КІБЕРБЕЗПЕКА: освіта, наука, техніка
CYBERSECURITY: EDUCATION, SCIENCE, TECHNIQUE
№ 2 (30), 2025
Спеціальний випуск ISSN 2663 – 4023
Special issue
DOI 10.28925/2663-4023.2025.30.921
УДК 004.738.5:004.056

Андрушко Дмитро Володимирович
к.т.н., старший викладач кафедри інфокомунікаційної інженерії імені В.В. Поповського
Харківський національний університет радіоелектроніки, Харків, Україна
ORCID: 0009-0004-0749-9729
dmytro.andrushko@nure.ua

Шедін Дмитро Андрійович
магістр кафедри інфокомунікаційної інженерії імені В.В. Поповського
Харківський національний університет радіоелектроніки, Харків, Україна
ORCID: 0009-0007-4597-5871
dmytro.shedin@nure.ua

Чакрян Вадим Хазарович
к.т.н., старший викладач кафедри інфокомунікаційної інженерії імені В.В. Поповського
Харківський національний університет радіоелектроніки, Харків, Україна
ORCID: 0000-0003-4827-9779
vadym.chakrian@nure.ua

МЕТОДИКА ВИЯВЛЕННЯ ШКІДЛИВОГО КОНТЕНТУ ЗА ДОПОМОГОЮ СЕНТИМЕНТАЛЬНОГО АНАЛІЗУ

Анотація. У статті досліджується проблема виявлення шкідливого контенту в персональних електронних комунікаціях, таких як електронна пошта, повідомлення в соціальних мережах або месенджерах. Сучасні системи фільтрації шкідливого контенту переважно базуються на аналізі ключових слів, структурних особливостей повідомлень та технічних параметрів, таких як IP-адреси чи домени походження повідомлень. Такі системи мають обмежену ефективність проти сучасних кіберзагроз, що використовують контент повідомлень з емоційними маніпуляціями чи фейковою інформацією. Крім того, ефективність таких підходів є обмеженою, особливо коли йдеться про цільові атаки та новітні методи обходу захисних механізмів, що постійно еволюціонують. В роботі запропоновано гібридну модель «MaliciousContentDetector», яка поєднує лексичні методи, методи машинного навчання (TF-IDF, Random Forest) та глибинне навчання (BERT) для аналізу сентименту тексту. Модель враховує лінгвістичні особливості української мови та контекстуальні емоційні тригери. Також запропоновану модель реалізовано у вигляді програмного забезпечення з можливістю інтеграції в браузери та інші застосунки для підвищення кіберзахищеності кінцевих користувачів. Розроблену модель було протестовано на базі реальних повідомлень в одній із соціальних мереж на прикладі популярного новинного каналу. В ході аналізу 162 дописів модель «MaliciousContentDetector» показала високу ефективність та точність виявлення потенційно шкідливого текстового контенту.

Ключові слова: сентиментальний аналіз, шкідливий контент, кібербезпека, штучний інтелект, BERT, електронні комунікації, емоційні тригери.

ВСТУП

У сучасному світі електронні комунікації та обмін текстовими повідомленнями стали невід’ємною частиною повсякденного життя, бізнес-процесів і соціальних взаємодій. За даними на 2025 рік, електронною поштою користується близько 4,6 мільярда людей, а частка шкідливих повідомлень сягає 45,6% [1]. Користувачі щодня стикаються з різноманітними формами шкідливого контенту: фішинговими атаками,

фейковими розіграшами, спам-повідомленнями, маніпуляціями та соціальною інженерією. Ці загрози призводять до фінансових втрат, витоку конфіденційної інформації, поширення дезінформації та інших негативних наслідків.

Традиційні системи фільтрації шкідливого контенту базуються переважно на аналізі ключових слів, структурних особливостей повідомлень та технічних параметрів, таких як IP-адреси, домени чи заголовки. Однак ефективність таких підходів обмежена, особливо щодо цільових атак, де зловмисники використовують легітимні домени, скомпрометовані облікові записи або різноманітні методи обходу захисних механізмів. Крім того, сучасні кібератаки часто застосовують емоційні аспекти, такі як страх, жадібність чи терміновість, що не виявляються стандартними методами.

