УДК 004.8

Юрченко Д.Ю., Мазурець О.В.

Хмельницький національний університет

МЕТОД ІНТЕРПРЕТАЦІЇ РЕЗУЛЬТАТІВ НЕЙРОМЕРЕЖЕВОГО ВИЯВЛЕННЯ ОЗНАК МАНІПУЛЯТИВНИХ ТЕХНІК ЗА ЕМОЦІЙНОЮ ТОНАЛЬНІСТЮ ТЕКСТІВ

Запропоновано метод інтерпретації результатів нейромережевого виявлення ознак маніпулятивних технік за емоційною тональністю текстів, який має перевагу у точності на понад 7% у порівнянні з відомими аналогами, та відрізняється використанням гібридної архітектури CNN та BiLSTM, що дозволяє ефективно виділяти локальні патерни та враховувати довгострокові залежності у тексті. Крім того, пропонується застосування моделі LIME для локальної інтерпретації результатів, що підвищує прозорість та інтерпретованість моделі, що сприяє довірі до нейромережевих підходів.

The method for interpreting the results of neural network detection of signs of manipulative techniques by the emotional tone of texts is proposed, which has an accuracy advantage of over 7% compared to known analogues, and is distinguished by the use of a hybrid CNN and BiLSTM architecture, which allows for effective local pattern identification and consideration of long-term dependencies in the text. In addition, the use of the LIME model for local interpretation of the results is proposed, which increases the transparency and interpretability of the model, which contributes to trust in neural network approaches.

Соціальні мережі, форуми та інші платформи стали основними каналами комунікації, де люди обмінюються думками, висловлюють свої емоції та реагують на події [1]. Важливість аналізу емоційної тональності таких повідомлень полягає у можливості виявлення загальних настроїв, тенденцій та потенційних соціальних ризиків [2]. Проте, результати нейромережевого аналізу часто є складними для сприйняття і вимагають додаткового пояснення. Створення ефективного методу візуалізації цих результатів допоможе полегшити розуміння виявлених емоційних станів, сприятиме більш інформованому прийняттю рішень та підвищенню прозорості аналізу [3, 4]. У цьому контексті використання методів візуального пояснення набуває особливого значення, оскільки дозволяє зменшити інформаційне навантаження на користувачів і забезпечити доступність аналітичних даних для широкого кола зацікавлених осіб.

Методи обробки природної мови демонструють значний потенціал у дослідженні маніпулятивних впливів у текстових повідомленнях [5, 6], зокрема

через аналіз емоційної тональності [7, 8]. Розвиток глибинних нейронних мереж [9], трансформерної архітектури [10] та контекстних ембеддингів [11] забезпечив можливість виявляти приховані патерни мовного впливу, які не піддаються фіксації традиційними статистичними підходами [12]. Емоційні індикатори – такі як інтенсивність позитивних чи негативних емоцій, наявність страхових, агресивних або мобілізаційних сигналів – стали окремими маркерами для розпізнавання маніпулятивних стратегій, зокрема залякування [13], емоційного тиску [14], викривлення інформації [15] та створення штучного емоційного контексту [16].

Водночас однією з ключових проблем є обмежена інтерпретованість нейромережевих моделей, що ускладнює достовірне пояснення того, які саме мовні ознаки стали підставою для класифікації повідомлення як маніпулятивного [17]. У цьому контексті особливого значення набувають методи Explainable AI (XAI) та інтерпретованого NLP, зокрема такі підходи, як attention-аналіз, оцінювання важливості слів (feature importance), методи LIME та SHAP для текстів, а також візуалізація емоційних патернів. Поєднання цих підходів дає змогу не лише підвищити прозорість моделей, але й виявити сталі мовні конструкції, що слугують індикаторами маніпулятивної риторики.

Проблема непрозорості моделей ШІ часто призводять до непередбачуваних результатів, включаючи упереджені рішення та недостатню інтерпретацію. Зазвичай відсутній блок пояснень результатів, які видає модель машинного навчання. І якщо мова йде про нейромережі – тут задача стає ще більш проблемною. Пропонується підхід на основі візуального пояснення результатів нейромережевого аналізу емоційної тональності, наведений на рисунку 1.

