КІБЕРБЕЗПЕКА: освіта, наука, техніка
CYBERSECURITY: EDUCATION, SCIENCE, TECHNIQUE
№ 1 (29), 2025
ISSN 2663 - 4023
DOI 10.28925/2663-4023.2025.29.839
УДК 004.934

Корченко Олександр Григорович
член-кореспондент НАН України, доктор технічних наук, професор, перший проректор
Державний університет інформаційно-комунікаційних технологій, Київ, Україна
ORCID ID: 0000-0003-3376-0631
agkorchenko@gmail.com

Терейковський Ігор Анатолійович
доктор технічних наук, професор, професор кафедри системного програмування і спеціалізованих комп'ютерних систем
Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського», Київ, Україна
ORCID ID: 0000-0003-4621-9668
terejkowski@ukr.net

Дичка Іван Андрійович
доктор технічних наук, професор, декан факультету прикладної математики
Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського», Київ, Україна
ORCID ID: 0000-0002-3446-3076
dychka@pzks.fpm.kpi.ua

Романкевич Віталій Олексійович
доктор технічних наук, професор, завідувач кафедри системного програмування і спеціалізованих комп'ютерних систем
Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського», Київ, Україна
ORCID ID: 0000-0003-4696-5935
zavkaf@scs.kpi.ua

Терейковська Людмила Олексіївна
доктор технічних наук, професор, професор кафедри інформаційних технологій проєктування та прикладної математики
Київський національний університет будівництва і архітектури, Київ, Україна
ORCID ID: 0000-0002-8830-0790
tereikovskal@ukr.net

ВИЯВЛЕННЯ МАНІПУЛЯТИВНОЇ СКЛАДОВОЇ У ТЕКСТОВИХ ПОВІДОМЛЕННЯХ ЗАСОБІВ МАСОВОЇ ІНФОРМАЦІЇ У КОНТЕКСТІ ЗАХИСТУ ВІТЧИЗНЯНОГО КІБЕРПРОСТОРУ

Анотація. Проблематикою статті є підвищення ефективності засобів забезпечення інформаційної безпеки у межах національного кіберпростору шляхом автоматизованого виявлення маніпулятивної складової у текстових повідомленнях засобів масової інформації. Показано, що одним із основних напрямків підвищення ефективності таких засобів є використання великих мовних моделей, які здатні здійснювати глибокий контекстуальний аналіз природномовного тексту з урахуванням емоційного, риторичного та семантичного наповнення. Встановлено, що більшість відомих рішень в області виявлення текстових маніпуляцій за допомогою великих мовних моделей достатньо складно адаптувати до умов практичного застосування в системах захисту вітчизняного кіберпростору внаслідок необхідності створення потужної апаратно-програмної інфраструктури для обслуговування відповідних LLM-засобів та необхідності формування спеціалізованих навчальних вибірок, що враховують основні типи маніпулятивних впливів, характерні для реалій інформаційного протистояння. Для подолання цих обмежень у статті запропоновано концепцію використання LLM-засобів (GPT, Gemini, DeepSeek, Grok тощо), яка базується на реалізації діалогової

взаємодії із заздалегідь стандартизованими формалізованими запитами, що враховують основні типи маніпулятивних впливів. До таких впливів, згідно з класифікацією, віднесено емоційно-маніпулятивні повідомлення, інформаційні підміни, дискредитаційні наративи, маніпуляції контекстом, пропагандистські конструкції, експлуатацію соціально чутливих тем та штучне формування суспільної думки шляхом бот-активності. Розроблено метод взаємодії з LLM, що охоплює етапи попередньої обробки тексту, формування запитів базового, типологічного та інтерпретаційного рівнів, а також інтерпретацію відповідей LLM у формалізованому вигляді. Проведені експериментальні дослідження із залученням трьох груп текстів (наукових, політичних та деструктивно-пропагандистських) засвідчили, що результати аналізу, отримані за допомогою LLM-засобу GPT-4-turbo, у середньому на 87% узгоджуються з експертними оцінками, що свідчить про високий рівень достовірності отриманих результатів та підтверджує практичну ефективність запропонованих рішень. Показано, що підвищити точність та стабільність оцінок маніпулятивної складової можливо за рахунок донавчання LLM-засобів шляхом включення до запиту прикладів коректних відповідей, що дозволяє підвищити узгодженість результатів без необхідності повного перенавчання моделі.

Ключові слова: маніпулятивна складова; велика мовна модель; кіберпростір; інформаційне протиборство; автоматизований аналіз тексту; безпека інформації; засоби масової інформації.

ВСТУП

В сучасних умовах український кіберпростір перебуває під постійним тиском деструктивних інформаційних впливів, спрямованих на дестабілізацію суспільства та підрив довіри до державних інституцій. Масштаби й інтенсивність таких впливів суттєво зросли після початку повномасштабного вторгнення, перетворивши кіберпростір на арену системного протистояння. Таким чином, на тлі актуальних подій забезпечення кібербезпеки держави залишається одним із ключових напрямів національної безпеки України, що знаходить відображення як у розпорядженні Кабінету Міністрів України «Про затвердження плану заходів на 2025 рік з реалізації Стратегії кібербезпеки України» [18], так і в низці наукових досліджень, зокрема у статті «Особливості забезпечення кібербезпеки як провідної складової національної безпеки України» [21]. Відповідно до результатів [2], [3], одним із основних механізмів ведення інформаційної війни є маніпулятивний вплив на суспільство, реалізований за допомогою текстових повідомлень у засобах масової інформації та соціальних мережах. Такі повідомлення можуть навмисно викривлювати факти, експлуатувати емоції аудиторії або формувати хибні причинно-наслідкові зв’язки з метою дестабілізації суспільно-політичної ситуації. Виявлення й нейтралізація подібних впливів є необхідною умовою інформаційної безпеки держави. Слід зазначити, що в Україні вже створено низку платформ, за допомогою яких громадяни можуть повідомляти про ресурси, що поширюють фейки та пропаганду [9], [10], однак ефективність цих інструментів обмежена, оскільки вони потребують обов’язкового залучення людини для подальшої верифікації інформації. У цьому контексті зростає потреба в ефективних методах автоматичного розпізнавання маніпуляцій у тексті, зокрема із залученням інструментів штучного інтелекту, що на сьогодні вважаються найбільш ефективними в області аналізу текстової інформації.

