МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
Київський національний університет імені Тараса Шевченка
Навчально-науковий інститут філології
Кафедра української мови та прикладної лінгвістики

Тональний аналіз українськомовних текстів

Кваліфікаційна робота бакалавра
студентки IV курсу
галузі знань 03 "Гуманітарні науки"
спеціальності 035 "Філологія",
спеціалізації 035.10 "Прикладна лінгвістика",
ОПП "Прикладна (комп'ютерна) лінгвістика
та англійська мова",
Марії Михайлівни ОНИЩЕНКО

Науковий керівник:
доктор філософії,
доцент кафедри української мови
та прикладної лінгвістики
Юлія ЦИГВІНЦЕВА

"Допущено до захисту"
Протокол засідання кафедри
української мови та прикладної лінгвістики
від _____ № ___
Завідувач кафедри _____ **Сергій РІЗНИК**

КИЇВ — 2025

Антоація

Кваліфікаційна робота присвячена актуальному завданню сучасної прикладної лінгвістики — аналізу тональності українськомовних текстів. Актуальність дослідження зумовлена стрімким зростанням обсягів текстових даних у цифровому просторі, зокрема відгуків, дописів у соціальних мережах і коментарів, що потребують автоматизованої обробки для виявлення емоційного ставлення автора. Об’єктом дослідження виступає тональність українськомовного тексту, а предметом — словниковий підхід і методи машинного навчання для її визначення. Метою роботи є дослідження можливостей словникового методу в поєднанні з сучасними нейромережевими моделями, аналіз релевантних лінгвістичних правил та створення українськомовного тонального датасету. У ході дослідження було розглянуто теоретичні засади сентимент-аналізу, специфіку синтаксичної структури української мови, особливості роботи з запереченнями та інтенсифікаторами. Методологічну основу склали аналіз лексикографічного ресурсу О. Толочко, розробка візуального тонального парсера, вебскрапінг та фільтрація відгуків, а також навчання моделі з використанням фреймворку TensorFlow. Результатом дослідження є створення гібридної системи аналізу тональності, яка поєднує словникову основу та нейромережевий підхід, а також розробка візуалізації тональних дерев. Новизна роботи полягає у синтаксично орієнтованому підході до визначення полярності, де вагомість лексем визначається залежно від їхнього синтаксичного зв’язку з присудком, а також у створенні українськомовного тонального датасету відгуків. Практична цінність проєкту полягає в можливості використання результатів як у сфері автоматизації бізнес-аналітики, так і в подальших наукових розробках з обробки природної мови.

Ключові слова: українська мова, тональність, аналіз тональності, словниковий підхід, машинне навчання, синтаксичний аналіз, тональні дерева, нейромережа, TensorFlow.

Abstract

This thesis is devoted to a topical issue in contemporary applied linguistics: the analysis of the sentiment of Ukrainian-language texts. The relevance of this research is driven by the rapid growth of text data in the digital space, particularly reviews, social media posts, and comments, which require automated processing to identify the author's emotional attitude. The object of the study is the tonality of Ukrainian-language text, and the subject is the lexical approach and machine learning methods for determining it. The aim of the work is to study the possibilities of the lexical method in combination with modern neural network models, analyze relevant linguistic rules, and create a Ukrainian-language tonal dataset. The study examined the theoretical foundations of sentiment analysis, the specifics of the syntactic structure of the Ukrainian language, and the peculiarities of working with negations and intensifiers. The methodological basis consisted of analyzing O. Tolochko's lexicographic resource, developing a visual tonal parser, web scraping and filtering reviews, and training the model using the TensorFlow framework. The result of the research is the creation of a hybrid sentiment analysis system that combines a dictionary-based and neural network approach, as well as the development of tonal tree visualization. The novelty of the work lies in a syntactically oriented approach to determining polarity, where the weight of lexemes is determined depending on their syntactic connection with the predicate, as well as in the creation of a Ukrainian-language tonal dataset of reviews. The practical value of the project lies in the possibility of using the results both in the field of business analytics automation and in further scientific developments in natural language processing.

Keywords: Ukrainian language, tonality, sentiment analysis, lexical approach, machine learning, syntactic analysis, sentiment trees, neural network, TensorFlow.

ЗМІСТ

ВСТУП

Актуальність дослідження. Аналіз тональності є завданням сентимент-аналізу (або майнінгу думок) — дослідження, ключовим у якому є обчислювальний підхід і яке спрямоване на дослідження настроїв, емоцій та ставлення людей до об’єктів реальності, таких як: продукти, послуги, проблеми, події, теми, їхні характеристики тощо. Це, у певному сенсі, мультидисциплінарна галузь, що охоплює психологію, соціологію, опрацювання природної мови та машинне навчання.

Сентимент-аналіз уможливлює відстежування узагальненого ставлення певної групи людей (яка визначається власне потребами конкретних застосувань) до певного об’єкта реальності з метою накопичення та подальшого прикладного використання отриманих даних у спеціальних галузевих дослідженнях (наприклад, соціологічних, психологічних тощо), бізнесі (аналіз рівня задоволеності продуктом за відгуками тощо).