Сентиментальний аналіз, як метод обробки природної мови (NLP), дозволяє виявляти суб'єктивні характеристики тексту – емоційне забарвлення, оцінки, установки та наміри автора [2]. Його застосування для виявлення шкідливого контенту у тексті може значно покращити точність систем кібербезпеки, доповнюючи технічний аналіз емоційним і контекстуальним. Це особливо актуально для україномовного контенту, де лінгвістичні особливості (варіативність словотворення, багата морфологія, культурний контекст) ускладнюють автоматизований аналіз текстових повідомлень.

Аналіз останніх досліджень і публікацій. Проблема виявлення шкідливого контенту в електронних комунікаціях активно досліджується. У роботі [2] описуються основні методи сентиментального аналізу: лексичні (за словниками, наприклад, «SentiWordNet»), методи машинного навчання (Naive Bayes, SVM, Random Forest) і гібридні підходи. Крім того, у роботі зазначається, що гібридні моделі зазвичай точніші за моделі, що використовують лише один підхід для аналізу тексту. При цьому фокус у роботі [2] та схожих працях зазвичай на аналізі англійськомовних текстів. Проте, проблема автоматичного аналізу українських текстів не є новою та досліджувалася в роботах Н. Дарчук, наприклад [4]. Схожі проблеми аналізу української мови та україномовних текстів розглянуті у роботах [5] та [6]. Таким чином, ці роботи демонструють доцільність використання методів штучного інтелекту (ШІ) та машинного навчання для класифікації та аналізу текстових повідомлень, що дозволяє підвищити кіберзахист кінцевих користувачів.

На сьогодні існує багато систем, які використовують ШІ для захисту кінцевих користувачів від кіберзагроз. Прикладами таких є:

- CrowdStrike Falcon: платформа захищає комп'ютери й мережі, використовуючи ШІ для аналізу поведінки користувачів. Вона може знаходити підозрілі дії, наприклад, спроби фішингу чи зараження вірусами [11];
- Darktrace: система діє як «імунна система» для мережі, шукаючи незвичайну поведінку. Вона не залежить від мови, бо аналізує не текст, а дії в мережі. Це робить її ефективною для великих компаній, де потрібен захист від прихованих загроз [12];
- Microsoft Security Copilot: інструмент, що використовує ШІ для швидкого реагувати на загрози. Він може обробляти запити різними мовами завдяки NLP. Система аналізує величезну кількість даних щодня, що робить її корисною для компаній [13];
- Sift: платформа захищає від шахрайства, наприклад, під час онлайн-покупок [14].

Більшість наведених вище та інших існуючих рішень, підтримують багатомовність, але найкраще працюють саме з англійською мовою.

Мета статті. Незважаючи на значний розвиток галузі штучного інтелекту в цілому, відповідних моделей та алгоритмів, практичне використання їх для аналізу шкідливого текстового контенту та кіберзахисту стикається з обмеженнями на кшталт:

- фокус на англійській мові та ігнорування україномовних патернів (наприклад, словозміни, емоційні конструкції);
- відсутність аналізу вмісту повідомлень, обмеження технічними параметрами;
- брак адаптації до нових загроз через статичний характер моделей;
- обмежена інтеграція з програмними застосунками (браузери, SIEM тощо);
- недостатнє тестування на реальних даних та невелика кількість україномовних тестових наборів даних (data set).

Тому метою цієї роботи є створення гібридної моделі «MaliciousContentDetector», яка комбінує лексичні методи, машинне навчання та глибинне навчання, що дозволяє аналізувати україномовні тексти та виявляти потенційно шкідливий контент. Крім того, у результаті дослідження створена програмна реалізація системи сентиментального аналізу для виявлення шкідливого текстового контенту в персональних електронних комунікаціях, таких як електрона пошта, соціальні мережі, тощо. Для розробленої моделі проведено кількісний та якісний аналіз ефективності розпізнавання шкідливого контенту.