Постановка проблеми. Вдосконалення методології застосування засобів штучного інтелекту для автоматизованого аналізу текстових повідомлень засобів масової інформації з огляду на необхідність виявлення маніпулятивної складової у контексті захисту вітчизняного кіберпростору.

Аналіз останніх досліджень і публікацій. Як свідчить аналіз науково-практичних робіт [2], [3], [14], сучасні дослідження в галузі виявлення маніпулятивної складової в текстових повідомленнях засобів масової інформації зосереджені на використанні технологій обробки природної мови та штучного інтелекту, які дозволяють автоматизувати аналіз великих обсягів інформації. При цьому в загальному випадку вважається, що маніпуляція — це прихований вплив на свідомість людини з метою змусити її діяти у вигідний для маніпулятора спосіб. На відміну від переконання, яке базується на відкритому і логічному поясненні, маніпуляція апелює до емоцій, когнітивних упереджень, стереотипів та неусвідомлених реакцій. У засобах масової інформації, в тому числі соціальних мережах, медіа та політичному дискурсі маніпулятивні техніки використовуються для формування думки, зміни поведінки або зміщення акцентів з важливих питань. Часто наявність маніпуляції не є очевидним фактом, що робить її особливо небезпечною в інформаційній війні чи політичній пропаганді. В текстових повідомленнях маніпулятивна складова реалізується за рахунок використання особливостей природньої мови з метою прихованого впливу на реципієнта. Маніпуляції можуть здійснюватися через підбір лексики, структуру речень, акценти, замовчування, емоційне забарвлення та інші стилістичні прийоми. Наприклад, в роботі [17] зазначається, що маніпулятивні повідомлення в соціальних мережах, зокрема в Telegram-каналах, часто використовують емоційно заряджену лексику, гіперболізацію фактів та логічні помилки для впливу на аудиторію. Тому одним із ключових напрямів досліджень в області розпізнавання маніпулятивної складової в тексті є аналіз лінгвістичних та психологічних маркерів маніпуляцій у текстах. Підкреслено, що маніпулятивні тексти мають характерні патерни, які можна виявляти за допомогою статистичних методів та моделей машинного навчання. При цьому показано, що традиційні методи, такі як моделі на основі правил або статистичні класифікатори (наприклад, SVM), мають обмежену точність через складність у визначенні контекстуальних особливостей маніпулятивного контенту. У цьому контексті особливу увагу приділено великим мовним моделям (Large Language Models, LLM), які демонструють значний потенціал у задачах класифікації текстів, виявлення емоційного забарвлення та ідентифікації маніпулятивних технік. Так, в роботах [3], [4], [6] відзначено, що розвиток LLM, таких як BERT та RoBERTa, надає нові можливості для аналізу текстової інформації. Продемонстровано, що модель BERT, яка базується на архітектурі трансформерів, здатна ефективно обробляти контекстуальні зв’язки в тексті задля здійснення емоційного впливу, що є критично важливим для виявлення маніпуляцій. Однак, як зазначають автори [15], [16], [19], для україномовних текстів необхідне донавчання моделей на спеціалізованих датасетах, адаптованих до особливостей української мови. Наведені результати свідчать про можливість підвищити точність класифікації текстів на 10–15%. Подібні рішення описані в роботах [7], [8], в яких для розпізнавання текстових маніпуляцій використано такі LLM, як GPT-4o, Gemini 1.5 Pro та Llama 3.1 8B. Декларується точність розпізнавання маніпуляцій близько 0,93, 0,8 та 0,6 відповідно. В роботі [1] показано перспективність розпізнавання маніпуляцій за допомогою мовних моделей SBERT, RoBERTa та mBERT, для навчання яких використано штучні дані, згенеровані ChatGPT. Питання захисту вітчизняного кіберпростору також активно досліджується в контексті нормативно-правового регулювання та технологічних рішень [20], [21]. Наголошується на необхідності застосування автоматизованих систем аналізу інформації для забезпечення реалізації стратегії кібербезпеки України. Зокрема, автори вказують на важливість використання засобів штучного інтелекту для моніторингу інформаційних потоків у реальному часі та

вказують на доцільність застосування автоматизованих систем аналізу інформації в межах реалізації стратегії кібербезпеки України. Таким чином, результати аналізу сучасних досліджень в області розпізнавання текстових маніпуляцій вказують на перспективність застосування в системі захисту вітчизняного кіберпростору нейромережевих засобів, що базуються на застосуванні LLM. Водночас відомі рішення мають в основному фрагментарний та дослідницький характер, що заважає їх оперативній адаптації до практичного впровадження в систему захисту вітчизняного кіберпростору, що характеризується високою динамікою маніпулятивних інформаційних впливів, необхідністю аналізу текстів, написаних різними природними мовами з використанням розмовних конструкцій і сленгових зворотів, а також обмеженими ресурсами для розгортання спеціалізованих інфраструктур. Враховуючи відому методологію розробки засобів оцінки параметрів захисту Інтернет-орієнтованих інформаційних систем, першим кроком на шляху виправлення цього недоліку являється розробка концепції використання LLM для виявлення маніпулятивної складової у текстових повідомленнях засобів масової інформації [5], [11].

Мета статті. Розробка концепції використання великих мовних моделей для автоматизованого виявлення маніпулятивної складової у текстових повідомленнях засобів масової інформації у контексті захисту вітчизняного кіберпростору.