Інтенсивне зростання кількості інтернет-платформ для обговорень, вебсайтів з оглядами на продукти, популяризація електронної комерції та соціальних мереж результує стабільним масштабуванням потоку думок. Це перманентне збільшення даних ускладнює отримання адекватної оцінки, загального ставлення до певного явища чи об’єкта. Майнінг думок наразі є звичайним інструментом у репертуарі аналізу соціальних мереж, який проводять компанії, маркетологи та політичні аналітики (Bloom, 2011). Інформація вилучається з власне слів у тексті, з контексту цих слів і з лінгвістичної структури тексту.

Тональність можна схарактеризувати як позитивну чи негативну оцінку явища, виражену, як правило, у текстовій формі (і ця кваліфікаційна робота зосереджена власне на текстовій формі). Тональність всього тексту в цілому можна визначити як функцію (в найпростішому випадку — суму)

лексичних тональностей його складників (речень) і правил їх поєднання (Mondher, Tomoaki, 2016). Найпопулярнішим застосуванням аналізу тональності є автоматичне визначення того, чи є відгук, опублікований в інтернеті (про фільм, книгу, споживчий продукт тощо), позитивним чи негативним. Також у деяких роботах допускають нейтральну оцінку.

Відповідні дані про продукт чи послугу, подані природною мовою, знаходяться у вільному доступі в Інтернеті. Ці нові джерела інформації є надзвичайно важливими для оптимізації маркетингових стратегій. Для того, щоб обробляти великі обсяги даних і видобувати цінну інформацію, необхідні методи інтелектуального аналізу даних і опрацювання природної мови. Оскільки процес вилучення цієї таргетної інформації вимагає глибокого розуміння явних і неявних, регулярних і нерегулярних, синтаксичних і семантичних особливостей мови, аналіз неструктурованих вебресурсів є досить складним завданням (Ligthart, Catal, Tekinerdogan, 2021).

А через складність словотвору слов'янських мов використання систем опрацювання природної мови загалом для розв'язання задач ускладнюється ще більше (Бабаскіна, 2019). Розгалужена система флексій в українській мові суттєво знижує ефективність методів, які працюють для германських мов (насамперед англійської як основної мови світових наукових досліджень) просто тому, що вони не враховують критично важливу інформацію, закладену в ці флексії — синтаксичні зв'язки. У теоретичній частині це розглянуто детальніше, адже глобально синтаксис є наріжним каменем в розподілі тональних коефіцієнтів.

Саме тому дослідження в цій галузі є цінним внеском і в розвиток опрацювання природної мови в цілому, і в україномовне програмне забезпечення різноманітних сфер, як наукового, так і бізнес-спрямування. Важливо зазначити, що серед джерел дослідження є й праці українських науковців.

Наразі існує дуже обмежена кількість тональних словників української мови. Як приклад можна навести "Тональний словник української мови" О. Сивоконя. Цей словник було укладено у 2016 році, він містить близько 3500 базових форм слів ненейтральної тональності (Сивоконь, 2016). Тобто, по-перше, тональна інформація є частково застарілою, по-друге, обсяг словника майже у два рази менший за "Тональний словник української мови" О. Толочко (2023). Створення та актуалізація словників тональності важливі, оскільки останні містять унікальну лінгвістичну інформацію. Відповідно, для опрацювання та подальшого використання цієї інформації первинно необхідний саме словниковий підхід.

Мета дослідження — визначення основних можливостей словникового підходу для встановлення тональності, ідентифікація релевантних для тональності правил, практична перевірка цих правил, укладення тонального датасету.

Завдання дослідження:

- з’ясувати поняття сентимент-аналізу;
- окреслити основні підходи до визначення тональності;
- схарактеризувати словниковий підхід;
- визначити та обґрунтувати релевантні для аналізу тональності правила;
- описати використаний словник загальновживаних слів О. Толочко;
- удосконалити робочу програму-аналізатор тональності коротких текстів на базі словника;
- увести функціонал візуалізації тональних дерев;
- укласти тональний датасет;
- навчити модель на укладеному датасеті.

Об’єктом дослідження є тональність українськомовного тексту. Загалом текст може бути у формі відгуку, допису в соціальній мережі,

повідомлення в месенджері, статті чи будь-яких інших текстових даних, які містять думки чи настрої для аналізу. У цій роботі використовуємо корпус українськомовних відгуків формі коротких текстів обсягом 2-4 речення.

Предмет дослідження — словниковий підхід та підхід із застосуванням машинного навчання до визначення тональності, тонально-релевантні правила, тональні класифікатори.

Методи дослідження — аналіз наукової літератури, словника тональності О.Толочко й моделювання тонального аналізатора з вбудованим функціоналом візуалізації тональної структури речення, побудова моделі машинного навчання.

Новизна роботи полягає в тому, що застосовано синтаксично орієнтований підхід до аналізу тональності українського тексту, коли за "наріжний камінь" береться корінь-присудок у дереві залежностей, а заперечення й інтенсифікатори локалізуються та враховуються завдяки їхньому синтаксичному зв’язку з головним дієсловом.Також новим є підхід, що передбачає інтеграцію словникового ресурсу О. Толочко з нейромережею на базі TensorFlow, який дав змогу поєднати переваги лексикону та машинного навчання у визначенні тональності. Уперше описано емоджі як маркери емоційного забарвлення. Створено й оприлюднено новий гібридний українськомовний датасет із детальною лематизацією, фільтрацією та статистичним аналізом, а також модель, навчену на цьому датасеті.

