DOI: https://doi.org/10.36910/6775-2524-0560-2025-60-21
УДК 004.415
Ломовацький Антон Андрійович, аспірант
https://orcid.org/0009-0004-5170-3272
Басюк Тарас Михайлович, к.т.н., доцент
https://orcid.org/0000-0003-0813-0785
Національний університет "Львівська політехніка", м. Львів, Україна

ІНТЕРПРЕТОВАНИЙ АНАЛІЗ СЕНТИМЕНТУ ДЛЯ УКРАЇНСЬКОЇ МОВИ, ЩО БАЗУЄТЬСЯ НА ПРАВИЛАХ

Ломовацький А. А., Басюк Т. М. Інтерпретований аналіз сентименту для української мови, що базується на правилах. Обробка природної мови (NLP) значною мірою покладається на аналіз емоційного забарвлення, що дає змогу автоматично оцінювати емоційний тон тексту різними мовами. Однак при використанні з такими мовами, як українська, що мають складну морфологію та синтаксис, відомі інструменти, такі як VADER, часто дають збій. Щоб подолати недоліки загальних моделей, орієнтованих на англійську мову, у цьому дослідженні запропоновано вдосконалений алгоритм аналізу емоційного забарвлення на основі правил, розроблений спеціально для контенту українською мовою. Для покращення виявлення емоційного забарвлення алгоритм використовує розширений лексикон, що включає модифікатори інтенсивності, оцінки полярності, відповідності емоційним знакам та словник емоцій EMOLEX. Для кращого відображення контекстуальних нюансів також використовуються складні техніки, такі як синтаксичний аналіз за залежністю та оцінка з урахуванням позиції. Ці вдосконалення необхідні для правильного розшифрування характерних мовних структур української мови, які часто створюють труднощі для традиційних систем аналізу емоцій. Алгоритм було оцінено за допомогою наборів даних українською мовою та порівняно з VADER. Згідно з результатами, спеціальна модель працює значно краще за VADER, особливо коли йдеться про виявлення сильно позитивних або негативних емоцій. Ці результати підкреслюють цінність мовних інструментів для неангломовного контенту, оскільки вони забезпечують більшу точність і контекстну обізнаність. Хоча результати є обнадійливими, необхідна подальша робота. З метою створення гібридної системи, здатної обробляти все більш складні та неоднозначні вирази з ще вищою точністю, у майбутніх дослідженнях можна розглянути можливість інтеграції технологій штучного інтелекту, таких як машинне навчання та глибоке навчання.
Ключові слова: класифікація сентиментів, виявлення емоцій, алгоритм на основі правил, українська NLP, EmoLex, Vader.

Lomovatskyi A., Basyuk T. Interpreted sentiment analysis for the Ukrainian language based on rules. Sentiment analysis plays a vital role in natural language processing (NLP), enabling automated evaluation of emotional tone in text across multiple languages. Yet, popular tools like VADER often fall short when applied to languages with complex morphology and syntax, such as Ukrainian. This study presents an enhanced rule-based sentiment analysis algorithm tailored specifically for Ukrainian-language content, addressing the limitations of generic, English-centric models. The algorithm incorporates an expanded lexicon that includes the EMOLEX sentiment dictionary, polarity scores, emoji sentiment mappings, and intensity modifiers to improve sentiment detection. It also leverages advanced techniques such as dependency parsing and position-aware scoring to better capture contextual nuances. These enhancements are necessary for correctly deciphering Ukrainian's distinctive linguistic structures, which frequently present difficulties for conventional sentiment analysis systems. The algorithm was evaluated using datasets in the Ukrainian language and compared to VADER. The custom model performs noticeably better than VADER, according to the results, especially when it comes to detecting strongly positive or negative sentiments. Because they provide greater accuracy and contextual awareness, these results emphasise the value of language-specific tools for non-English content. Even though the results are encouraging, more work is necessary. In order to create a hybrid system that can handle increasingly complex and ambiguous expressions with even higher accuracy, future research may investigate integrating AI techniques, such as machine learning and deep learning.
Keywords: Sentiment classification, emotion detection, rule-based algorithm, Ukrainian NLP, EmoLex, Vader.

Постановка проблеми. Обробка природної мови (NLP) значною мірою спирається на аналіз настроїв, що дозволяє виокремлювати з тексту ставлення, почуття та думки. Це важливо в таких сферах, як аналіз відгуків клієнтів, моніторинг соціальних мереж та маркетингові дослідження.

Українська мова стає все більш поширеною в цифровій комунікації, особливо на платформах соціальних мереж, в новинних джерелах та відгуках користувачів, де мільйони україномовних людей по всьому світу. Однак сучасні алгоритми аналізу настроїв стикаються зі значними труднощами через її складну лінгвістичну структуру, яка характеризується багатою морфологією, гнучким синтаксисом, частими запереченнями та ідіоматичними виразами. Більшість популярних інструментів, таких як VADER, при роботі з українськими текстами є менш точними, оскільки вони розроблені переважно для англійської мови і не враховують граматичні та лексичні тонкощі української мови.