РОЗРОБКА КОНЦЕПЦІЇ ВИКОРИСТАННЯ ВЕЛИКИХ МОВНИХ МОДЕЛЕЙ ДЛЯ ВИЯВЛЕННЯ МАНІПУЛЯТИВНОЇ СКЛАДОВОЇ

Основна ідея концепції полягає у використанні LLM як готового аналітичного інструменту для аналізу текстів з метою виявлення ознак маніпуляції. З огляду на доступність апробованого інструментального забезпечення, яке забезпечує зручну взаємодію з такими моделями (наприклад, ChatGPT, Grok, Gemini, DeepSeek), їх застосування можна вважати доцільним як з технічної, так і з методологічної точки зору. При цьому означені LLM-засоби здатні здійснювати якісний аналіз контенту багатьма природними мовами з урахуванням сленгових зворотів, що в свою чергу дозволяє уникнути значних витрат часу та ресурсів на розробку, навчання та підтримку як власних LLM, так і на розробку та підтримку відповідного апаратно-програмного забезпечення. Натомість увага може бути зосереджена на побудові ефективних методів виявлення та класифікації маніпулятивних повідомлень. Таким чином, у межах запропонованої концепції LLM розглядається не лише як класифікатор, а як аналітичний засіб, здатний розпізнавати риторичні прийоми, емоційні маркери, семантичні зсуви та інші ознаки маніпуляції, ґрунтуючись на контекстному аналізі. Використання LLM-засобу у діалоговому форматі запит-відповідь (prompt-based interaction) забезпечує можливість оцінювати повідомлення на різних рівнях (від лексичного до прагматичного) і отримувати інтерпретовані оцінки щодо маніпулятивності змісту, та дозволяє формалізувати опис концепції за допомогою наступного виразу:

$$F_{LLM}(Text, \boldsymbol{Pr}) = \boldsymbol{R}, \quad (1)$$

де F_{LLM} — функція розпізнавання маніпуляцій за допомогою LLM; $Text$ — піддослідний текст; $\boldsymbol{Pr}$ — множина запитів до LLM; $\boldsymbol{R}$ — множина відповідей LLM стосовно виявлених маніпуляцій.

Також в концепції акцентована увага на формуванні множини $\boldsymbol{Pr}$, адже відповідно до результатів [1], [6] достовірність відповідей LLM-засобів в значній мірі залежить від

врахування в запиті інтелектуальних можливостей LLM та коректності формулювання запиту щодо очікуваного результату. Базуючись на висновках [12], [14], визначено, що до складу множини $\boldsymbol{R}$ повинні входити відповіді, які в першому наближенні (базовий варіант аналізу) відображають розраховану LLM-засобом числову оцінку наявності маніпуляції та її спрямованості, а в другому (типологічний аналіз) та третьому наближенні (інтерпретаційний аналіз) — додатково числову оцінку кожного з основних типів маніпуляцій та аргументоване пояснення отриманих результатів. Селекція очікуваних відповідей, тобто складових множини $\boldsymbol{R}$, реалізована з позицій допустимого терміну очікування відповіді від LLM-засобу, доступного обсягу обчислювальних ресурсів та цілей розпізнавання маніпуляцій. Також з огляду на потреби формування множини відповідей, враховуючи результати [13], [20] в контексті сучасних реалій, визначено низку ключових типів маніпулятивних впливів, які доцільно враховувати під час автоматизованого аналізу текстів. Їх стисла характеристика наведена в табл. 1.

Таблиця 1

Характеристики основних типів маніпуляцій

Позначення	Тип маніпуляції	Коротка характеристика
m_1	Емоційно-маніпулятивні повідомлення	Використання мови, що провокує страх, тривогу, гнів або відразу — часто базується на перебільшенні загроз
m_2	Інформаційні підміни	Підміна понять, виривання цитат з контексту, зміщення акцентів з метою спотворення змісту
m_3	Дискредитаційні наративи	Фейкові або маніпулятивні звинувачення щодо державних органів, ЗСУ, волонтерів тощо
m_4	Формування хибного контексту	Подання правдивих фактів у такому інформаційному оточенні, що змінює їхнє первинне значення
m_5	Агітаційно-пропагандистські конструкції	Використання імперських, релігійних чи ностальгійних мотивів в інтересах ворожої пропаганди
m_6	Маніпуляції соціальними темами	Експлуатація питань мови, історії, релігії, економіки для розпалювання суспільної ворожнечі
m_7	Токсичні дискусії та бот-активність	Імітація суспільної думки через провокаційні коментарі, кероване поширення образливих або збурювальних тез

Таким чином, множина типів маніпуляцій, що підлягають виявленню, описується за допомогою виразу:

$$\boldsymbol{M} = \{m\}_{\boldsymbol{7}}\,. \tag{2}$$

Враховуючи функціональні можливості апробованих LLM, пропонується, що в множині запитів, відповідних заданій виразом (2) множині $\boldsymbol{M}$ слід відобразити особливості кожної маніпуляції. Тобто

$$z_i \to m_i\,, m_i \in \boldsymbol{M}, z_i \in \boldsymbol{Z}, i = 1,\ldots,7, \tag{3}$$

де z_i — фрагмент тексту запиту, що відповідає m_i типу маніпуляції; $\boldsymbol{Z}$ — множина фрагментів текстів запитів на розпізнавання основних типів маніпуляцій.

Фрагменти текстів запитів для виявлення основних типів маніпуляцій наведено в табл. 2. Передбачається, що в більшості випадків, взаємодія з LLM-засобами реалізується з використанням технології zero-shot, тобто модель отримує лише запит (prompt) без жодного прикладу виконання завдання.