Практичне значення роботи — розробка може слугувати ефективним інструментом бізнес-моніторингу, а також може бути використана як відправна точка для подальших наукових досліджень.

Структура роботи. Робота складається зі вступу, двох розділів (16 підрозділів) — теоретичних аспектів сентимент-аналізу в цілому та аналізу тональності зокрема, практичної реалізації аналізу тональності на основі словника та висновків, списку використаних джерел та додатків. Розділ 1

описує базові поняття та ключові компоненти дослідження: від визначення сентимент-аналізу та аналізу полярної тональності тексту, огляду нетекстових маркерів емоцій — емоджі, до ролі синтаксису й побудови "тональних дерев" залежностей. Далі здійснюється огляд сучасних машинно-навчальних підходів для розв’язання цього завдання, зокрема теоретичні засади нейромережевої моделі на платформі *TensorFlow*, і обґрунтовується вибір відгуків як оптимального джерела даних для тренування й тестування системи. Розділ 2 описує весь життєвий цикл розробки від інструментальної бази до готового продукту: детальний аналіз і структуризацію лексичного ресурсу О. Толочко, реалізацію власного аналізатора коротких текстів із візуальним відображенням результатів, збір вихідних даних через вебскрапінг із наступною фільтрацією й анонімізацією, етапи підготовки даних до моделювання, безпосереднє тренування нейромережі, доповнення корпусу автоматично отриманими тональними оцінками та опис основних статистичних характеристик сформованого датасету. Розділ завершується обговоренням можливих практичних застосувань створеного аналізатора, тренованої моделі й самої бази даних. В Додатках наведене покликання на репозиторій з описаним у роботі практичним доробком.

РОЗДІЛ 1. ТЕОРЕТИЧНІ АСПЕКТИ СЕНТИМЕНТ-АНАЛІЗУ В ЦІЛОМУ ТА АНАЛІЗУ ТОНАЛЬНОСТІ ЗОКРЕМА

У цьому розділі представлена структура сентимент-аналізу як глобальної задачі, а також огляд підходів до її вирішення на основі опрацьованої наукової літератури. Основну увагу приділено специфіці словникового підходу, його перевагам і недолікам.

1.1 Поняття сентимент-аналізу

Сентимент-аналіз і майнінг думок часто вживають як взаємозамінні терміни, хоча деякі дослідники все ж вбачають невелику різницю між ними, зазначаючи, що сентимент-аналіз стосується більшою мірою емоційних реакцій, а майнінг думок — реакцій мисленнєвих. Однак безсумнівним є той факт, що обидва поняття пов’язані між собою — через те, що реакції мисленнєві та емоційні є спряженими. У цій роботі вищезазначені терміни вживаються як синоніми (Ligthart, Catal, Tekinerdogan, 2021).

Перші академічні дослідження, що охоплювали "вимірювання" громадської думки, були проведені під час та після Другої світової війни; мотивація цих досліджень мала виражений політичний характер. Новий спалах інтересу до сентимент-аналізу стався лише в середині 2000-х років, концентруючись на оглядах продуктів, доступних в Інтернеті. З того часу сентимент-аналіз суттєво виріс поняттєво, охоплюючи досить багато різних завдань, підходів та видів власне аналізу.

Сфера досліджень зазнавала змістових трансформацій протягом багатьох років. До появи величезної кількості тексту та думок в інтернеті дослідження в основному покладалися на методи опитування та були зосереджені на громадських чи експертних думках, а не на думках користувачів чи клієнтів. Одна з перших статей у цій галузі була

опублікована в 1940 році під назвою "Техніка викреслення як метод аналізу громадської думки" (Mäntylä, Graziotin, Kuutila, 2018).

До 2000 року було опубліковано лише 37 статей, а вже до 2005 року їх кількість становила 101. До кінця 2015 року кількість статей досягла значення 5699. Саме можливість автоматично збирати й аналізувати великі масиви думок за допомогою інструментів аналізу тексту зробила сентимент-аналіз настільки актуальною темою дослідження (Mäntylä, Graziotin, Kuutila, 2018).

Активно дослідженням аналізу тональності займалися та продовжують займатися й українські науковці. Від аналізу мовних засобів оцінності в сучасній українській мові через поєднання системно-структурного й семантико-функціонального підходів (Онищенко, 2005) та досліджень лінгвістичних засад створення системи автоматичного аналізу тональності публіцистичних та інформаційних повідомлень українськомовних джерел (Дарчук, 2019) до розробки окремого модулю автоматичної системи статистичної параметризації українськомовних текстів "TextAttributor 1.0", спрямованого на лінгвістичну експертизу токсичності тексту (Дарчук, Зубань, Робейко, Цигвінцева, 2024) — питанням аналізу тональності присвячений цілий спектр наукових робіт.

Планування аналізу тональності як практичного завдання ілюструє рис. 1, де наведено структуру майнінгу думок із блоків завдань, підходів та рівнів аналізу. Ця структура є спрямованою: формулювання задачі аналізу тональності здійснюється шляхом поетапного проходження зазначених блоків.