Методи аналізу настроїв з часом еволюціонували від простих підходів, заснованих на правилах, до складних моделей глибокого навчання, які можуть ідентифікувати емоції в контексті,

- від простого підходу, заснованого на правилах, до складних моделей глибокого навчання. Хоча ранні методи на основі лексики добре працювали для мов з простою структурою, вони часто не спрацьовують у ситуаціях зі складною морфологією, як-от українська мова. Завдяки останнім досягненням в обробці природної мови з'явилися потужні моделі на основі трансформаторів, такі як Ukr-RoBERTa та Multilingual BERT (mBERT), які використовують глибокі контекстні вбудовування для покращення класифікації настроїв. Навіть з урахуванням цих досягнень, ці моделі продовжують мати проблеми зі специфічними для української мови ідіомами, запереченнями та вираженням почуттів.

NLP для української мови розвинулося, але аналіз настроїв цією мовою все ще перебуває в зародковому стані. Оптимізовані під англійську мову інструменти, які зараз використовуються, недостатньо гнучкі, щоб врахувати лінгвістичні складнощі української мови, такі як різноманітний порядок слів та культурно унікальні способи вираження почуттів. Це підкреслює потребу в більш цілеспрямованих дослідженнях та розробці україномовних інструментів аналізу настроїв [1].

Аналіз останніх досліджень і публікацій. Для підвищення ефективності аналізу настроїв у контексті лінгвістичної складності природної мови, сучасні дослідження зосереджені на інтеграції rule-based підходів із методами машинного навчання. Такі гібридні моделі спрямовані на покращення точності класифікації шляхом поєднання формальної інтерпретованості правил із адаптивністю та узагальнюючою здатністю статистичних моделей. Це дозволяє ефективніше обробляти мовні винятки, включно з контекстуальними модифікаторами, синонімією та прагматичними аспектами. З огляду на те, проведено аналіз основних досліджень.

У статті M. Hu та B. Liu "Mining and Summarizing Customer Reviews" [2] запропоновано метод автоматичного вилучення сентименту з клієнтських відгуків шляхом виявлення ключових атрибутів продукту та асоційованих з ними думок. Методика передбачає класифікацію точок зору як позитивних або негативних на рівні аспектів (aspect-based sentiment analysis) та формування узагальненого представлення тексту шляхом групування емоційно забарвлених фраз. Для реалізації було використано підхід неконтрольованого навчання, що зробив цю роботу однією з перших фундаментальних реалізацій аналізу настроїв з врахуванням особливостей (feature-level sentiment analysis), яка послужила основою для численних наступних досліджень у цій галузі.

Дотичне до аналізу настроїв, проте сфокусоване на людино-машинній взаємодії, дослідження "Determining the Sentiment of Opinions" [3] описує систему розпізнавання жестів, яка дозволяє здійснювати природну та інтуїтивну комунікацію між користувачем і комп'ютером. Хоча дана система безпосередньо не реалізує текстовий аналіз настроїв, вона демонструє загальний вектор розвитку технологій у бік мультиканального сприйняття, включаючи вербальні, невербальні та візуальні сигнали.

У дослідженні "Sentiment Classification Using Machine Learning Techniques" [4] B. Pang, L. Lee та S. Vaithyanathan виконали одне з перших систематичних порівнянь методів машинного навчання для задачі сентимент-класифікації. На прикладі рецензій на фільми вони порівняли ефективність наївного байєсівського класифікатора, моделі максимальної ентропії та машин опорних векторів (SVM). Результати показали, що SVM перевершує альтернативні підходи, продемонструвавши кращу здатність до розділення позитивних і негативних текстів, що зробило цей підхід домінуючим на ранньому етапі розвитку автоматичної сентимент-аналізу.

Виділення невирішених раніше частин загальної проблеми. З огляду на зазначені особливості основними завданнями є:

- Розробити та інтегрувати розширений лексикон настроїв для української мови, що охоплює ідіоматичні вирази, сленг, специфічну для українського контексту лексику, а також елементи цифрової комунікації, зокрема емодзі як засіб вираження емоцій.
- Створити механізм синтаксичного аналізу на основі залежностей, здатний враховувати гнучкий порядок слів в українській мові, точно виявляти контекстуальні зв'язки між словами, інтерпретувати модифікатори настрою (такі як «дуже», «майже», «надзвичайно») та застосовувати позиційне зважування для коригування інтенсивності емоційної оцінки.
- Провести порівняльну оцінку ефективності запропонованої моделі аналізу настроїв шляхом кількісного та якісного аналізу результатів у порівнянні з алгоритмом VADER з метою виявлення переваг підходу, що базується на правилах, у порівнянні з методами глибокого навчання для текстів українською мовою.

Вирішення зазначених задач дозволить підвищити точність задачі сентимент-аналізу для

україномовного текстового контенту за рахунок адаптації моделей до особливостей морфологічно складної та малоресурсної мови. Результати дослідження сприятимуть розвитку інструментарію обробки природної мови (NLP) для мов із високим рівнем флексії та вільним порядком слів. Запропонована методологія сформує концептуальну базу для побудови гібридних систем, що комбінують правила та методи машинного навчання для досягнення ще вищої точності.

Виклад основного матеріалу дослідження. Для того, щоб висвітлити основні ідеї теми дослідження, було створено схему, що описує основні кроки, необхідні для реалізації системи аналізу настроїв (див. рис. 1).