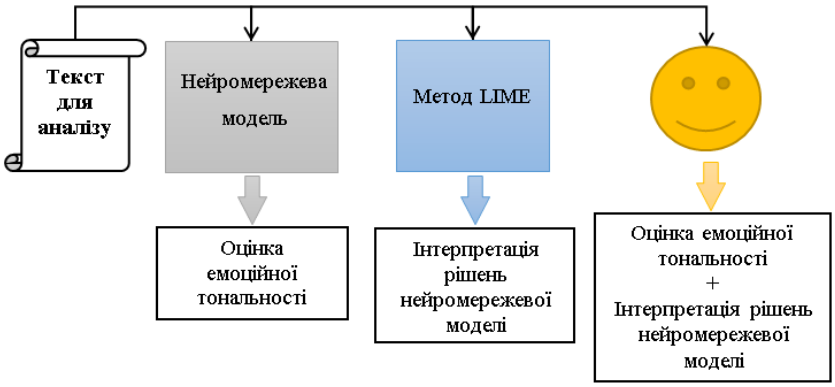

Рисунок 1 – Схема підходу на основі інтерпретації результатів нейромережевого виявлення ознак маніпулятивних технік

У межах запропонованого підходу для візуалізації пояснень результатів нейромережевого аналізу емоційної тональності пропонується застосувати модель LIME, яка є локальною моделлю для інтерпретації пояснень, незалежних від конкретної моделі [18].

Отже, такий підхід дозволить зберегти всі переваги нейромережевих рішень, одночасно надаючи користувачеві зрозуміле пояснення, що вплинуло на ці рішення. Це підвищить довіру до результатів нейромережі та дасть змогу виявляти помилки, які вона може допускати.

Метод візуального пояснення результатів нейромережевого аналізу емоційної тональності повідомлень у соціально-орієнтованих сервісах, що базується на використанні гібридної нейронної мережі архітектур CNN та BiLSTM, з локальною інтерпретацією моделлю машинного навчання LIME [19]. Схема та етапи методу наведені на рисунку 2.

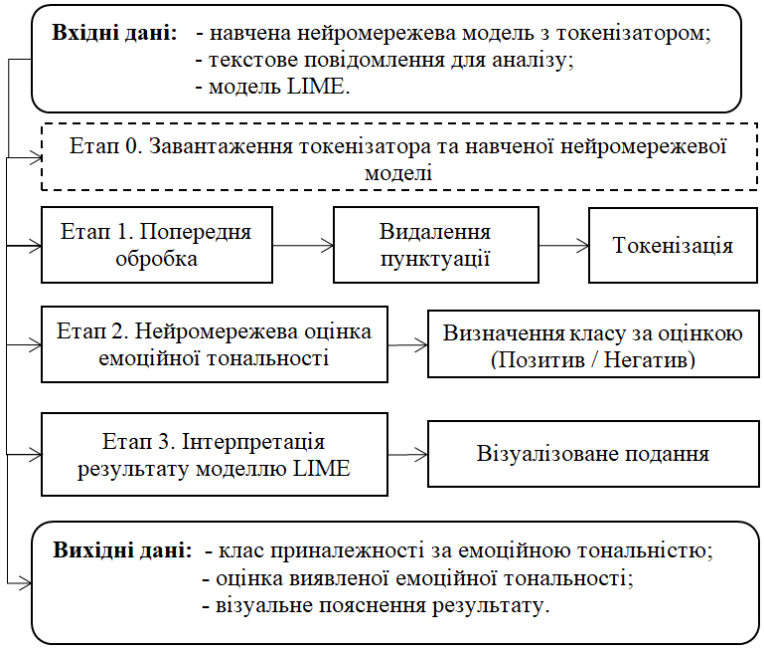

Рисунок 2 – Етапи методу інтерпретації результатів нейромережевого виявлення ознак маніпулятивних технік за емоційною тональністю текстів

Комбінація архітектур CNN та BiLSTM дозволяє ефективно виявляти локальні патерни в тексті (CNN), такі як важливі слова або фрази, що можуть бути

ключовими для визначення емоційної тональності. Bidirectional LSTM, у свою чергу, дає змогу аналізувати довгострокові залежності в тексті з обох напрямків, що сприяє кращому розумінню контексту та підвищує точність прогнозів.