Крім того, майнінг думок можна розглядати як трирівневу структуру (синтаксичний, семантичний та прагматичний рівні), що охоплює п’ятнадцять підзадач опрацювання природної мови (Ligthart, Catal, Tekinerdogan, 2021). Синтаксичний рівень передбачає нормалізацію мікротексту, сегментацію на рівні речень (тобто визначення початку та

кінця речення), розмічування частин мови, розбиття тексту на синтаксично співвіднесені частини слів та лематизацію (Project DeepDive, 2017). Семантичний рівень передбачає зняття лексичної неоднозначності, видобування концептів, розпізнавання іменованих сутностей, розпізнавання анафори (процес інтерпретації зв'язку між анафорою, тобто повторюваним посиланням) і її антецедентом, тобто попередньою згадкою об'єкта) та визначення суб'єктивності. Прагматичний рівень зосереджений на розпізнаванні особистості автора, виявленні сарказму, розпізнаванні метафор, видобуванні аспектів і виявленні полярності.

Підходи до сентимент-аналізу можна розділити на три основні типи: підходи, засновані на експертних знаннях (на лексиконі), статистичні підходи (такі як машинне навчання та глибоке навчання) і гібридні підходи, які поєднують два вищезазначених (Taboada, 2016).

Розробляючи модель сентимент-аналізу, використовують різні методи попереднього опрацювання, методи вибору специфічних ознак чи певних правил (це залежить від обраного підходу). Попереднє опрацювання передбачає перетворення тексту в нормалізовані лексеми, що може включати видалення стоп-слів і застосування стемінгу або лематизації; не варто забувати й про приведення всіх слів до одного регістру (Taboada, 2016).

Вибір ознак стосується визначення вимірюваних властивостей, що слугуватимуть вхідними даними. Вибір (та створення) правил відрізнятиметься залежно від лексикону та кінцевої мети.

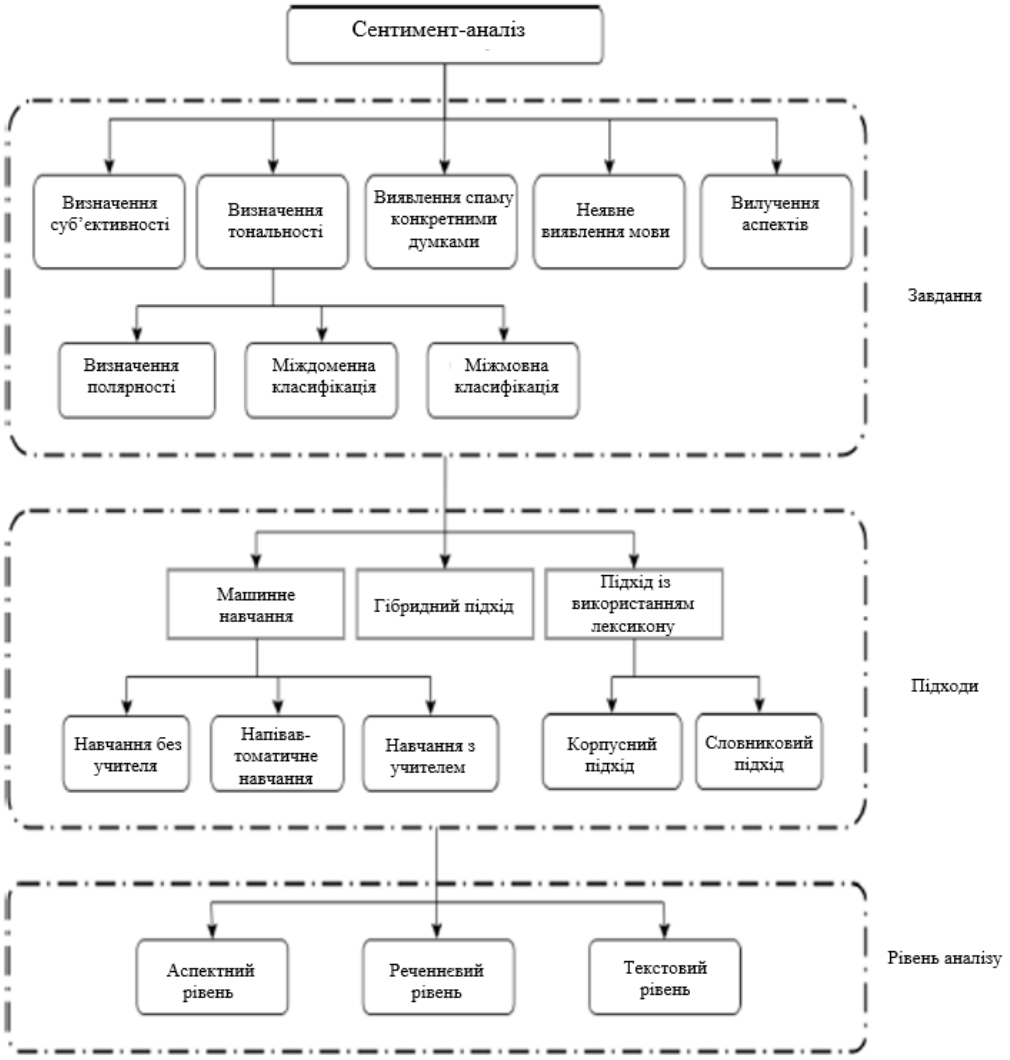

Рис. 1 Огляд концепції сентимент-аналізу (Taboada, 2016)