РЕЗУЛЬТАТИ ДОСЛІДЖЕННЯ

Існують різноманітні методи для визначення шкідливого текстового контенту. Їх можна умовно поділити на лексичні підходи, методи машинного навчання, методи глибинного навчання та гібридні.

Лексичні методи базуються на словниках ключових слів і фраз, що вказують на маніпуляції (наприклад, «терміново», «ваш рахунок заблоковано»). Вони швидкі, але не враховують контекст повідомлення в цілому. 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. Математично це можна виразити як:

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

де *score(w)* – сентиментальний бал слова *w*, отриманий із лексикону.

Основними перевагами лексичних методів є їх простота реалізації та інтерпретованість результатів. Водночас недоліками є неможливість врахування контексту та складних лінгвістичних конструкцій, таких як іронія, сарказм або заперечення. На рис. 1 наведено приклад задання лексичних списків для розробленої моделі «MaliciousContentDetector».

```
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', 'зареєструйтесь', '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())
```

Рис. 1. Зняток екрану програмної реалізації коду для встановлення лексичних списків для аналізу шкідливого контенту

На відміну від лексичних, методи на основі машинного навчання дозволяють автоматично виявляти закономірності в даних без попереднього визначення правил. У такому методі текстові повідомлення необхідно представити в числовому вигляді. Для цього застосовуються підходи вилучення ознак, такі як TF-IDF [2,3]. TF або частота слова – це відношення кількості входжень вибраного слова до загальної кількості слів у тексті:

$$TF = \frac{n_i}{\sum_k n_k}, \quad (2)$$

де n_i– кількість входжень слова в документ, $\sum_k n_k$– загальна кількість слів в тексті.

Ознака IDF або обернена частота документа – це інверсія частоти, з якою слово трапляється в документах колекції:

$$IDF = log \frac{|D|}{|d_i \supset t_i|} \quad (3)$$

де $|D|$ – кількість документів колекції;

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

Загальний же показник TF-IDF визначається як добуток *TF* (2) та *IDF* (3):

$$TF\text{-}IDF = TF \cdot IDF \quad (4)$$

Розроблена модель реалізує гібридний підхід через об’єднання лексичних, TF-IDF і сентиментальних ознак у єдиний вектор, який обробляється потім обробляється класифікатором. Також у моделі «MaliciousContentDetector» використовується попередньо навчена BERT-модель «nlptown/bert-base-multilingual-uncased-sentiment» для сентиментального аналізу. Вона підтримує багатомовний аналіз, включаючи українську мову. Лексичні ознаки включають підрахунок таких характеристик, як довжина тексту, кількість слів, середня довжина слова, співвідношення великих літер, кількість знаків питання та оклику, кількість емодзі, а також тригери, такі як терміновість, заклики до дії, винагорода, загрози, та співвідношення між кількістю термінів винагороди й загроз.

У програмній реалізації моделі «MaliciousContentDetector» після створення TF-IDF ознак відбувається їх комбінація з іншими характеристикими тексту, такими як довжина, кількість великих літер чи емоційних тригерів. В результаті за допомогою «TfidfVectorizer» з бібліотеки «scikit-learn» [10] формується векторне представлення тексту. Наостанок усі ознаки об’єднуються в єдиний вектор шляхом застосування методу «concat()» бібліотеки «pandas» для лексичних ознак, TF-IDF векторів і ймовірностей від BERT, які додаються як числові значення для п’яти класів емоційного забарвлення. Отриманий вектор передається до «RandomForestClassifier» класифікатора [10], для фінальної класифікації на основі всіх вилучених ознак. Для параметрів «TfidfVectorizer» було обрано наступні значення:

- «*max_features=1000*»: обмеження кількості ознак до 1000 найважливіших слів чи біграм (пар послідовних слів) для зменшення обчислювальної складності та уникнення перенавчання. Обране значення – це компроміс між точністю і швидкодією, підтверджений експериментально;
- «*ngram_range=(1, 2)*»: дозволяє враховувати не лише окремі слова, але й біграми, що дає змогу частково враховувати контекст та покращує якість класифікації. Це

особливо важливо для виявлення шаблонів шкідливого контенту, де певні послідовності слів можуть бути більш показовими, ніж окремі лексеми.