Таблиця 2

Фрагменти запитів для розпізнавання основних типів маніпуляцій

Позначення фрагменту запиту	Позначення типу маніпуляції	Фрагмент тексту запиту
z_1	m_1	«емоційного маніпулятивного впливу»
z_2	m_2	«інформаційних підмін: підміни понять, зміщення акцентів або виривання з контексту»
z_3	m_3	«дискредитації органів влади, ЗСУ або волонтерів»
z_4	m_4	«зміни значень достовірних фактів»
z_5	m_5	«пропагандистських елементів (ностальгія за СРСР, імперські, релігійні мотиви тощо)»
z_6	m_6	«соціально чутливих тем з маніпулятивною метою (наприклад, мова, історія, релігія, економіка)»
z_7	m_7	«маркерів токсичності та ознак координації через бото- чи тролеферми»

Також, з позицій забезпечення універсальності, передбачено стандартизувати склад запиту. Так, наприклад, першу частину запиту, що відповідає фокусній рамці, пропонується формулювати у вигляді (4), а фрагменти запиту щодо загальної оцінки наявності маніпуляцій, оцінки інтенсивності впливу та ймовірної цілі впливу — у вигляді (5–7) відповідно.

$$bg = \text{"Проаналізуй текст і дай максимально строгу відповідь, без творчих відхилень, без жодних додаткових коментарів, лише у вказаному форматіу форматі JSON, де вкажи: ",} \quad (4)$$

$$ms = \text{"manip\_sce} - \text{числова оцінка ймовірності наявності маніпуляції (0–1), ",} \quad (5)$$

$$is = \text{"infl\_sc} - \text{числова оцінка інтенсивності впливу на читача (0–1), ",} \quad (6)$$

$$gs = \text{"}g\text{\_sc} - \text{ймовірну ціль впливу. ",} \quad (7)$$

$$legal\_risk - true/false \text{ (чи є ознаки порушення законодавства України).} \quad (8)$$

Варіанти запитів для реалізації експрес-аналізу (Pr_1), типологічного аналізу (Pr_2) та інтерпретаційного аналізу (Pr_3), розроблені з урахуванням (4–8), пропонується формувати за допомогою виразів (9-14).

$$Pr_1 = bg + ms + is + gs + t, \quad (9)$$

$$Pr_2 = bg + ms + is + gs + ts + Z + t, \quad (10)$$

$$t = \text{ " Текст: [" текст для аналізу "]",} \quad (11)$$

$$ts = \text{"}t\text{\_}s - \text{об'єкт із оцінками (0–1) для кожного типу маніпуляцій: ",} \quad (12)$$

$$Z = z_1 + z_2 + z_3 + z_4 + z_5 + z_6 + z_7, \quad (13)$$

$$Pr_3 = bg + dm + ms + is + gs + ts + Z + t, \quad (14)$$

$$dm = \text{"}d\text{\_}s - \text{пояснення цілі впливу для кожного типу маніпуляцій, "} \quad (15)$$

$$exp = \text{"}exp - \text{опис логіки виявлення для кожного типу маніпуляцій. "} \quad (16)$$

Базуючись на результатах аналізу функціональних можливостей сучасних апробованих LLM-засобів, враховуючи наведені в табл. 1 характеристики основних типів маніпуляцій та задані виразами (2–16) запити щодо розпізнавання маніпуляцій в концепції, передбачається, що в базовому випадку метод взаємодії з LLM поділяється на наступні етапи:

1. Введення текстового повідомлення, що може бути реалізоване за рахунок «ручного введення», так і за рахунок підключення до засобів автоматичного моніторингу онлайн ресурсів.
2. Попередня обробка текстового повідомлення. Реалізується, коли обсяг текстового повідомлення перевищує допустиму межу. Для цього введене

текстове повідомлення підлягає редагуванню, що повинне призвести до його скорочення без втрати змісту.

3. Формування запиту до LLM-засобу, що реалізується в залежності від обраного виду аналізу.
4. Проведення аналізу текстового повідомлення за допомогою LLM-засобу.
5. Інтерпретація результатів аналізу, що полягає у перетворенні відповіді LLM-засобу для подальшого використання.
6. Донавчання LLM. Даний етап виконується у випадку залучення експертів до оцінки відповідей LLM-засобів та забезпечує можливість як вдосконалення запитів до LLM, так і підвищення точності відповідей LLM в межах сеансу.

Узагальнюючи отримані результати, слід зазначити, що побудована концепція інтегрує розроблений метод взаємодії з LLM-засобами, характеристики основних типів маніпуляцій (див. табл. 1) та визначені за допомогою виразів (4–14) формалізовані запити, що забезпечує адаптивний підхід до виявлення маніпуляцій у текстових повідомленнях за допомогою LLM-засобів.

ДОСЛІДЖЕННЯ ОБМЕЖЕНЬ ЩОДО ЗАСТОСУВАННЯ LLM-ЗАСОБІВ ДЛЯ РОЗПІЗНАВАННЯ МАНІПУЛЯЦІЙ

Незважаючи на широке поширення LLM-засобів як інструментів контекстного аналізу, їх застосування у завданнях, пов’язаних із виявленням маніпулятивного змісту текстової інформації, супроводжується низкою труднощів, зумовлених як новизною самих технологій, так і специфікою поставлених задач [1], [14]. До них належать нестійкість відповідей до незначних змін у формулюванні запиту, варіативність інтерпретації одного й того самого тексту, а також відсутність контрольованого механізму оцінки достовірності отриманих результатів. Зважаючи на вказані труднощі, а також з урахуванням розробленого механізму формування однотипних запитів (вирази 4–16) і попереднього оціночного характеру визначення концептуальних обмежень щодо застосування LLM-засобів для розпізнавання маніпуляцій, планом експерименту було передбачено верифікацію запропонованих рішень шляхом аналізу відповідей LLM-засобів на типові запити щодо виявлення маніпулятивної складової в текстах, зібраних авторським колективом із відкритих джерел інформації. Розглянуто три групи текстів. До першої групи віднесено тексти наукових статей, присвячених різноманітним питанням захисту інформації, опубліковані у фахових журналах України. До другої — публічні виступи вітчизняних політичних діячів, що містять риторичні конструкції з чітко вираженим емоційним забарвленням. До третьої групи віднесено тексти, що поширюються через інформаційні платформи з деструктивною риторикою стосовно України. Тексти третьої групи зібрано шляхом моніторингу текстових повідомлень вітчизняних засобів масової інформації, що містили фрагменти або повні цитати матеріалів з інформаційних ресурсів, ідентифікованих як джерела ворожої пропаганди, дезінформації або інформаційно-психологічного впливу.