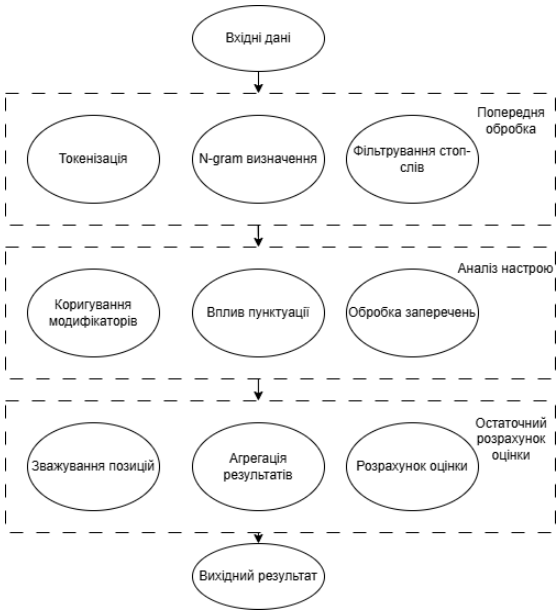

Рис. 1. Послідовність етапів аналізу настроїв

Процес аналізу настроїв у запропонованій системі реалізується через три ключові етапи обробки:

1. Попередня обробка текстових даних – на цьому етапі виконується очищення вхідного тексту від шумових елементів, таких як пунктуація, стоп-слова, HTML-теги та інші нефункціональні символи. Також здійснюється токенізація, нормалізація (лематизація або стемінг) та приведення тексту до уніфікованого вигляду, придатного для подальшого аналізу.
2. Аналіз настрою – центральний етап, на якому оцінюється емоційний тон кожного токена. Оцінювання проводиться з врахуванням контекстуальних модифікаторів, що дозволяє скоригувати емоційну полярність окремих лексем.
3. Агрегація результатів – на завершальному етапі здійснюється об'єднання всіх оцінок настрою в єдиний кінцевий результат.

Основою запропонованого аналізу настроїв на основі правил виступає спеціалізований лексикон настроїв, адаптований для української мови, який враховує морфологічну варіативність та лексичні особливості. Центральним ресурсом є розширена україномовна версія EMOLEX, де кожне слово анотовано відповідно до емоційних категорій (радість, гнів, страх, довіра тощо) та загальної тональності (позитивна, негативна, нейтральна). Для покращення охоплення термінів, зокрема рідковживаних або доменно-специфічних, до системи інтегрується лексикон полярності, який надає числову шкалу емоційної валентності – від сильно негативної до сильно позитивної, що дозволяє виконувати градацію настроїв.

Додатково реалізовано семантичне мапування емодзі: кожному емодзі призначено категорію тональності (позитивна, негативна, нейтральна), що суттєво покращує точність аналізу в умовах цифрового неформального спілкування, типового для соціальних медіа. Наприклад, емодзі " 😂 " та "❤️" трактуються як позитивні, " 😢 " та " 😡 " – як негативні, а " 🙂 " – як нейтральні.

Для підвищення точності семантичного аналізу неформального цифрового тексту запропонований алгоритм інтегрує емодзі-мепінг, який виконує категоризацію емодзі відповідно до їх емоційної полярності. Кожен емодзі класифікується як позитивний (наприклад, 😂, ❤️), негативний ("😢", "😡") або нейтральний (🙂). Це дозволяє системі коректно інтерпретувати зміст

емоційно-насичених повідомлень у соціальних мережах.

Крім того, до системи включено підсистему обробки інтенсивіфікаторів, яка оперує лексиконом посилювачів і пом'якшувачів емоційного тону. Наприклад, слова типу дуже, значно підсилюють емоційну валентність, тоді як трохи, дещо — знижують її. Цей модуль виконує вагове масштабування базових оцінок полярності відповідно до сили впливу модифікатора.

Для обробки ідіоматичних виразів і стійких словосполучень застосовується попередньо скомпільований лексикон фразової семантики, що дозволяє уникнути хибних оцінок у випадках, коли значення виразу не можна вивести зі значень його складових частин на кшталт "на межі розпачу" ("на межі відчаю").

Крім того, високочастотні неінформативні слова, такі як "і" ("і") та "або" ("або"), відфільтровуються за допомогою уточненого списку українських стоп-слів. Алгоритм зосереджує увагу на змістових термінах, вилучаючи їх з аналізу, що підвищує точність і ефективність обчислень.

Важливим етапом у запропонованому алгоритмі аналізу настроїв на основі правил є точна попередня обробка тексту. Цей етап гарантує збереження лінгвістичних та контекстуальних нюансів української мови при перетворенні неструктурованого вхідного тексту у формат, придатний для аналізу.

Токенізація - це перший крок у процесі попередньої обробки, на якому вхідний текст розбивається на менші одиниці, які називаються токенами. Як токени можуть використовуватися смайлики, слова та розділові знаки. Щоб ефективно обробляти вхідні дані українською мовою, процес токенізації модифікується шляхом: збереження морфологічної структури складних слів, збереження розділових знаків, таких як "!" та ".", які згодом досліджуються на предмет їхнього впливу на емоції, відокремлення та розпізнавання емодзі, які розглядаються як окремі компоненти, що передають емоції [6].

Токенізація речення "Це було неймовірно красиво, але трохи сумно ☹!" виглядає наступним чином: ["Це", "було", "неймовірно", "красиво", ",", "але", "трохи", "сумно", "☹", "!"]

Алгоритм використовує аналіз N-грам, щоб вловити контекстний настрій та ідіоматичні вирази на додаток до окремих значень слів. N-грами, які є наборами з N послідовних лексем, корисні для пошуку сентиментальних багатослівних фраз:

- Уніграми досліджують окремі слова, такі як "красиво" (що означає "гарно"),
- Контекстуальні пари "трохи сумно" (трохи сумно) вловлюються біграмами.
- Триграми здатні розпізнавати більш складні вирази (наприклад, "на межі розпачу" або "на межі відчаю").