Такий гібридний підхід дозволяє моделі виявляти як явні маніпулятивні патерни (наприклад, фразу «терміново натисніть»), так і складні контекстуальні сигнали (наприклад, контрастність сентименту між реченнями), які є потенційними ознаками шкідливого контенту. Перевагою цього підходу є також його здатність адаптуватися до різних типів текстів.

Для оцінки ефективності розробленої моделі «MaliciousContentDetector» та її програмної реалізації, застосовуються стандартні метрики бінарної класифікації, які базуються на матриці помилок. Вона включає такі характеристики:

- Кількість правильно класифікованих шкідливих текстів (true positives);
- Кількість правильно класифікованих безпечних текстів (true negatives);
- Кількість безпечних текстів, помилково класифікованих як шкідливі (false negatives);
- Кількість шкідливих текстів, помилково класифікованих як безпечні.

Розроблену модель було протестована для текстових повідомлень у одній із соціальній мереж на прикладі популярного новинного каналу. За допомогою моделі «MaliciousContentDetector» було проаналізовано 162 дописів. На рис. 2 наведено статистичний аналіз тестування моделі на базі дописів новинного каналу.

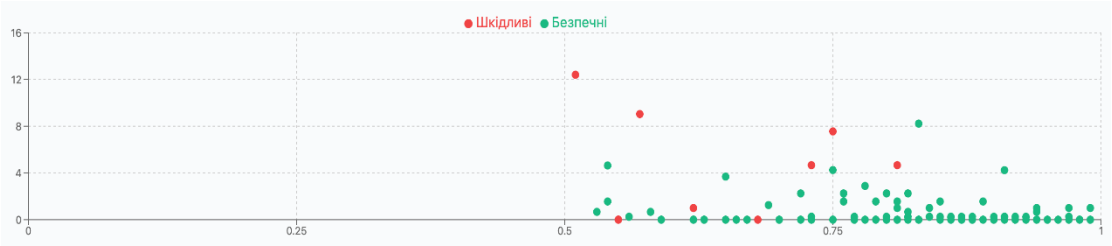

Рис. 2. Розподіл між впевненістю передбачення та варіацією сентименту

У таблиці 1 наведено підсумкові статистичні результати тестування розробленої моделі «MaliciousContentDetector» на реальних текстових повідомленнях новинного каналу.

Таблиця 1

Статистичні показники тестування у «Telegram» каналі моделі «MaliciousContentDetector»

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

Як видно із таблиці, гібридний підхід, який поєднує лексичні методи, TF-IDF і BERT, забезпечує високу середню впевненість (0,84) і швидку обробку текстових повідомлень (середній час 256,57 мс). Проте 6 хибнонегативних результатів (3,7 %) та 5 (3,1 %) хибнопозитивних результатів вказують на необхідність подальшого навчання та оптимізації параметрів моделі.

ВИСНОВКИ ТА ПЕРСПЕКТИВИ ПОДАЛЬШИХ ДОСЛІДЖЕНЬ

У роботі було запропоновано та реалізовано модель на основі сентиментального аналізу «MaliciousContentDetector» для виявлення шкідливого контенту в системах електронних комунікацій, таких як соціальні мережі, месенджери та електрона пошта. Тестування розробленої моделі на основі аналізу тестових повідомлень соціальних мереж показало високу ефективність розпізнання шкідливого контенту. Крім того, тестування виявило обмеження моделі, пов’язані з обробкою коротких україномовних текстів, сарказму та неоднозначних патернів (наприклад, сленгу), що призводять до хибнонегативних (3,7 %) і хибнопозитивних (3,1 %) результатів. Для підвищення точності прогнозів протягом подальших досліджень плануються наступні вдосконалення моделі та програмної реалізації:

- Розширення тренувального набору. Поточний набір даних ефективний для стандартних фішингових і спам повідомлень, але недостатній для коротких текстів (<10 слів) і нестандартних україномовних виразів;
- Адаптація до нових загроз, наприклад шляхом інтеграції з зовнішніми базами, такими як PhishTank, VirusTotal, тощо;
- Покращення класифікації типів загроз та розширення словника загроз з додаванням специфічних україномовних маркерів;
- Аналіз мультимодального контенту. Оскільки шкідливий контент може часто містити мультимедійний контент, одним із напрямків розвитку програмної реалізації є інтеграція моделей комп’ютерного зору для аналізу такого контенту.