Отже, сентимент-аналіз і майнінг думок, хоча іноді розрізняються акцентами на емоціях чи мисленнєвих реакціях, у сучасній практиці розглядаються як синоніми й реалізуються через поетапну структуру завдань, підходів і рівнів обробки (синтаксичний, семантичний, прагматичний). Історично від перших воєнних опитувань до вибуху інтернет-оглядів у 2000-х та стрімкого зростання наукових публікацій, розвиток sentiment analysis підтримується трьома основними підходами (на основі лексикону, із застосуванням машинного навчання та гібридний) і вимагає глибинного попереднього опрацювання тексту й точного вибору

ознак. В українському контексті дослідження охоплюють як лінгвістичний опис оцінності, емоційного ставлення, вираженого в тексті, у різних стилях, так і практичну реалізацію автоматизованих систем для аналізу тональності й токсичності текстів.

1.2 Аналіз тональності як підзавдання сентимент-аналізу

Одним з найбільш відомих і досліджуваних завдань сентимент-аналізу є аналіз тональності тексту (також називають аналізом полярної тональності, аналізом полярності) (Mäntylä, Graziotin, Kuutila, 2018). Визначення полярності є лише підзавданням сентимент-аналізу, але часто ці терміни вживаються як синоніми, що є дещо застарілим підходом.

Традиційно полярність класифікується як позитивна або негативна. Деякі дослідження включають третій клас, який називається нейтральним (Mondher, Tomoaki, 2016). Також дослідження може передбачати багатосмугову шкалу:

- дуже позитивно;
- позитивно;
- нейтрально;
- негативно;
- дуже негативно (Скрипка, 2019, с. 10-11).

Зазвичай дослідження з урахуванням такої кількості категорій називається "дрібнозернистим" аналізом настроїв. Також можливе вираження настроїв конкретними почуттями: негативними — гнівом, смутком або позитивними — щастям, захопленням тощо (Скрипка, 2019, с. 10-11).

Міждоменна та міжмовна класифікація є підзавданнями класифікації полярності, які спрямовані на передачу знань із домену джерела, багатого на дані, до цільового домену, де дані та розмітка обмежені. Міждоменний

аналіз визначає тональність цільового домену за допомогою моделі, (частково) навченої на більш насиченому даними вихідному домені. Популярним методом є виділення інваріантних ознак домену, розподіл яких у вихідному домені близький до розподілу цільового домену. Модель може бути розширена інформацією, що стосується цільової області. Міжмовний аналіз також реалізується подібним чином — шляхом навчання моделі на наборі даних вихідної мови та її тестування іншою мовою, дані якої обмежені. Тестування може відбуватися, наприклад, шляхом перекладу цільової мови на вихідну мову перед обробкою (Xhymshiti, 2020).

Однією з основних проблем аналізу тональності є те, що слово може мати різну полярність у різних контекстах та сферах уживання. Наприклад, слово *жахи* в рецензії на фільм не означає, що рецензія загалом щодо фільму є негативною — оскільки це, з високою ймовірністю, просто назва жанру. З іншого боку, якщо слово *жах* трапляється в огляді на якийсь продукт, це, скоріш за все, негативний відгук (Xhymshiti, 2020).

Таким чином, підхід на основі лексикону повинен мати реалізацію, яка розрізняє слова на основі контексту їх використання. З іншого боку, система машинного навчання повинна мати велику кількість маркованих даних у цьому конкретному домені, щоб давати результат високого рівня. Якщо система машинного навчання вже навчена в певному домені, можна використовувати його для аналізу тональності за межами цього навчального домену, наприклад, у доменах, де бракує великої кількості або "хорошої якості" маркованих даних. Система не буде давати визначних результатів за межами навчальної області, але при цьому ці результати не можна назвати поганими порівняно з системою на основі лексикону (Xhymshiti, 2020). Ще з іншого боку, якщо систему машинного навчання потрібно навчити з нуля, і при цьому відсутня достатня кількість маркованих даних для домену, у якому потрібно реалізувати аналіз

тональності, найкращим вибором буде підхід з лексиконом, адже це швидше та простіше.

На рис. 2 схематично представлені етапи та елементи обох підходів у порівняльному аспекті. Бачимо, що сентимент-аналіз на основі використання лексикону відбувається швидше.

Підходи, засновані на використанні лексикону, спираються на припущення, що семантична орієнтація тексту строго пов’язана з полярністю слів і фраз, які в ньому зустрічаються. Це стосується семантично навантажених слів, а саме прикметників, прислівників, іменників і дієслів, а також фраз і речень, які їх містять (Catelli, Pelosi, Esposito, 2022).