Алгоритм здатен розшифровувати емоційне значення фраз, а також сенс окремих термінів, поєднуючи токенізацію та аналіз N-грам. Це значно покращує його розуміння тонких лінгвістичних закономірностей, які не були б помічені при словесному підході.

Запропонований алгоритм аналізу настроїв включає модуль аналізу залежностей, який враховує лінгвістичну складність української мови. Визначаючи граматичні зв'язки між словами, цей компонент допомагає алгоритму краще зрозуміти контекст і взаємодію в кожному реченні. Досліджуючи синтаксичні залежності, система може краще обробляти важливі мовні елементи, що впливають на настрій, такі як заперечення, модифікатори та розділові знаки.

На полярність настрою суттєво впливає, зокрема, заперечення. Алгоритм розпізнає заперечення, такі як "не" або "ні", і відповідно змінює значення слів, які вони модифікують. Це гарантує точну інтерпретацію емоційного тону речення, навіть якщо заперечення стоять перед словами, що несуть смислове навантаження.

Приклад:

- Вхідні дані: "Це не гарно".
- Без обробки заперечення. Позитивне завдяки "гарно".
- З обробкою заперечення. Заперечення через "не".

Алгоритм використовує синтаксичні залежності, щоб зв'язати заперечення з цільовими словами, гарантуючи точну зміну настрою.

Наприклад:

- *S(w)* – оцінка настрою слова w_t.
- w_{neg} слово-заперечення (наприклад: "не", "ні").
- $dep(w_{neg}, w_t)$ – синтаксична залежність, що зв'язує w_{neg} з цільовим словом w_t.

- $S'(w_t)$ – скоригована оцінка настрою цільового слова w_t.

Скоригований показник настроїв розраховується за формулою 1:

$$S'(w_t) = -k \cdot S(w_t) \quad (1)$$

Де:

- k – коефіцієнт підсилення (наприклад, k =1.5) щоб збільшити ефект заперечення.

Розділові знаки, такі як знаки оклику ("!") та багатокрапки ("..."), часто передають додатковий емоційний контекст. Алгоритм коригує оцінку настрою на основі наявності та типу розділових знаків:

- *Знаки оклику.* Підсилюють інтенсивність емоцій. Приклад: "Це чудово!" має вищий бал за рахунок знаку оклику.
- *Зменшують інтенсивність висловлювання,* вказуючи на вагання або невпевненість. Приклад: "Це цікаво..." має нижчу оцінку за рахунок багатокрапки.

Нехай:

- $S(w)$ – оцінка настрою слова w.
- P розділовий знак, пов'язаний з реченням (наприклад, "!" або "...").
- $S'(w)$ – скоригована оцінка сентименту слова www.
- $\gamma(P)$ – *пунктуаційний множник*, де: $\gamma("!") > 1$ (посилює інтенсивність сентименту), $0 < \gamma("...") < 1$ (зменшує інтенсивність сентименту).

Скоригований показник настроїв розраховується як формула 2:

$$S'(w) = \gamma(P) * S(w) \quad (2)$$

Де:

- $\gamma(!) = 1.5$ (приклад коефіцієнта посилення для знаків оклику).
- $\gamma("...") = 0.8$ (приклад коефіцієнта зменшення для еліпсів).

Алгоритм використовує позиційні ваги та множники інтенсивності для подальшого вдосконалення класифікації настрою, що дозволяє йому більш точно змінювати емоційний вплив слів і фраз. Цей метод гарантує, що при визначенні загального настрою враховується як синтаксична позиція, так і семантичне значення слів у реченні [7].

Використання підсилювачів настрою - слів, які збільшують або зменшують емоційну інтенсивність пов'язаних з ними термінів - є важливим покращенням. Цим модифікаторам алгоритм надає попередньо встановлену вагу відповідно до ступеня їхнього впливу та напряму. Наприклад, сила емоційності суміжних слів збільшується за допомогою таких підсилювачів, як "дуже" або "абсолютно". Наявність "дуже" в реченні "Це дуже гарно" ("Це дуже красиво") збільшує оцінку "гарно" в 1,5 рази. Зменшувачі "трохи" (що означає "трохи") або "майже" (що означає "майже"), навпаки, зменшують інтенсивність оцінки. Наприклад, оцінка почуття "сумно" зменшується на 0,7 у фразі "Це трохи сумно" ("Це трохи сумно").

Для визначення цих модифікаторів використовується синтаксичний аналіз залежностей, що дозволяє алгоритму динамічно коригувати оцінку настрою в контексті, застосовуючи відповідні множники, засновані на синтаксичних зв'язках.

Розміщення слова або словосполучення в реченні може мати великий вплив на його значення на додаток до лінгвістичних модифікаторів. Алгоритм розподіляє позиційні ваги, щоб врахувати це:

- Слова, що починають речення, часто вважаються тематично значущими, тому їм присвоюється більша вага. Наприклад, слово "Чудово" підкреслено через його позицію на початку речення в "Чудово, але трохи складно".
- Слова в кінці речення мають дещо більшу вагу, ніж слова в середині, і зазвичай справляють більше враження. У "Це було добре, але складно" слово "Складно" має більшу вагу через його останню позицію.

Алгоритм покращує контекстну точність і деталізацію оцінки настроїв, поєднуючи позиційне зважування та розпізнавання бустерів.