В якості LLM-засобу використано GPT-4-turbo, що пояснюється його доступністю у рамках платформи ChatGPT, високою якістю контекстного аналізу, а також здатністю забезпечувати стабільні результати при обробці запитів з великим обсягом текстових даних. Ілюстрацією застосування GPT-4-turbo для відповіді на запити щодо маніпулятивної складової являється рис. 1, на якому показано варіант запиту у випадку базового аналізу тексту, який є частиною даної статті.

ISSN 2663 - 4023

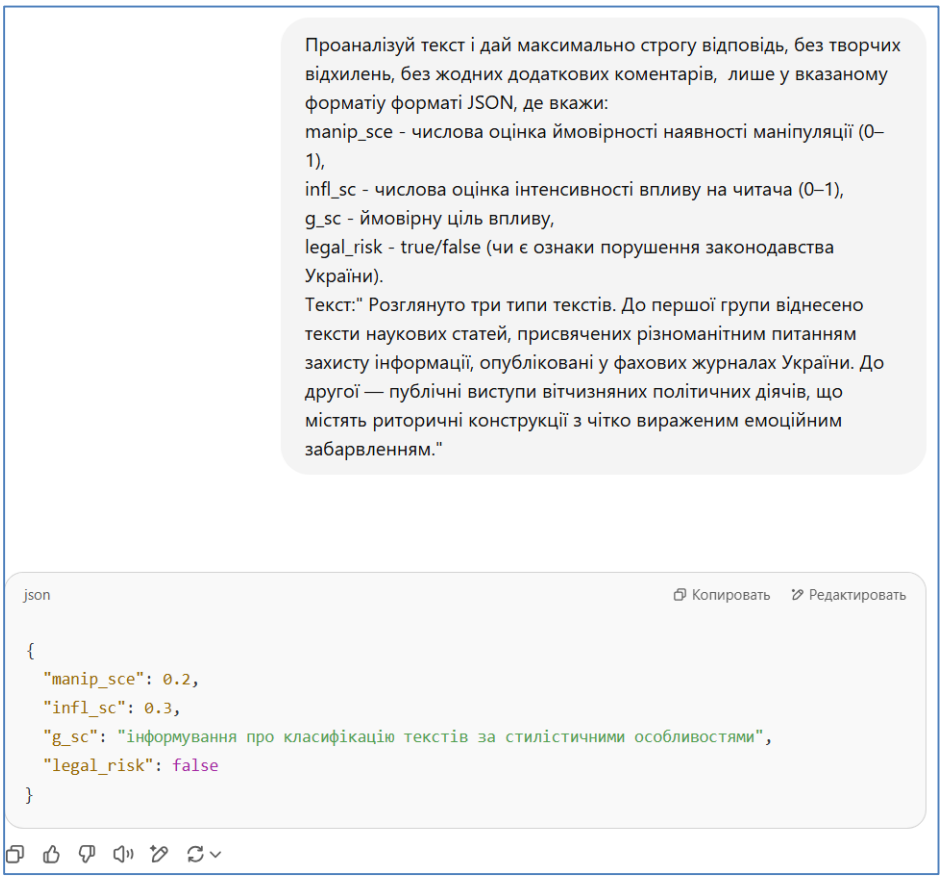

Рис. 1. Ілюстрація застосування GPT-4-turbo у випадку базового аналізу

Як і передбачено запитом, GPT-4-turbo сформував відповідь у форматі JSON достатньо якісно оцінивши як маніпулятивну, емоційну та семантичну складову тексту, так і його відповідність нормам вітчизняного законодавства.

Результати експертного оцінювання отриманих від GPT-4-turbo відповідей щодо маніпулятивної складової в текстових повідомленнях для всіх трьох груп текстів для всіх передбачених варіантів аналізу вказують на його достатньо високу ефективність для виявлення маніпулятивної складової у текстових повідомленнях засобів масової інформації у контексті захисту вітчизняного кіберпростору. За результатами порівняння з узагальненими експертними оцінками, середня точність класифікації становила 87%. При цьому GPT-4-turbo забезпечила можливість не лише категоризації повідомлень, але формування обґрунтованих пояснень щодо виявлення маніпулятивної складової текстових повідомлень. Відповідно до [3], [4], [7], перспективним шляхом підвищення точності розпізнавання маніпулятивної складової LLM-засобами вважається застосування донавчання, яке, хоча й потребує розробки відповідних наборів даних, досить ефективно реалізується за рахунок технології навчання з небагатьма прикладами (few-shot learning), коли разом із запитом моделі надаються кілька прикладів коректних відповідей для формування контексту. Разом з тим, під час проведення експериментів було встановлено, що відповіді LLM-засобів на однакові запити щодо маніпулятивної складової одного й того самого тексту можуть дещо відрізнятися, хоча при повторних запитах частина відповіді, що стосувалася можливих ознак порушення законодавства України, залишалася стабільною. Нестабільність відповідей виявляється у варіативності числових оцінок параметрів маніпуляцій, що вимірюються за шкалою від 0 (повна

відсутність маніпуляції) до 1 (максимальна ймовірність маніпуляції). У деяких випадках розбіжності між відповідями сягали до 10%, що може впливати на узгодженість оцінювання. Водночас, відповідно до [1], [7], вказаний недолік ефективно нівелюється за рахунок застосування апробованих методів, що базуються на усередненні результатів декількох повторних запусків та/або встановленні порогового рівня числової оцінки, перевищення якого є умовою прийняття рішення про наявність маніпулятивної складової. Таким чином, хоча результати експериментальних досліджень і вказують на необхідність розробки додаткових рішень, пов'язаних з підвищенням точності та стабільності відповідей LLM-засобів, однак в цілому підтверджують доцільність застосування розробленої концепції використання великих мовних моделей для автоматизованого виявлення маніпулятивної складової в текстових повідомленнях засобів масової інформації у контексті захисту вітчизняного кіберпростору.