Таким чином, продемонстровано, що використання сентиментального аналізу є ефективним методом для ідентифікації шкідливого контенту в системах електронних комунікацій, таких як електронна пошта, месенджери, соціальні мережі. Розроблена модель штучного інтелекту показала високу ефективність на практиці, підтверджуючи її здатність покращити кібербезпеку кінцевих користувачів, шляхом фільтрації шкідливого контенту.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. DeBounce. (2025). *Email spam statistics 2025*. Retrieved May 1, 2025, from https://debounce.io/email-spam-statistics
2. Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal.* Retrieved March 1, 2025, from https://www.sciencedirect.com/science/article/pii/S2090447914000550
3. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). *Attention is all you need. Advances in Neural Information Processing Systems, 30.*
4. Darchuk, N. (2019). Linguistic approach for development of computer-based sentiment analysis in the Ukrainian language. *Science and Education a New Dimension, VII(189)*(55), 10–13. Retrieved April 10, 2025, from https://seanewdim.com/wp-content/uploads/2021/04/Linguistic-approach-for-development-of-computer-based-sentiment-analysis-in-the-Ukrainian-language-N.-Darchuk.pdf
5. Кириченко, Р. (2021). Типологія задач машинного аналізу текстів у сучасній соціології. *Sociological studios, 2(19),* 53–62. Retrieved July 15, 2025, from https://journals.indexcopernicus.com/api/file/viewByFileId/1564492

6. Рябишев, О., Єрохін, А., & Бахмет, А. (2021). Аналіз тональності тексту українською мовою. *Біоніка інтелекту: науково-технічний журнал, (96),* 15–21. Retrieved August 12, 2025, from https://openarchive.nure.ua/server/api/core/bitstreams/f59561cb-66a5-4f99-b6ff-cf30a5162f91/content
7. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). *BERT: Pre-training of deep bidirectional transformers for language understanding.* Cornell University. Retrieved from https://arxiv.org/abs/1810.04805
8. Amazon Web Services. (2025). *What is sentiment analysis? Sentiment analysis explained.* Retrieved May 1, 2025, from https://aws.amazon.com/what-is/sentiment-analysis
9. Baccianella, S., Esuli, A., & Sebastiani, F. (2010). SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)* (pp. 2000–2004).
10. scikit-learn. (2025). *Machine learning in Python.* Retrieved May 15, 2025, from https://scikit-learn.org
11. CrowdStrike. (2025). *Falcon Endpoint Protection Platform (EPP).* Retrieved May 12, 2025, from https://www.crowdstrike.co.uk/falcon-platform
12. Darktrace. (2025). *ActiveAI security platform | Darktrace: The essential AI cybersecurity platform.* Retrieved June 1, 2025, from https://www.darktrace.com/platform
13. Microsoft. (2025). *Microsoft security copilot.* Retrieved June 1, 2025, from https://www.microsoft.com/en-us/security/business/ai-machine-learning/microsoft-security-copilot
14. Sift. (2025). *Why Sift.* Retrieved June 1, 2025, from https://sift.com/why-sift

Dmytro V. Andrushko
PhD (Eng.),
Senior Lecturer of the Department of Infocommunication Engineering named after V.V. Popovskyi
Kharkiv National University of Radio Electronics, Kharkiv, Ukraine
ORCID: 0009-0004-0749-9729
dmytro.andrushko@nure.ua

Dmytro A. Shedin
M. Sc. (Eng.), Department of Infocommunication Engineering named after V.V. Popovskyi
Kharkiv National University of Radio Electronics, Kharkiv, Ukraine
ORCID: 0009-0007-4597-5871
dmytro.shedin@nure.ua