Нехай:

- $S(w)$ – оцінка настрою слова w.
- $W_{pos}(w)$ – позиційна вага присвоєна слову w, виходячи з його позиції в реченні.
- $W_{pos}(w_{start}) > W_{pos}(w_{end}) > W_{pos}(w_{middle})$ – більша вага присвоюється початковій та кінцевій позиції).

- $S'(w)$ – скоригована оцінка настрою слова w.

Скоригована оцінка настрою розраховується як (формула 3):

$$S'(w) = W_{pos}(w) * S(w) \quad (3)$$

Де:

- $W_{pos}(w_{start}) = 1.5$ (приклад ваги слів спочатку речення).
- $W_{pos}(w_{end}) = 1.2$ (приклад ваги слів в кінці речення).
- $W_{pos}(w_{middle}) = 1.0$ (приклад ваги слів в середині речення).

Можна створити власну систему аналізу настроїв для україномовного контенту, використовуючи лексичні словники, інструменти попередньої обробки, алгоритми аналізу настроїв та інструменти для створення результатів.

Ефективність запропонованого алгоритму аналізу настроїв на основі правил було оцінено з використанням VADER (Valence Aware Dictionary and Sentiment Reasoner) як базової лінії. Його власний дизайн на основі правил, який включає такі компоненти, як пунктуація, емодзі та модифікатори інтенсивності - функції, які дуже схожі на ті, що є в користувацькому алгоритмі, - робить його підходящим еталоном. Однак безпосереднє застосування VADER до українських текстів створює значні проблеми, оскільки він оптимізований насамперед для англомовного контенту [8].

На вибір VADER для порівняння вплинула низка факторів. Перш за все, це перевірений метод, який часто використовується як у галузевих, так і в наукових дослідженнях, особливо для вивчення коротких текстів у соціальних мережах. Це популярний варіант для завдань аналізу настроїв завдяки своїй ефективності в обробці неформальної цифрової комунікації.

По друге, VADER пропонує підхід, який є чітким і простим для розуміння. Він ідеально підходить для порівняння методів, заснованих на правилах, без складності та непрозорості, які часто притаманні моделям глибокого навчання, оскільки, як і запропонована система, він базується на попередньо визначених лексиконах та правилах оцінювання.

Для оцінки обох алгоритмів використовувався один і той самий набір даних україномовних текстів, який містив попередньо позначені зразки, класифіковані як позитивні, нейтральні або негативні. Для оцінки відносної ефективності обох моделей використовували однакові вхідні дані, щоб забезпечити об'єктивне порівняння.

Оцінювання проводилося за допомогою методичної, багатоступеневої процедури у кілька етапів:

1. Попередня обробка. З усіх текстів було видалено зайві пробіли та спеціальні символи. Перед обробкою українські тексти були перекладені за допомогою бібліотеки translate з Python, оскільки VADER призначений лише для англійської мови. З іншого боку, унікальний алгоритм зберіг лінгвістичну структуру оригінальних україномовних текстів, працюючи безпосередньо з ними.
2. Класифікація настроїв. Обидва алгоритми виводили складний бал, що відображає загальну інтенсивність настроїв, після того, як вони окремо класифікували кожен текст за однією з трьох категорій настроїв: позитивний, нейтральний або негативний.
3. Метрики для оцінювання. Ефективність оцінювали за допомогою низки кількісних метрик:
    a. Точність: Відсоток правильно класифікованих зразків.
    b. Оцінка F1: Середнє гармонійне значення точності та пригадування для кожного класу настроїв.
    c. Результати моделі порівнюються з людськими анотаціями настроїв, використовуючи середній складний бал, який є середньою оцінкою настрою для кожної категорії.
4. Аналіз помилок був використаний для вивчення неправильно класифікованих зразків, щоб знайти тенденції та граничні випадки, коли один алгоритм працював краще, ніж інший, надаючи інформацію про переваги та недоліки кожної моделі.
5. Візуалізація. Статистичні зведення середніх комплексних оцінок за обома моделями відображалися поряд із гістограмами, які показували розподіл настроїв (позитивні, нейтральні та негативні).

Для більш наочного зображення процесу порівняння було створено контекстну діаграму (рис. 2).