ВИСНОВКИ ТА ПЕРСПЕКТИВИ ПОДАЛЬШИХ ДОСЛІДЖЕНЬ

В результаті проведених досліджень вперше розроблено концепцію використання великих мовних моделей для автоматизованого виявлення маніпулятивної складової у текстових повідомленнях засобів масової інформації у контексті захисту вітчизняного кіберпростору, що за рахунок реалізації взаємодії з LLM-засобами у форматі діалогу на основі системи формалізованих запитів забезпечує можливість виявлення ознак маніпулятивного впливу, характерного для сучасних умов інформаційної протидії в Україні. Зокрема, йдеться про такі типи маніпуляцій, як емоційно-маніпулятивні повідомлення, інформаційні підміни, дискредитаційні наративи, формування хибного контексту, агітаційно-пропагандистські конструкції, маніпуляції на соціально чутливих темах та токсична бот-активність. В межах запропонованої концепції розроблено метод взаємодії з LLM-засобами, що описує структуровану послідовність дій для автоматизованого аналізу текстових повідомлень та забезпечує оцінювання виявлених типів маніпуляцій з урахуванням їх характерних мовних і риторичних ознак. Метод передбачає використання доступних та апробованих LLM-засобів, здатних аналізувати текстові повідомлення без попереднього навчання чи надання прикладів. Це усуває потребу у створенні спеціалізованих навчальних вибірок і розробці власного апаратно-програмного забезпечення, що, у свою чергу, спрощує практичне впровадження рішення в умовах обмежених ресурсів і потреби в оперативному реагуванні, особливо в контексті сучасного інформаційного протиборства, з яким стикається Україна. Проведені експериментальні дослідження засвідчили, що результати аналізу, отримані за допомогою LLM-засобу GPT-4-turbo, у середньому на 87% узгоджуються з експертними оцінками, що свідчить про високий рівень достовірності отриманих результатів та підтверджує практичну ефективність запропонованих рішень. Таким чином, на відміну від відомих підходів, які передбачають створення україномовних навчальних вибірок, донавчання моделей або розгортання спеціалізованого програмного забезпечення, запропоноване рішення передбачає використання доступних LLM-засобів у режимі діалогової взаємодії з формуванням запитів типологічного характеру, що не потребує попереднього навчання моделі чи створення спеціалізованих датасетів. Це дозволяє зменшити витрати, пов'язані з розробкою нейромережевих засобів виявлення маніпулятивної складової, сприяє зниженню залежності від спеціалізованих апаратно-програмних ресурсів і дозволяє підвищити оперативність інтеграції рішення у практичні системи кіберзахисту. Крім того, на відміну від методів, орієнтованих лише на загальну оцінку тексту, у розробленій

концепції реалізовано багаторівневу типологічну та інтерпретаційну діагностику маніпуляцій, що забезпечує інтерпретацію типу маніпулятивного впливу та попередню оцінку правової допустимості контенту відповідно до законодавства України.

Перспективи подальших досліджень полягають в розробці методу формування типологічних запитів, який забезпечить їх адаптацію до нових форм маніпуляцій, що еволюціонують у відповідь на інформаційне протистояння. Також доцільно дослідити можливості підвищення точності класифікації за рахунок інтегрованого використання декількох LLM-засобів та шляхом застосування підходу навчання з небагатьма прикладами (few-shot learning), який дозволяє підвищити стабільність та точність відповідей без потреби у великих навчальних вибірках, подаючи моделі кілька якісно підібраних прикладів під час формування запиту. Ще один перспективний напрямок досліджень доцільно пов'язати з розробкою засобів інтеграції LLM-рішень у вітчизняні системи моніторингу інформаційного простору в режимі реального часу, що дозволить не лише виявляти маніпуляції, але й оперативно оцінювати динаміку їх поширення та потенційний вплив.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Alsaedi, N., & Alsaedi, A. (2023). Improving multiclass classification of fake news using BERT-based models and ChatGPT-augmented data. *Inventions, 8(5)*, 112. https://doi.org/10.3390/inventions8050112
2. Da San Martino, G., Barrón-Cedeño, A., Petrov, R., & Nakov, P. (2019). Fine-grained analysis of propaganda in news articles [Preprint]. *arXiv*. https://arxiv.org/abs/1910.02517
3. Da San Martino, G., Yu, S., Barrón-Cedeño, A., Petrov, R., & Nakov, P. (2020). SemEval-2020 task 11: Detection of propaganda techniques in news articles [Preprint]. *arXiv*. https://arxiv.org/abs/2009.02696
4. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding [Preprint]. *arXiv*. https://arxiv.org/abs/1810.04805
5. Korchenko, O., Tereikovskyi, I., Ziubina, R., Tereikovska, L., Korystin, O., Tereikovskyi, O., & Karpinskyi, V. (2025). Modular neural network model for biometric authentication of personnel in critical infrastructure facilities based on facial images. *Applied Sciences, 15*, 2553. https://doi.org/10.3390/app15052553
6. Li, L., Ma, R., Guo, Q., & Qiu, X. (2020). BERT-ATTACK: Adversarial attack against BERT using BERT. *arXiv*. https://arxiv.org/abs/2004.09984
7. Lin, S., Hilton, J., & Evans, O. (2021). TruthfulQA: Measuring how models mimic human falsehoods [Preprint]. *arXiv*. https://arxiv.org/abs/2109.07958
8. Smith, J., & Brown, L. (2022). Fake news detection using deep learning models. *Journal of Computational Linguistics, 28(4)*, 345–360.
9. StopRussia | MRIYA. (n.d.). Official chatbot of Ukraine. *Internet Archive*. https://web.archive.org/web/20220601115228/https://t.me/stopdrugsbot
10. StopRussiaChannel | MRIYA. (n.d.). Official Telegram channel of Ukraine for checking and blocking resources that spread fake news and propaganda. *Internet Archive*. https://web.archive.org/web/20220601090556/https://t.me/stoprussiachannel
11. Tereikovskyi, I., Korchenko, O., Bushuyev, S., Tereikovskyi, O., Ziubina, R., & Veselska, O. (2023). A neural network model for object mask detection in medical images. *International Journal of Electronics and Telecommunications, 69(1)*, 41–46. https://doi.org/10.24425/ijet.2023.144329
12. Tereikovskyi, I., AlShboul, R., Mussiraliyeva, Sh., Tereikovska, L., Bagitova, K., Tereikovskyi, O., & Hu, Zh. (2024). Method for constructing neural network means for recognizing scenes of political extremism in graphic materials of online social networks. *International Journal of Computer Network and Information Security, 16(3)*, 52–69. https://doi.org/10.5815/ijcnis.2024.03.05
13. Tereikovskyi, I., Hu, Zh., Chernyshev, D., Tereikovska, L., Korystin, O., & Tereikovskyi, O. (2022). The method of semantic image segmentation using neural networks. *International Journal of Image, Graphics and Signal Processing, 14(6)*, 1–14. https://doi.org/10.5815/ijigsp.2022.06.01
14. Wang, Z., Liu, Y., & Chen, X. (2023). Implementing BERT and fine-tuned RoBERTa to detect AI generated news by ChatGPT [Preprint]. *arXiv*. https://doi.org/10.48550/arXiv.2306.07401