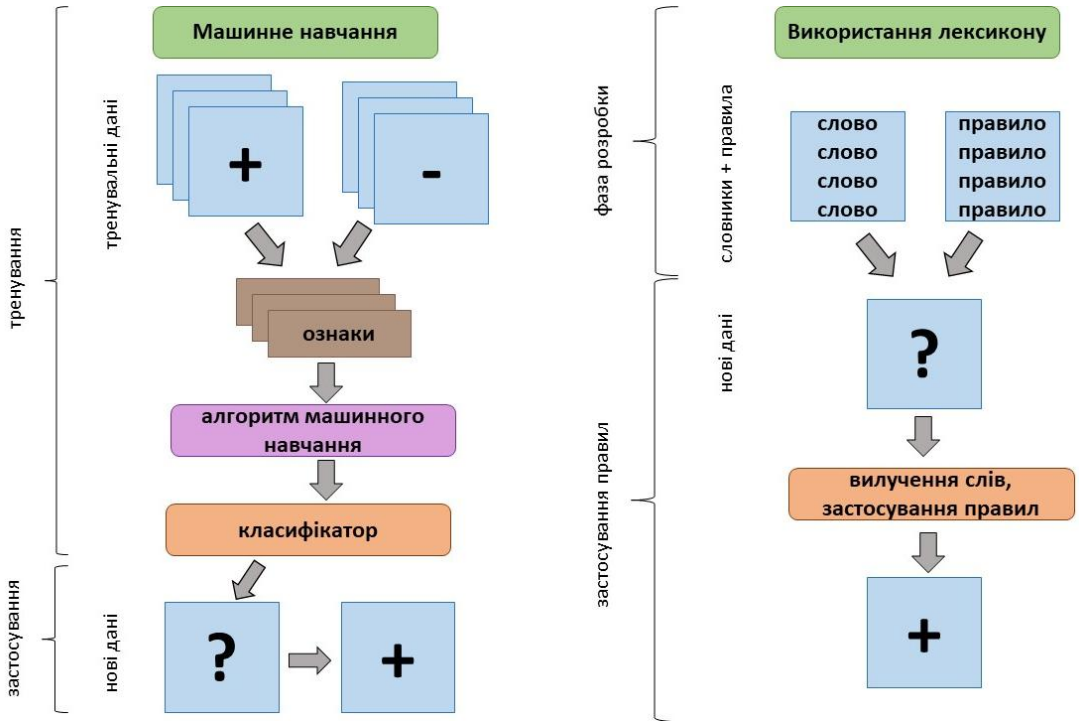

Рис. 2 Машинне навчання та підхід з використанням лексикону в сентимент-аналізі (Taboada, 2016)

Хоча лексикони, створені вручну, очевидно точніші, ніж ті, що створені автоматично, особливо в завданнях міждоменного аналізу настроїв, анотація вручну є недешевою з погляду людських ресурсів і часу (Bloom, 2011). Це стало причиною популярності досліджень автоматичного створення лексиконів полярної тональності, які виконують аналіз за допомогою морфологічних методів, використовуючи семантичні зв'язки тезаурусів і алгоритми спільного входження у великі корпуси.

Автоматично створені лексикони зазвичай менш стабільні, але об'ємніші за ті, що створені вручну. Та обсяг не дорівнює якості — великі лексикони зазвичай містять мало детальної інформації. Окрім того, велика кількість записів може означати більші обсяги шуму (Catelli, Pelosi, Esposito, 2022).

Існує два основні підходи до аналізу тональності на основі лексикону: на основі словника та на основі корпусу. Підхід з використанням словника починається з укладання невеликої вибірки слів, а далі ця вибірка поступово розширюється синонімами та антонімами з актуальних словників. У більшості випадків підхід, заснований на словнику, найкраще працює для неспецифічних цілей. Лексикони на основі корпусу можна адаптувати до конкретних доменів. Спочатку укладається список тонально забарвлених слів загального вжитку, а потім на основі цього списку розпізнаються інші слова з корпусу домену за допомогою словесних шаблонів, що збігаються (Ligthart, Catal, Tekinerdogan, 2021).

Існують також різноманітні гібридні підходи. Деякі з них орієнтовані на розширення моделей машинного навчання за допомогою знань, що ґрунтуються на лексиконі (Behera, 2016). Мета полягає в тому, щоб об'єднати обидва методи для отримання оптимальних результатів з використанням ефективної комбінації функцій як лексикону, так і методів, заснованих на машинному навчанні. У такий спосіб можна подолати недоліки та обмеження обох підходів.

Рис. 3 Загальний алгоритм аналізу тональності на основі лексикону (Sadia, Khan, Bashir, 2018)

На рис. 3 представлений загальний алгоритм аналізу тональності на основі лексикону. Основними етапами є попереднє опрацювання даних, яке результує чистими нормалізованими токенами, і конекція з лексиконом цих очищених даних (Sadia, Khan, Bashir, 2018).

Отже, аналіз тональності зосереджений на визначенні полярності тексту — позитивної, негативної або нейтральної. Підзавданнями є міждоменна та міжмовна класифікація, які переносять знання з насичених" доменів у "бідні" за даними, використовуючи інваріантні ознаки або переклад перед обробкою. Головна складність полягає в контекстній мінливості полярності слів. Підходи на основі лексикону швидкі та ефективні в умовах обмежених анотованих даних, але потребують контекстного врахування, тоді як статистичні моделі машинного навчання дають вищу точність у "своєму" домені внаслідок великих обсягів маркованих даних. Гібридні підходи поєднують обидва методи для

компенсації їхніх недоліків, а загальний алгоритм аналізу на основі лексикону складається з етапів попередньої обробки тексту та зіставлення нормалізованих токенів із полярними словниками.