Вхідними даними є попередньо навчена нейромережева модель з токенізатором, текст для аналізу та модель LIME для пояснення результатів. Спочатку необхідно завантажити токенізатор і навчану нейромережеву модель, що дасть змогу надалі подавати текстові повідомлення з соціально орієнтованих сервісів для подальшого аналізу нейронною мережею. На етапі 1 також виконується попередня обробка тексту, яка повинна відповідати тій, що застосовувалась під час навчання моделі. Вона охоплює видалення пунктуації, стоп-слів та виконання токенізації.

Наступним кроком є оцінка емоційної тональності за допомогою нейронної мережі з гібридною архітектурою, що об'єднує CNN та BiLSTM. Ця нейромережа має один вихід, і повідомлення вважається позитивним, якщо показник позитивної тональності перевищує 0,5, в іншому випадку – негативним. Третій етап передбачає інтерпретацію результату за допомогою моделі LIME.

Інтерпретація відбувається у вигляді візуалізації, яка демонструє ключові слова та ознаки, що вплинули на прийняте моделлю рішення, а також оцінку їх важливості.

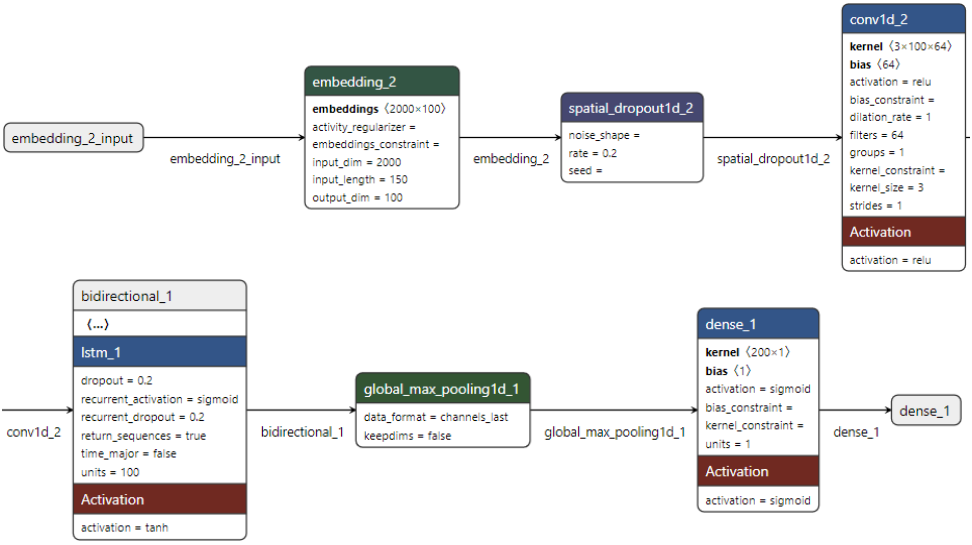

Рисунок 3 – Архітектура нейромережі для виявлення ознак маніпулятивних технік за емоційною тональністю текстів

Вихідними даними методу є клас приналежності за емоційною тональністю, оцінка виявленої емоційної тональності; візуальне пояснення результату.

Для нейромережевого аналізу емоційної тональності повідомлень запропоновано використати нейромережу гібридної архітектури, що поєднує одночасну переваги архітектур CNN та BiLSTM. Архітектура запропонованої нейромережі наведена на рисунку 3.

Модель починається з шару Embedding, який перетворює вхідний текст у числові вектори заданої розмірності. Потім використовується шар SpatialDropout1D, який випадковим чином «відключає» частину нейронів (20%) для запобігання перенавчанню.

Наступним йде шар Conv1D, який застосовує одинарні згортки до вхідних даних, виявляючи локальні патерни. Після цього модель має двонаправлений шар LSTM, який здатний запам'ятовувати інформацію з обох кінців послідовності. Цей шар також має вбудовані механізми випадкового відключення нейронів для кращої узагальнювальної здатності.