15. Volokhovskyi, V., Khovrat, A., Kobziev, V., & Nazarov, O. (2024). Domain specific text analysis via decoder-only large language models. *Grail of Science*, *43*, 313–321. https://doi.org/10.36074/grail-of-science.06.09.2024.041
16. Vorochek O.H., & Solovei I.V. (2024). Using artificial intelligence speech models to generate social media posts. *Technical Engineering, 1(93)*, 128–134. https://doi.org/10.26642/ten-2024-1(93)-128-134
17. Zaporozhets, V., & Opirskyy, I. (2024). The danger of using Telegram and its impact on ukrainian society. *Cybersecurity: Education, Science, Technique*, *1*(25), 59–78. https://doi.org/10.28925/2663-4023.2024.25.5978.
18. Cabinet of Ministers of Ukraine. (2025). On approval of the action plan for 2025 for the implementation of the cybersecurity strategy of Ukraine: Order No. 204-p. https://www.kmu.gov.ua/npas/pro-zatverdzhennia-planu-zakhodiv-na-2025-rik-z-realizatsii-stratehii-kiberbezpeky-t70325
19. Korovii O., & Tereikovskyi I. (2024). Conceptual model of the process of determining the emotional tonality of the text. *Computer-integrated technologies: education, science, production*, *55*, 115–123. https://doi.org/10.36910/6775-2524-0560-2024-55-14
20. Lesko N. V., & Kira S. O. (2023). Cyber security as a part of the national security of Ukraine in the conditions of war. *Juridical scientific and electronic journal*, *5*, 112–120. https://doi.org/10.32782/2524-0374/2023-5/55
21. Chervyakov O. (2024). Peculiarities of ensuring cyber security as a leading component of Ukraine's national security. *Bulletin of criminological association of Ukraine*, *33(3)*, 511–518. https://doi.org/10.32631/vca.2024.3.47

КІБЕРБЕЗПЕКА: освіта, наука, техніка
CYBERSECURITY: EDUCATION, SCIENCE, TECHNIQUE
№ 4 (28), 2025
ISSN 2663 - 4023

Oleksandr Korchenko
Corresponding Member of the National Academy of Sciences of Ukraine,
Doctor of Technical Sciences, Professor, First vice-rector
State University of Information and Communication Technologies, Kyiv, Ukraine
ORCID ID: 0000-0003-3376-0631
agkorchenko@gmail.com

Ihor Tereikovskyi
Doctor of Technical Sciences, Professor, Professor of the Department of
Systems Programming and Specialized Computer Systems
National Technical University of Ukraine
"Igor Sikorsky Kyiv Polytechnic Institute", Kyiv, Ukraine
ORCID ID: 0000-0003-4621-9668
terejkowski@ukr.net

Ivan Dychka
Doctor of Technical Sciences, Professor, Dean of the Faculty of Applied Mathematics
National Technical University of Ukraine
"Igor Sikorsky Kyiv Polytechnic Institute", Kyiv, Ukraine
ORCID ID: 0000-0002-3446-3076
dychka@pzks.fpm.kpi.ua

Vitaliy Romankevich
Doctor of Technical Sciences, Professor, Head of the Department of
Systems Programming and Specialized Computer Systems
National Technical University of Ukraine
"Igor Sikorsky Kyiv Polytechnic Institute", Kyiv, Ukraine
ORCID ID: 0000-0003-4696-5935
zavkaf@scs.kpi.ua

Liudmyla Tereikovska
Doctor of Technical Sciences, Professor, Professor of the Department of
Information Technology Design and Applied Mathematics
Kyiv National University of Construction and Architecture, Kyiv, Ukraine
ORCID ID: 0000-0002-8830-0790
tereikovskal@ukr.net

DETECTION OF MANIPULATIVE COMPONENT IN TEXT MESSAGES OF MASS MEDIA IN THE CONTEXT OF PROTECTION OF DOMESTIC CYBERSPACE

Abstract. The problem of the article is to increase the effectiveness of information security means within the national cyberspace by automated detection of the manipulative component in text messages of the mass media. It is shown that one of the main directions of increasing the effectiveness of such means is the use of large language models, which are capable of performing a deep contextual analysis of a natural language text, taking into account emotional, rhetorical and semantic content. It is established that most of the known solutions in the field of detecting text manipulations using large language models are quite difficult to adapt to the conditions of practical application in systems for protecting domestic cyberspace due to the need to create a powerful hardware and software infrastructure to service the corresponding LLM-means and the need to form specialized training samples that take into account the main types of manipulative influences characteristic of the realities of information confrontation. To overcome these limitations, the article proposes a concept for using LLM tools (GPT, Gemini, DeepSeek, Grok, etc.), which is based on the implementation of dialogic interaction with pre-standardized formalized queries that take into account the main types of manipulative influences. Such influences, according to the classification, include emotional-manipulative messages, information substitutions, discrediting narratives, context manipulation, propaganda constructs, exploitation of socially sensitive topics, and artificial formation of public opinion through bot activity. A method of interaction with LLM has been

developed, which includes the stages of text pre-processing, the formation of queries at the basic, typological, and interpretative levels, as well as the interpretation of LLM responses in a formalized form. Experimental studies involving three groups of texts (scientific, political, and destructive-propaganda) have shown that the analysis results obtained using the GPT-4-turbo LLM tool agree with expert assessments by an average of 87%, which indicates a high level of reliability of the results obtained and confirms the practical effectiveness of the proposed solutions. It is shown that it is possible to increase the accuracy and stability of assessments of the manipulative component by retraining the LLM tools by including examples of correct answers in the query, which allows to increase the consistency of the results without the need for complete retraining of the model.