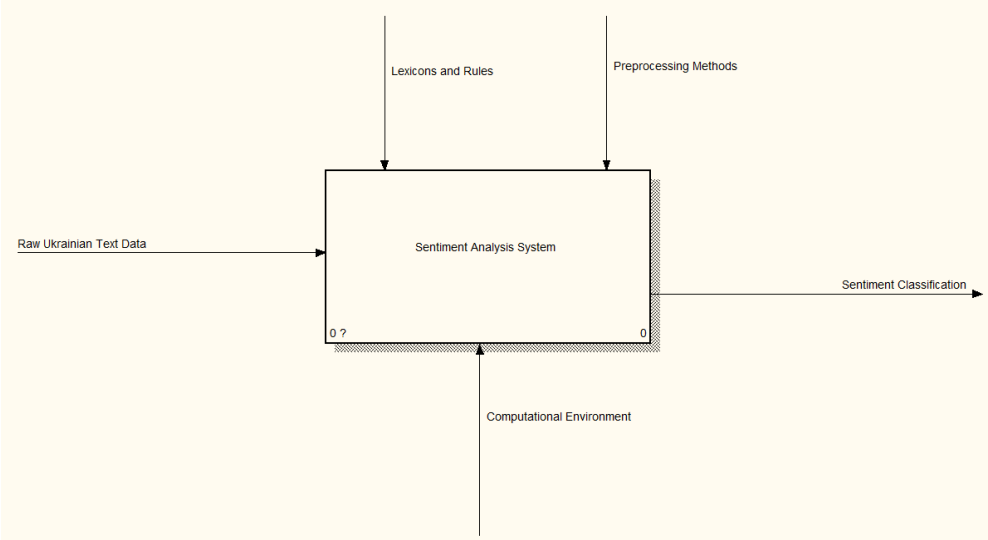

Рис. 2. Контекстна діаграма системи

Поєднуючи ці вхідні дані та елементи керування, система аналізу настроїв створює точні класифікації настроїв, використовуючи лінгвістичні ресурси та заздалегідь встановлені правила. Контекстну діаграму було розбито на кілька підпроцесів, щоб краще зрозуміти робочий процес аналізу настроїв (рис. 3).

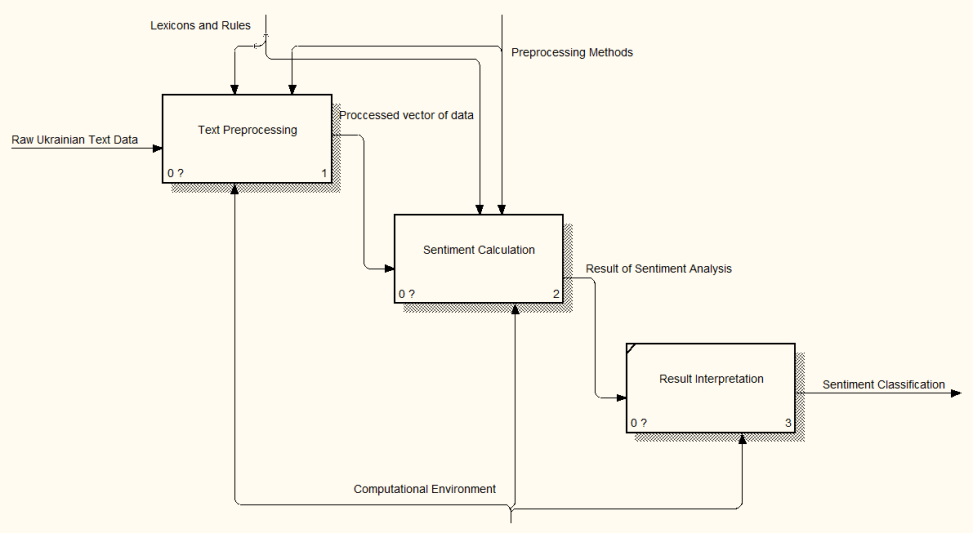

Рис. 3. Діаграма декомпозиції системи

Система показана на діаграмі декомпозиції як серія пов'язаних кроків, кожен з яких відіграє унікальну роль у робочому процесі аналізу настроїв:

- Попередня обробка тексту. Цей перший крок відповідає за перетворення неструктурованого українського тексту в акуратний та організований формат. Використовуючи такі методи, як токенізація, видалення стоп-слів та нормалізація, він готує вхідні дані для подальшого аналізу.
- Обчислення настроїв. Після завершення попередньої обробки очищений текст надсилається до модуля обчислення настроїв. Після контент-аналізу цей компонент генерує результат настрою, який може бути виражений категорично (наприклад, позитивний, нейтральний або негативний) або чисельно (у вигляді балів).
- Інтерпретація результатів. Обчислені результати опитування остаточно трансформуються у формат, який може бути прочитаний людиною. Після цього користувачеві або показують ці класифікації безпосередньо, або вони надсилаються до інших систем для подальшої інтеграції чи обробки.

Розроблена система має VaderAnalyzer для порівняльного аналізу, який використовує метод аналізу настроїв VADER, що особливо корисно після перекладу українського тексту на англійську

мову. Усе це об'єднує SentimentComparer, який порівнює результати VaderAnalyzer і користувацького SentimentAnalyzer. Крім того, він керує процедурами перекладу і використовує графіки для відображення розподілу настроїв, що спрощує порівняння ефективності двох підходів.

Для порівняння цих двох підходів було взято два набори даних, що містять слова "добре" і "погано" відповідно. Для слова "добре", на рисунку 4 показано, що 83% твітів зі словом "добре" були класифіковані як позитивні користувацьким аналізатором, порівняно з 58%, класифікованими VADER. Крім того, середній складний бал користувацької моделі становив 0,66, що більш ніж удвічі вище, ніж у VADER (0,28). Цей результат демонструє, як інтеграція специфічної української лексики та вагових коефіцієнтів інтенсивності в кастомний алгоритм підвищує його здатність визначати контекстуальну позитивність.