Після шару LSTM йде шар GlobalMaxPooling1D, який обирає максимальні значення з усіх отриманих ознак, зменшуючи розмірність даних. Завершує модель щільний шар Dense з одним нейроном та сигмоїдною функцією активації, який видає ймовірність належності вхідного тексту до класу «Позитивна тональність». Для навчання нейромережі необхідно виконати ще підготовчий етап по роботі з датасетом. Схема навчання нейромережі гібридної архітектури наведено на рис. 4.

Вхідними даними для отримання навченої нейромережевої моделі є навчальна вибірка та ненатренована нейромережева модель гібридної архітектури.

Перший етап відповідає за попередню обробку всіх навчальних текстів, що присутні в датасеті. Попередня обробка включає видалення стоп-слів та знаків пунктуації, а також токенізацію тексту.

На другому етапі відбувається поділ навчальної вибірки на тренувальну та валідаційну у співвідношенні 80 на 20, де 80% тренувальні дані та 20 % валідаційні.

Третій етап відповідає за визначення параметрів навчання нейромережі, таких як кількість епох та розмір партії навчальних зразків. Ці параметри змінюються у залежності від результатів метрик, якщо результати навченої нейронної мережі не задовольняють потребам, змінюються дані параметри та повторюється процес навчання. За замовчуванням кількість епох 5, а розмір навчальної партії 32 зразка.

Наступним етапом відбувається тренування нейромережі із визначенням метрик для кожної епохи навчання. Використовуються метрики accuracy та loss. Це дозволяє оцінювати результат і бачити статистику. Якщо по завершенні епох

accuracy все ще зростає, а loss спадає, це означає що можна спробувати тренування з більшою кількістю епох.

Рисунок 4 – Схема навчання нейромережі гібридної архітектури

П'ятим етапом відбувається збереження навченої моделі та токенізатора, якщо метрики accuracy та loss показують результат, більше 0.9 для метрики accuracy та менше 0.1 для метрики loss. Якщо результату не досягнуто, необхідно повернутись на етап 3 та перенавчити нейромережу.

Вихідними даними є збережена нейромережева модель з оцінками accuracy та loss і токенізатор.

Як показав проведений експеримент, точність нейромережі з гібридною архітектурою CNN та BiLSTM перевищує 97 %, що пояснюється здатністю цієї архітектури ефективно виявляти локальні патерни та враховувати довгострокові залежності в тексті. Графіки статистики навчання використовуваної архітектури наведені на рисунку 5.

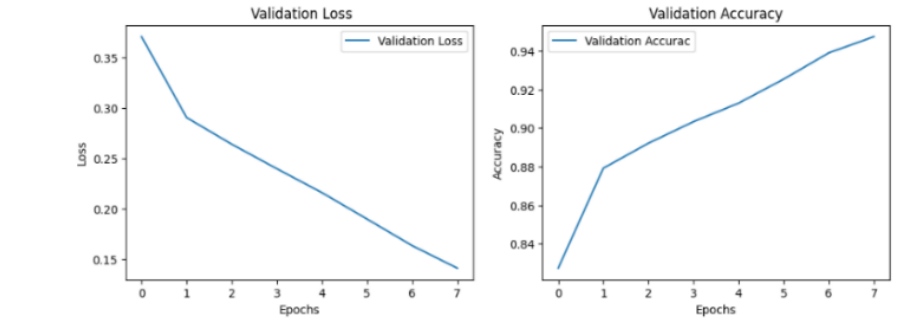

Рисунок 5 – Графіки навчання нейромережі

Окрім високої точності, у порівнянні з реалізованими аналогами (аналоги мають точність 86 – 89 %), протестованими на запропонованому наборі даних, даний метод має надбудову для інтерпретації отриманих рішень. Приклад інтерпретації наведено на рисунку 6.