Keywords: manipulative component; large language model; cyberspace; information confrontation; automated text analysis; information security; mass media.

REFERENCES (TRANSLATED AND TRANSLITERATED)

1. Alsaedi, N., & Alsaedi, A. (2023). Improving multiclass classification of fake news using BERT-based models and ChatGPT-augmented data. *Inventions, 8(5)*, 112. https://doi.org/10.3390/inventions8050112
2. Da San Martino, G., Barrón-Cedeño, A., Petrov, R., & Nakov, P. (2019). Fine-grained analysis of propaganda in news articles [Preprint]. *arXiv*. https://arxiv.org/abs/1910.02517
3. Da San Martino, G., Yu, S., Barrón-Cedeño, A., Petrov, R., & Nakov, P. (2020). SemEval-2020 task 11: Detection of propaganda techniques in news articles [Preprint]. *arXiv*. https://arxiv.org/abs/2009.02696
4. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding [Preprint]. *arXiv*. https://arxiv.org/abs/1810.04805
5. Korchenko, O., Tereikovskyi, I., Ziubina, R., Tereikovska, L., Korystin, O., Tereikovskyi, O., & Karpinskyi, V. (2025). Modular neural network model for biometric authentication of personnel in critical infrastructure facilities based on facial images. *Applied Sciences, 15*, 2553. https://doi.org/10.3390/app15052553
6. Li, L., Ma, R., Guo, Q., & Qiu, X. (2020). BERT-ATTACK: Adversarial attack against BERT using BERT. *arXiv*. https://arxiv.org/abs/2004.09984
7. Lin, S., Hilton, J., & Evans, O. (2021). TruthfulQA: Measuring how models mimic human falsehoods [Preprint]. *arXiv*. https://arxiv.org/abs/2109.07958
8. Smith, J., & Brown, L. (2022). Fake news detection using deep learning models. *Journal of Computational Linguistics, 28(4)*, 345–360.
9. StopRussia | MRIYA. (n.d.). Official chatbot of Ukraine. *Internet Archive*. https://web.archive.org/web/20220601115228/https://t.me/stopdrugsbot
10. StopRussiaChannel | MRIYA. (n.d.). Official Telegram channel of Ukraine for checking and blocking resources that spread fake news and propaganda. *Internet Archive*. https://web.archive.org/web/20220601090556/https://t.me/stoprussiachannel
11. Tereikovskyi, I., Korchenko, O., Bushuyev, S., Tereikovskyi, O., Ziubina, R., & Veselska, O. (2023). A neural network model for object mask detection in medical images. *International Journal of Electronics and Telecommunications, 69(1)*, 41–46. https://doi.org/10.24425/ijet.2023.144329
12. Tereikovskyi, I., AlShboul, R., Mussiraliyeva, Sh., Tereikovska, L., Bagitova, K., Tereikovskyi, O., & Hu, Zh. (2024). Method for constructing neural network means for recognizing scenes of political extremism in graphic materials of online social networks. *International Journal of Computer Network and Information Security, 16(3)*, 52–69. https://doi.org/10.5815/ijcnis.2024.03.05
13. Tereikovskyi, I., Hu, Zh., Chernyshev, D., Tereikovska, L., Korystin, O., & Tereikovskyi, O. (2022). The method of semantic image segmentation using neural networks. *International Journal of Image, Graphics and Signal Processing, 14(6)*, 1–14. https://doi.org/10.5815/ijigsp.2022.06.01
14. Wang, Z., Liu, Y., & Chen, X. (2023). Implementing BERT and fine-tuned RoBERTa to detect AI generated news by ChatGPT [Preprint]. *arXiv*. https://doi.org/10.48550/arXiv.2306.07401
15. Volokhovskyi, V., Khovrat, A., Kobziev, V., & Nazarov, O. (2024). Domain specific text analysis via decoder-only large language models. *Grail of Science, 43*, 313–321. https://doi.org/10.36074/grail-of-science.06.09.2024.041
16. Vorochek O.H., & Solovei I.V. (2024). Using artificial intelligence speech models to generate social media posts. *Technical Engineering, 1(93)*, 128–134. https://doi.org/10.26642/ten-2024-1(93)-128-134

17. Zaporozhets, V., & Opirskyy, I. (2024). The danger of using Telegram and its impact on ukrainian society. *Cybersecurity: Education, Science, Technique, 1*(25), 59–78. https://doi.org/10.28925/2663-4023.2024.25.5978.
18. Cabinet of Ministers of Ukraine. (2025). On approval of the action plan for 2025 for the implementation of the cybersecurity strategy of Ukraine: Order No. 204-p. https://www.kmu.gov.ua/npas/pro-zatverdzhennia-planu-zakhodiv-na-2025-rik-z-realizatsii-stratehii-kiberbezpeky-t70325
19. Korovii O., & Tereikovskyi I. (2024). Conceptual model of the process of determining the emotional tonality of the text. *Computer-integrated technologies: education, science, production*, *55*, 115–123. https://doi.org/10.36910/6775-2524-0560-2024-55-14
20. Lesko N. V., & Kira S. O. (2023). Cyber security as a part of the national security of Ukraine in the conditions of war. *Juridical scientific and electronic journal*, *5*, 112–120. https://doi.org/10.32782/2524-0374/2023-5/55
21. Chervyakov O. (2024). Peculiarities of ensuring cyber security as a leading component of Ukraine's national security. *Bulletin of criminological association of Ukraine*, *33(3)*, 511–518. https://doi.org/10.32631/vca.2024.3.47