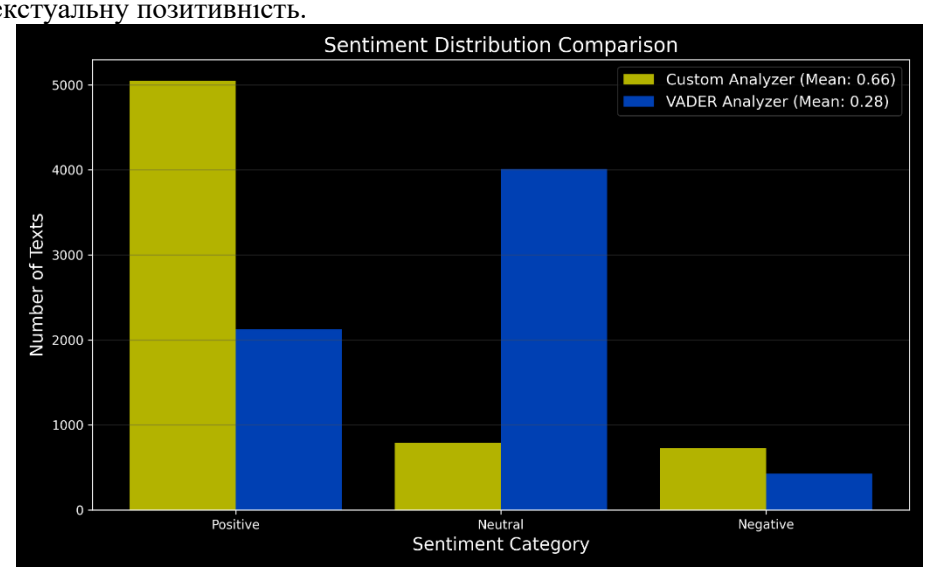

Рис. 4. Порівняння розподілу настроїв для набору даних зі словом "добре"

У випадку зі словом "погано" користувацький аналізатор також перевершив VADER у виявленні негативних настроїв. Як показано на рисунку 5, він класифікував 65% твітів як негативні, порівняно з 49% у VADER. Більше того, середній складний бал, отриманий користувацькою моделлю, становив -0,44, що свідчить про сильніші негативні настрої і більше відповідає фактичному емоційному тонусу, тоді як середній бал VADER становив лише -0,22. Це свідчить про вищу чутливість кастомного алгоритму до негативних висловлювань в україномовному контенті.

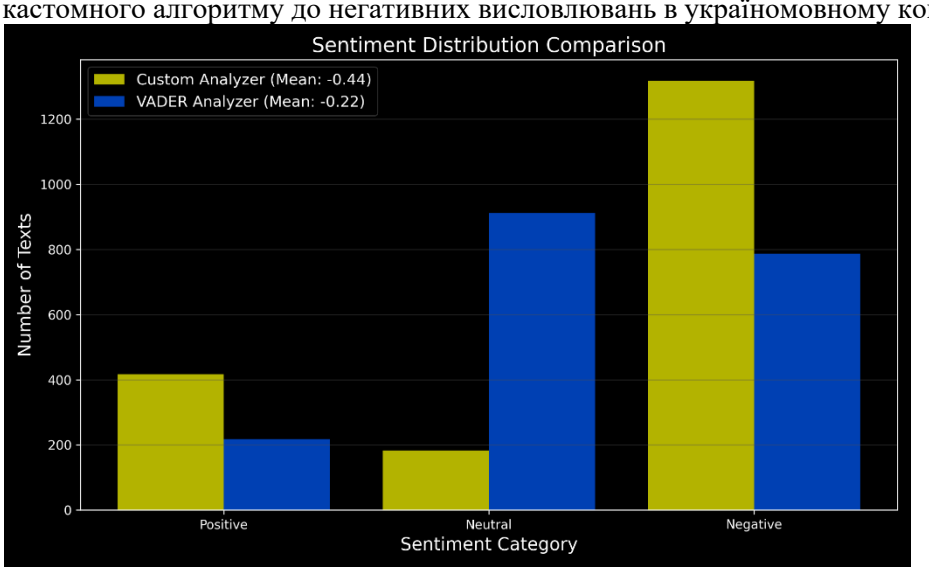

Рис. 5. Порівняння розподілу настроїв для набору даних зі словом "погано"

Експериментальні результати, наведені в Таблиці 1, дають змогу порівняти результати роботи аналізатора VADER і користувацького аналізатора. Користувацька модель на основі правил демонструє стабільну перевагу в продуктивності над VADER на всіх протестованих наборах даних,

як показано в Таблиці 1. Модель демонструє чудове виявлення сильних позитивних і негативних емоцій завдяки кільком важливим перевагам:

- Алгоритм інтегрує численні лінгвістичні ресурси, які включають лексику EMOLEX і бази даних оцінок полярності, а також модифікатори інтенсивності та велике сховище фраз, що містять емоції. Розширений лексичний запас забезпечує кращі можливості для аналізу настрою.
- Кастомний аналізатор використовує синтаксичний розбір залежностей і зважування на основі позиції, щоб правильно обробляти мовні особливості, такі як заперечення, модифікатори і розділові знаки. Впровадження цих методів призводить до суттєвого покращення оцінки настрою в контекстному контексті.
- Система точно інтерпретує неформальні елементи цифрового тексту завдяки впровадженню лексикону емоцій емодзі. Ця можливість є забутою функцією, яку традиційні алгоритми не враховують.

Поєднання цих переваг дає змогу кастомному алгоритму ефективно обробляти специфічні особливості україномовного контенту та краще аналізувати сучасні емоційні тексти.