1.3 Емоджі та тональність

Важливою проблемою в аналізі тональності є так звані емоджі — невіддільна частина сучасної письмової комунікації. І хоча це спірне питання, більшість лінгвістів вважають, що емоджі не є окремою мовою, а лише складною системою піктограм, які розширюють наше спілкування завдяки нюансам та емоціям. Однак аналіз емоційної ролі емоджі є дуже складним завданням. Це пов’язано з їхньою залежністю від різних факторів, таких як текстовий контекст, культурна перспектива, особистісні риси співрозмовника, стосунки між співрозмовниками або функціональні особливості платформи.

Залежно від контексту, емоджі може відігравати різні смислові ролі. Вони можуть виступати в ролі акценту, індикатора, пом’якшувача, реверсу або тригера негативного чи позитивного настрою в тексті. Крім того, емоджі можуть мати нейтральний ефект (тобто не впливати) на тональність тексту.

Аналіз тональності використовується для виявлення думок, емоцій в тексті. Відповідно, визначають емоджі як форму розвиненої пунктуації (спосіб кодування невербальних просодичних сигналів у системах письма), яка доповнює письмову мову, щоб полегшити автору артикуляцію своїх емоцій у текстовій комунікації. Крім того, розглядають використання емоджі як аналогічно закодованих символів, чутливих до зв’язку "відправник-одержувач" і повністю інтегрованих із супровідними словами (тобто, видимих актів значення) (Shatha, Hakami, 2022).

Ураховуючи все вищезазначене, за умови текстів, насичених емоджі (наприклад, дописів у соціальних мережах), рекомендовано створювати додатково словник тонально значущих емоджі, та врахувати їх при обчисленні загальної тональності. Тонально значущими емоджі вважаємо такі, що однозначно та незалежно транслюють позитивну/негативну тональність: це конкретні емоції (обличчя + сльози/посмішка тощо) та серця різних кольорів. Вибір обґрунтований тим, що абсолютна більшість інших емоджі мають нейтральну тональність, яка приймає певне забарвлення вже в контексті самого повідомлення, вторинно закріплюючи певну емоцію. В цій роботі було прийняте рішення відмовитись від введення цього функціоналу через практичну відсутність емоджі в відгуках.

Отже, емоджі зазвичай розглядають як розширену пунктуацію, що передає емоційні нюанси; їхнє значення залежить від контексту, культури та стосунків між співрозмовниками. Для текстів із великою кількістю емоджі доцільно мати окремий словник тонально значущих символів (позитивних чи негативних), але через практичну відсутність емоджі в аналізованих відгуках в цій роботі від цієї функції відмовилися.

1.4 Синтаксис та визначення тональності тексту. Тональні дерева

Підхід до аналізу тональності на основі правил дозволяє глибоко проаналізувати зміст рецензій, тобто можна знайти не лише загальну тональність рецензії, але й тональність окремих речень та підрядних частин. Це, своєю чергою, надає достатньо інформації, щоб виокремити, наприклад, позитивні та негативні моменти про особу чи подію.

Аналіз тональності зазвичай розглядають як завдання, орієнтоване на конкретну галузь. Дуже складно створити систему аналізу тональності, яка

б добре працювала для будь-якої теми, оскільки певні слова та фрази можуть бути позитивними в одній галузі, але негативними в іншій. Наприклад, слово "маленький". Якщо говорити про ґаджети, це слово, швидше за все, матиме позитивне забарвлення, але якщо мається на увазі розмір порцій у ресторані, з великою ймовірністю це негативна тональність.

Крім того, навіть коли обрано галузь аналізу настроїв, будуть деякі слова та фрази, які самі по собі не можуть бути віднесені до певної категорії тональності та/або емоцій, але в контексті передають певний визначений настрій та/або емоцію (Дідун, 2019). Тут також необхідно додати, що дуже важливо врахувати таргетну аудиторію. У домені новин, наприклад, критично важливими є політичні преференції як автора текстів, так і читачів/слухачів, бо це, по суті, зовнішній контекст тональності. Наприклад, речення *Поранено п'ятеро солдатів* — якщо додати перед цим *Новини з бєлгороду*, то тональність для українського читача буде дуже позитивною; а *Десна* (прим. українська військова частина): *поранено п'ятеро солдатів* має очевидно негативну тональність. Так само очевидно, що для середньостатистичного споживача-росіянина ці приклади матимуть полярно протилежні тональності; а для когось з віддаленої географічно частини світу тональність обох варіантів буде ближче до нейтральної.

Тому ідея універсального тонального аналізатора є дещо утопічною, особливо якщо аналізуються тексти з соціально-політичним забарвленням. Можна було б створити систему з декількома тематичними модулями, і підключати їх за необхідності, щоб тональні акценти максимально відповідали таргетній аудиторії, або ж, як це наразі й практикується, створювати системи, підлаштовані під конкретні цілі та області застосування, як-от під відгуки на товари.

