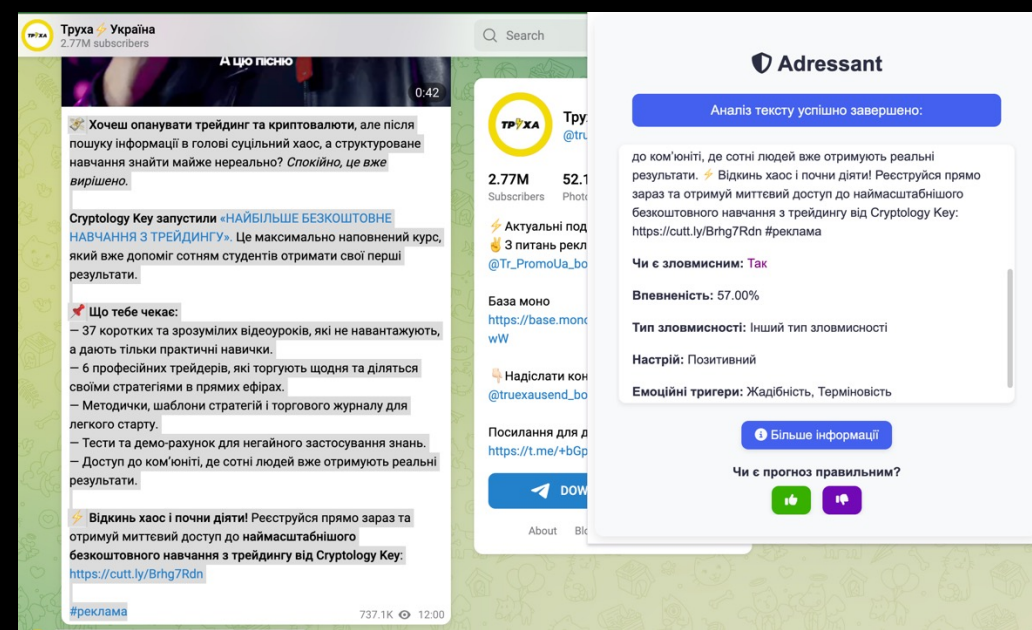


Рисунок 14 – Приклад виявленого шкідливого повідомлення

Тестування моделі в соціальній мережі

Показник	Значення
Загальна кількість повідомлень	162
Шкідливі повідомлення	9 (5,6 %)
Хибнопозитивні результати	5 (3,1 %)
Хибнонегативні результати	6 (3,7 %)
Середня впевненість	0,84
Середній час відповіді (мс)	256,57
Мінімальний час відповіді (мс)	70,34
Максимальний час відповіді (мс)	884,54
Загальна кількість виявлених тригерів	10
Середня кількість тригерів на повідомлення	0,06
Повідомлення з високою щільністю тригерів ($>0,1$)	8 (4,9 %)
Повідомлення з контрастом емоцій	61 (37,7 %)
Повідомлення з високою впевненістю ($>0,9$)	56 (34,6 %)
Середня кількість шкідливих повідомлень за день	1,8

Таблиця 2 – Статистичні показники тестування у «Telegram» каналі моделі «MaliciousContentDetector»

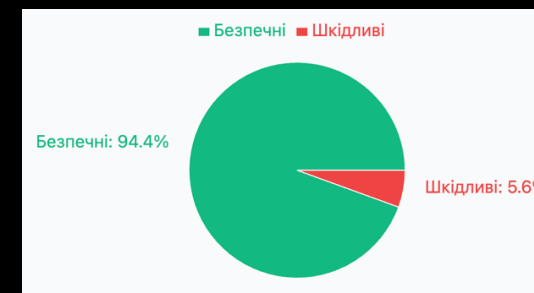


Рисунок 15 – Відношення безпечних повідомлень до шкідливих

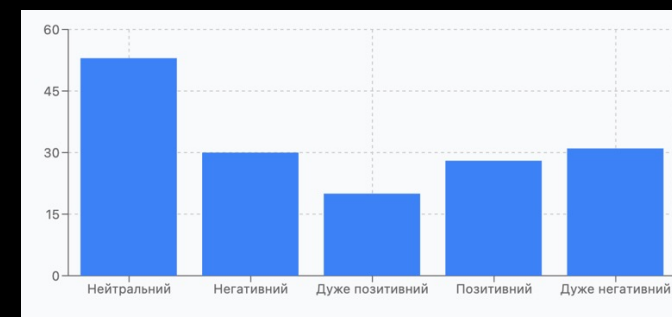


Рисунок 16 – Кількість повідомлень за настроєм

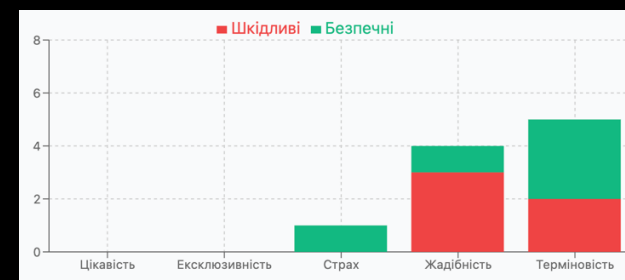


Рисунок 17 – Емоційні тригери в шкідливих та безпечних повідомленнях

Висновки

Основні результати роботи

- розроблено та впроваджено модель штучного інтелекту «MaliciousContentDetector» для виявлення шкідливого контенту в електронних комунікаціях на основі сентиментального аналізу з використанням гібридного підходу;
- інтегровано модель до системи «Adressant 2.0». Подолано обмеження попередньої версії;
- досягнуто високих показників ефективності як для навчальних даних, так і при тестуванні на електронній пошті та в соціальній мережі.

Наукова новизна

практично показано ефективність використання гібридного підходу сентиментального аналізу для перевірки україномовних текстів, який поєднує лексичні методи, TF-IDF та BERT для виявлення маніпулятивних шаблонів із врахуванням лінгвістичних особливостей української мови.

Висновки

Практична значимість

- підвищення безпеки електронних комунікацій через виявлення фішингу, спаму, шахрайства тощо;
- універсальність застосування – аналіз текстів з соціальних мереж, новинних сайтів, форумів, месенджерів тощо;
- можливість використання API для інтеграції зі зовнішніми системами;
- великий набір текстових даних, які можна використовувати в інших комерційних та некомерційних рішеннях.

Публікації

Окремі результати роботи доповідались на 9-й Міжнародній науково-технічній конференції «Інформаційно-комунікаційні технології та кібербезпека (ІКТК-2023)» (секція «Кібербезпека та захист інформації»), а також на 12-й Міжнародній науково-технічній конференції «Інформатика, управління та штучний інтелект (ІУШІ-2025)».

Ба більше, були здобуті призові місця за доповіді на секції «Управління інформаційною безпекою» 29-го Міжнародного молодіжного форуму «Радіoeлектроніка та молодь у ХХІ столітті» та виставці форуму в номінації «Програмне забезпечення. Програми».