Vadym K. Chakrian
PhD (Eng.),
Senior Lecturer of the Department of Infocommunication Engineering named after V.V. Popovskyi
Kharkiv National University of Radio Electronics, Kharkiv, Ukraine
ORCID: 0000-0003-4827-9779
vadym.chakrian@nure.ua

METHODOLOGY FOR DETECTING MALICIOUS CONTENT USING SENTIMENT ANALYSIS

Abstract. This paper investigates the problem of detecting malicious content in personal electronic communications, such as emails, social media messages, or messengers. Modern systems for filtering malicious content are mainly based on the analysis of keywords, structural features of messages, and technical parameters, such as IP addresses or message origin domains. Such systems have limited effectiveness against modern cyber threats that use message content with emotional manipulation or fake information. In addition, the effectiveness of such approaches is limited, especially when it comes to targeted attacks and new methods of bypassing protection mechanisms that are constantly evolving. The paper proposes a hybrid model, "MaliciousContentDetector," which combines lexical methods, machine learning methods (TF-IDF, Random Forest), and deep learning (BERT) for text sentiment analysis. The model takes into account the linguistic features of the Ukrainian language and contextual emotional triggers. The proposed model is also implemented as a software which may be integrated into into browsers and other applications to enhance the cybersecurity of end-users. The developed model was tested on real messages from a popular news channel on a social network. During the analysis of 162 posts in that news channel, the "MaliciousContentDetector" model showed high efficiency and accuracy in detecting potentially malicious text content.

Keywords: sentiment analysis, malicious content, cybersecurity, artificial intelligence, BERT, electronic communications, emotional triggers.

REFERENCES (TRANSLATED AND TRANSLITERATED)

1. DeBounce. (2025). *Email spam statistics 2025.* Retrieved from https://debounce.io/email-spam-statistics
2. Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal.* Retrieved from https://www.sciencedirect.com/science/article/pii/S2090447914000550
3. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems, 30.*
4. Darchuk, N. (2019). Linguistic approach for development of computer-based sentiment analysis in the Ukrainian language. *Science and Education a New Dimension, VII(189)*(55), 10–13.* Retrieved from https://seanewdim.com/wp-content/uploads/2021/04/Linguistic-approach-for-development-of-computer-based-sentiment-analysis-in-the-Ukrainian-language-N.-Darchuk.pdf

5. Kyrychenko, R. (2021). Typology of tasks of machine analysis of texts in contemporary sociology. *Sociological Studios, 2(19),* 53–62. Retrieved from https://journals.indexcopernicus.com/api/file/viewByFileId/1564492
6. Riabyshev, O., Yerokhin, A., & Bakhmet, A. (2021). Analysis of the sentiment of the text in the Ukrainian language. *Bionics of Intelligence, (96),* 15–21. Retrieved from https://openarchive.nure.ua/server/api/core/bitstreams/f59561cb-66a5-4f99-b6ff-cf30a5162f91/content
7. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). *BERT: Pre-training of deep bidirectional transformers for language understanding.* Cornell University. Retrieved from https://arxiv.org/abs/1810.04805
8. Amazon Web Services. (2025). *What is sentiment analysis? Sentiment analysis explained.* Retrieved from https://aws.amazon.com/what-is/sentiment-analysis
9. Baccianella, S., Esuli, A., & Sebastiani, F. (2010). SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)* (pp. 2000–2004).
10. scikit-learn. (n.d.). *Machine learning in Python.* Retrieved from https://scikit-learn.org
11. CrowdStrike. (2025). *Falcon Endpoint Protection Platform (EPP).* Retrieved from https://www.crowdstrike.co.uk/falcon-platform
12. Darktrace. (2025). *ActiveAI Security Platform | The essential AI cybersecurity platform.* Retrieved from https://www.darktrace.com/platform
13. Microsoft. (2025). *Microsoft Security Copilot.* Retrieved from https://www.microsoft.com/en-us/security/business/ai-machine-learning/microsoft-security-copilot
14. Sift. (2025). *Why Sift.* Retrieved from https://sift.com/why-sift