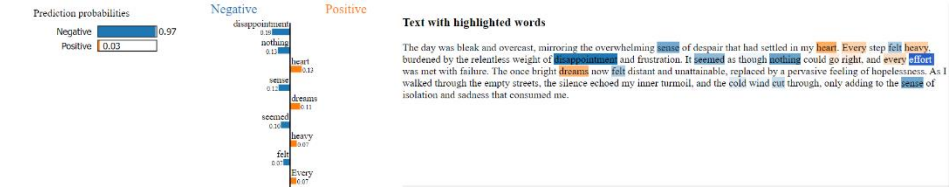

Рисунок 6 – Приклад застосування інтерпретації до результатів нейромережевого виявлення ознак маніпулятивних технік

Як видно з рисунку 6, повідомлення має негативну тональність. Окрім оцінки тональності, представлено слова, які мають вплив на прийняте мережею рішення із вагами. Наприклад, такі слова як «disappointment» (розчарування), «nothing» (нічого), «seemed» (здалося), «felt» (відчувати себе далеким) дійсно мають негативне спрямування.

Отже, запропонований метод візуального пояснення результатів нейромережевого аналізу емоційної тональності повідомлень у соціально-орієнтованих сервісах має перевагу у точності на понад 7% у порівнянні з відомими аналогами, та відрізняється використанням гібридної архітектури CNN та BiLSTM, що дозволяє ефективно виділяти локальні патерни та враховувати довгострокові залежності у тексті. Крім того, пропонується застосування моделі LIME для локальної інтерпретації результатів, що підвищує прозорість та інтерпретованість моделі, що сприяє довірі до нейромережевих підходів

Подальші дослідження будуть спрямовані експерименти з гібридною архітектурою, на кшталт зміни кількості та розмірності шарів, що націлені на підвищення точності ідентифікації. Також планується використання додаткових методі візуалізації з метою підвищення рівня поясненності прийнятих нейромережею рішень.

Перелік посилань

1. Huffaker, J. S., Kummerfeld, J. K., Lasecki, W. S., & Ackerman, M. S. (2020, April). Crowdsourced detection of emotionally manipulative language. In Proceedings of the 2020 CHI conference on human factors in computing systems (pp. 1-14).
2. Leary, R. (2022). Manipulation techniques: Discover covert manipulation techniques to persuade and influence anyone, how to analyze people and Reading body language. BoD–Books on Demand.
3. Bakpokpayev, A., Razaque, A., Kalpeeva, Z., & Ayapbergenova, A. (2025). Combating Social Network Manipulations: A Machine Learning Approach to Enhance Digital Literacy and Emotional Awareness. Computing & Engineering, 3(1), 1-8.
4. Phirangee, K., & Hewitt, J. (2016). Loving this dialogue!!!!: Expressing emotion through the strategic manipulation of limited non-verbal cues in online learning environments. In Emotions, technology, and learning (pp. 69-85). Academic Press.
5. Залуцька О.О., Молчанова М.О., Мазурець О.В., Мельник О.І., Скрипник Т.К. Метод інтелектуального аналізу емоційної тональності текстової інформації для визначення поведінкових намірів нейромережевими засобами. Науковий журнал «Вісник Хмельницького національного університету» серія: Технічні науки. Хмельницький, 2023. №5 (325). Т.1. С. 67-73.
6. Мазурець О.В., Козенко О.В., Собко О.В. Метод автоматизованого підбору відповідей на користувацькі запитання за семантичною подібністю. Матеріали XII Всеукраїнської науково-практичної конференції «Глушковські читання». Київ – 2023. с. 106-109.
7. Shevchuk P., Molchanova M., Mazurets O. Software for Text Messages Reliability Analysis Based on the Machine Learning Models Ensemble. Proceedings of IV International Scientific and Practical Conference «Innovative research and perspectives of the development of science and technology». January 29-31, 2024. Stockholm, Sweden. 2024. Pp. 347-354.
8. Молчанова М.О., Мазурець О.В., Собко О.В., Віт Р.В., Назаров В.В. Алгоритм виявлення аб'юзивного вмісту в україномовному аудіоконтенті для імплементації в об'єктно-орієтовану інформаційну систему. Науковий журнал «Вісник Хмельницького національного університету» серія: Технічні науки. Хмельницький, 2024. №1 (331). С. 101-106.
9. Молчанова М.О., Мазурець О.В., Собко О.В., Кліменко В.І., Андрощук В.І. Метод нейромережевого виявлення кібербулінгу з використанням хмарних сервісів та об'єктно-орієнтованої моделі. Науковий журнал «Вісник Хмельницького національного університету» серія: Технічні науки. Хмельницький, 2024. №2 (333). С. 200-206.
10. Mazurets O., Sobko O., Vit R., Pasternak V. Practical Approach for Detection by Deep Learning of Target Objects of Subject Area Based on Semantic Connectivity Indicators in Audio Database. Proceedings of XXIV International Scientific and Practical Conference «Modern Scientific