Таблиця 1. Результати експерименту

Слово	Custom Analyzer (Позитивний/Негативний %)	VADER (Позитивний /Негативний %)	Custom середній результат	VADER середній результат
добре	83% (позитивний)	58% (позитивний)	0.66	0.28
погано	65% (негативний)	49% (негативний)	-0.44	-0.22

Висновки та перспективи подальших досліджень. У даному дослідженні представлено новий алгоритм аналізу настроїв на основі правил, розроблений для української мови. Мапування емоцій за допомогою емодзі, оцінки полярності, модифікатори інтенсивності та лексикон EMOLEX - це лише деякі з лінгвістичних ресурсів, які інтегровані в запропонований метод для успішного подолання особливих труднощів, пов'язаних з аналізом емоцій в українській мові. Здатність моделі керувати складними граматичними структурами та контекстуальними варіаціями ще більше покращується завдяки додаванню складних функцій, таких як синтаксичний аналіз залежностей та підрахунок балів з урахуванням позиції.

Ці результати підкреслюють важливість індивідуальних лінгвістичних підходів у розвитку аналізу настроїв для недостатньо представлених мов і закладають міцний фундамент для майбутніх інновацій у цій галузі. Хоча продуктивність алгоритму є багатообіцяючою, він пропонує кілька можливостей для подальшого розвитку. Порівняльний аналіз підтверджує, що добре розроблений алгоритм, який базується на правилах для конкретного домену і підтримується структурованими лексиконами та точними лінгвістичними правилами, може конкурувати з такими популярними інструментами, як VADER.

Незважаючи на обнадійливі результати, алгоритм має низку перспектив для подальшого розвитку:

- Інтеграція з технологіями штучного інтелекту. Здатність моделі, заснованої на правилах, обробляти неоднозначні або контекстуально нюансовані вирази може бути покращена шляхом поєднання її з методами машинного навчання або глибокого навчання. Найкращі результати можуть дати гібридні моделі, які поєднують статистичні та символьні методи.
- Поширення на інші мови. Застосовність підходу та його суспільний вплив можна збільшити, адаптувавши його до інших малоресурсних або морфологічно складних мов після успіху в українській мові.
- Динамічне оновлення лексики. Система може адаптуватися до нового сленгу, тенденцій та специфічної мови домену шляхом автоматичного оновлення лексики за допомогою веб-скрепінгу та обробки природної мови.

З огляду на перелічені чинники, дане дослідження робить значний внесок у розробку більш точного та зрозумілого аналізу настроїв для української мови. Воно підкреслює, як лінгвістична структура та нові можливості штучного інтелекту можуть бути поєднані для створення потужного балансу між інтерпретованістю та адаптивністю, що є ключовою метою в сучасній обробці

природної мови.

Список бібліографічного опису

1. E. Riloff, J. Wiebe, Learning Extraction Patterns for Subjective Expressions, Матеріали конференції 2003 року з Empirical Methods in Natural Language, 2003, pp. 105-112.
2. M. Hu, B. Liu, Mining and Summarizing Customer Reviews, Матеріали конференції з 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2004, pp. 168-177.
3. S. M. Kim, E. Hovy, Determining the Sentiment of Opinions, Матеріали конференції з 20th International Conference on Computational Linguistics, 2004, pp. 1367-1373.
4. B. Pang, L. Lee, S. Vaithyanathan, Thumbs Up? Sentiment Classification Using Machine Learning Techniques, Матеріали конференції з ACL-02 Conference on Empirical Methods in Natural Language Processing, 2002, pp. 79-86.
5. Basyuk T., Vasyliuk A. Approach to a subject area ontology visualization system creating // CEUR Workshop Proceedings. – 2021. – Vol. 2870: Матеріали конференції з the 5th International conference on computational linguistics and intelligent systems (COLINS 2021), Lviv, Ukraine, April 22–23, 2021. Том I: основна конференція. – P. 528–540.
6. S. M. Kim, E. Hovy, Identifying and Analyzing Judgment Opinions, Матеріали конференції з the Human Language Technology Conference of the NAACL, 2006, pp. 200-207.
7. Y. Kim, Convolutional Neural Networks for Sentence Classification, Матеріали конференції з 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014, pp. 1746-1751.
8. P. Bojanowski, E. Grave, A. Joulin, T. Mikolov, Enriching Word Vectors with Subword Information, Transactions of the Association for Computational Linguistics, 5, 2017, pp. 135-146.

References

1. E. Riloff, J. Wiebe, Learning Extraction Patterns for Subjective Expressions, Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing, 2003, pp. 105-112.
2. M. Hu, B. Liu, Mining and Summarizing Customer Reviews, Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2004, pp. 168-177.
3. S. M. Kim, E. Hovy, Determining the Sentiment of Opinions, Proceedings of the 20th International Conference on Computational Linguistics, 2004, pp. 1367-1373.
4. B. Pang, L. Lee, S. Vaithyanathan, Thumbs Up? Sentiment Classification Using Machine Learning Techniques, Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing, 2002, pp. 79-86.
5. Basyuk T., Vasyliuk A. Approach to a subject area ontology visualization system creating // CEUR Workshop Proceedings. – 2021. – Vol. 2870: Proceedings of the 5th International conference on computational linguistics and intelligent systems (COLINS 2021), Lviv, Ukraine, April 22–23, 2021. Volume I: main conference. – P. 528–540.
6. S. M. Kim, E. Hovy, Identifying and Analyzing Judgment Opinions, Proceedings of the Human Language Technology Conference of the NAACL, 2006, pp. 200-207.
7. Y. Kim, Convolutional Neural Networks for Sentence Classification, Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014, pp. 1746-1751.
8. P. Bojanowski, E. Grave, A. Joulin, T. Mikolov, Enriching Word Vectors with Subword Information, Transactions of the Association for Computational Linguistics, 5, 2017, pp. 135-146.