Challenges are the Driving Force of the Development of Scientific Research». May 22-24, 2024. Bruges, Belgium. International Scientific Unity. 2024. Pp. 91-96.

11. Mazurets O.V., Sobko O.V., Molchanova M.O., Zalutska O.O., Yurchak A.V. Practical Implementation of Neural Network Method for Stress Features Detection by Social Internet Networks Posts. Global Science: Prospects and Innovations. Proceedings of the II International Scientific and Theoretical Conference «Scientific Review of the Actual Events, Achievements and Problems». May 31, 2024. Berlin, Federal Republic of Germany: International Center of Scientific Research. 2024. Pp. 160-167.

12. Sobko O., Mazurets O., Didur V., Chervonchuk I. Recurrent Neural Network Model Architecture for Detecting a Tendency to Atypical Behavior Of Individuals by Text Posts. Theoretical and Practical Aspects of Modern Research. Proceedings of XXVI International scientific and practical conference. June 5-7, 2024. International Scientific Unity. Ottawa, Canada. 2024. Pp. 113-117.

13. Mazurets O., Molchanova M., Klimenko V., Prosvitliuk M Practice Implementation of Neural Network Model BART-Large-CNN for Text Annotation. Prospects of Scientific Research in the Conditions of the Modern World. Proceedings of XXVII International scientific and practical conference. June 12-14, 2024. Rotterdam, Netherlands. 2024. Pp. 97-102.

14. Мазурець О.В., Віт Р.В. Інтелектуальний метод виявлення цільових об'єктів предметної області за показниками семантичної зв'язності для класифікації текстової інформації. Розвитки інформаційно-керуючих систем та технологій.: монографія. Львів-Торунь : Lina-Pres, 2024. – С.223-244.

15. Mazurets O., Vit R. Practical Application of Method of Thematic Classification of Text Information Using LDA. Information Technology and Implementation (Satellite). Proceedings 11th International Conference. November 21, 2024. Kyiv, Ukraine. 2024. Pp. 151-152.

16. Овчарук О.М., Мазурець О.В. Нейромережеве діагностування проявів ПТСР у текстовому контенті з використанням помилко-орієнтованого навчального набору даних. Науковий журнал «Вісник Хмельницького національного університету» серія: Технічні науки. Хмельницький, 2024. №6, Т.1 (343). С. 195-200.

17. Yurchenko D., Mazurets O., Didur V., Molchanova M. Approach to Using Cloud Services for Visual Analytics of Neural Network Analysis of Texts Emotional Tonality. The Future of Scientific Discoveries: New Trends and Technologies. Proceedings of the XLVII International scientific and practical conference. November 13-15, 2024. Marseille, France. 2024. Pp. 108-113.

18. Юрченко Д.Ю., Мазурець О.В., Залуцька О.О, Безпрозвана Ю.Г. Підхід до візуального пояснення результатів нейромережевого аналізу емоційної тональності повідомлень у соціальних мережах. Збірник наукових праць за матеріалами XVI Всеукраїнської науково-практичної конференції «Актуальні проблеми комп'ютерних наук АПКН-2024». 15-16 листопада 2024. Хмельницький, 2024. с. 565-571.

19. Юрченко Д.Ю., Овчарук О.М., Мазурець О.В., Шевчук П.О. Метод використання нейромережі гібридної архітектури для визначення емоційної тональності текстових повідомлень. Міжнародний науково-технічний журнал «Вимірювальна та обчислювальна техніка в технологічних процесах». № 2, 2025. с. 136-141.