Українська мова має вільний порядок слів, тому, наприклад, для речення *Зовсім не впоралась з курсовою* важливо врахувати, що *не* та *зовсім*, як представники негації та інтенсифікації відповідно, стосуються дієслова

впоралась, інвертуючи та підсилюючи його тональність. І якщо з негацією це в більшості випадків не проблема, бо вона "поруч" з таргетним словом, інтенсифікатори можуть знаходитися на відстані.

За Н.П. Дарчук, при побудові автоматизованих систем аналізу тональності для української мови потрібно враховувати кілька взаємопов’язаних лінгвістичних аспектів. По-перше, на рівні лексики важливо виділити саме оцінну лексику: прикметники та прислівники із позитивною або негативною конотацією (наприклад, *чудовий, прекрасно, катастрофічно*). До того ж важливо фіксувати морфологічні маркери, які утворюють оцінні похідні слова — наприклад, суфікси *-еньк- чи -ющ-,* — а також ідіоматичні вирази з прихованим емоційним забарвленням, такі як *на сьомому небі* чи *до біса*. На синтаксичному рівні досліджуються трансформативні конструкції, де предикатна частина містить оцінне визначення (наприклад, *це було дуже цікаво*), а також присутність модальних слів і часток (*лише, ж, би*), які змінюють інтенсивність або напрямок оцінки. Окрему увагу варто приділити заперечним сполучникам *але, проте,* що можуть інвертувати полярність висловлювання. Семантичний рівень передбачає розв’язання проблеми лексичної неоднозначності: одне й те саме слово може мати різну полярність залежно від контексту. Вагомими факторами є також інтенсифікатори (*дуже, мало, ледве* тощо), модифікатори (частка *не*). Важливо розпізнавати омонімію й метафори, аби уникнути хибних інтерпретацій (*гарячий* у значенні *модний* проти *у вогні*). Прагматичний рівень фокусується на розпізнаванні авторської інтенції: виявленні сарказму та іронії (поєднання позитивних слів із негативним контекстом), експресивних засобів (вигуків, повторів — *ой-ой-ой!, дуже-дуже*) і врахуванні функціонального стилю (публіцистичний чи розмовний), що визначає ступінь суб’єктивності.

Щоб упевнитися в коректності всіх цих аналізів, необхідно передусім провести якісну лематизацію та розмітку частин мови — це дає змогу

зіставляти слова з полярними словниками. Грамотно організовані процеси сегментації та синтаксичного парсингу, які виділяють значущі фразові одиниці, підвищують точність визначення того, які саме сегменти тексту несуть оцінне забарвлення. Нарешті, ефективність системи значно зростає, якщо застосувати домен-специфічні лексикони, адаптовані до конкретної тематики (новини, відгуки, соцмережі тощо). Усі ці лінгвістичні рівні — від окремого слова до прагматичного контексту — формують багато-рівневу модель, яка здатна правдиво фіксувати й класифікувати тональність українського тексту (Дарчук, 2019).

В інших наукових працях також звертають увагу на різнорівневі засоби експлікації інтенсивності: активно оперують термінами інтенсифікатор, інтенсифікат. Так, С.Є. Родіонова, Т.М. Сидорова (цит. у Дідун, 2019) визначають інтенсифікатори як експліцитні засоби підсилення (напр., *дуже, надто, надзвичайно*). Універсальним інтенсифікатором є прислівник *дуже* та його синоніми — *надто* та *вельми*. Ці прислівники показують небайдужість мовця до предмета мовлення, дозволяють передати сильне враження, викликане предметом мовлення. До "індикаторів елемента інтенсивності" Т.В. Гриднєва (цит. у Дідун, 2019) зараховує лексеми: *дуже, сильно, украй, надзвичайно*.

На основі вищенаписаного було складено список інтенсифікаторів та введене правило множення: тональне значення, утворене на момент розбору множиться на значення інтенсифікатора в словнику.

Список інтенсифікаторів: *дуже, надто, надзвичайно, неймовірно, занадто, зовсім, вельми, украй, вкрай, сильно*.

Заперечення інвертують полярність пов’язаної лексеми (наприклад, *не хороший ≈ поганий)*. У складніших випадках заперечення вбудовується глибоко у фразу, змінюючи смисл всього речення.

Вставні конструкції можуть як деталізувати, так і заперечувати основну думку, впливаючи на тональність. Наприклад, у реченні *Хоча*

фільм і був смішним, загалом він нудний протиставлення підрядної та головної частин формує негативну оцінку.

Сполучники *але, проте, однак, ні…ні* часто змінюють полярність оцінки: перша частина речення може бути позитивною, а після *але* — негативною. Багаторівневий синтаксичний аналіз допомагає відстежувати такі контексти (Socher, 2013).

Урахування вищезазначених граматичних явищ потребує формального синтаксичного аналізу (парсингу) тексту, оскільки тільки деревоподібна структура речення дозволяє точно визначити зв’язки між словами (напр., область дії заперечення чи межі підрядного речення). Таким чином, теоретичне обґрунтування показує, що без урахування синтаксису системи аналізу тональності будуть часто припускатися помилок при роботі зі складними конструкціями (запереченнями, інтенсифікаторами тощо).

Приклад результату роботи створеної програми для речення *Я не дуже й радію*:

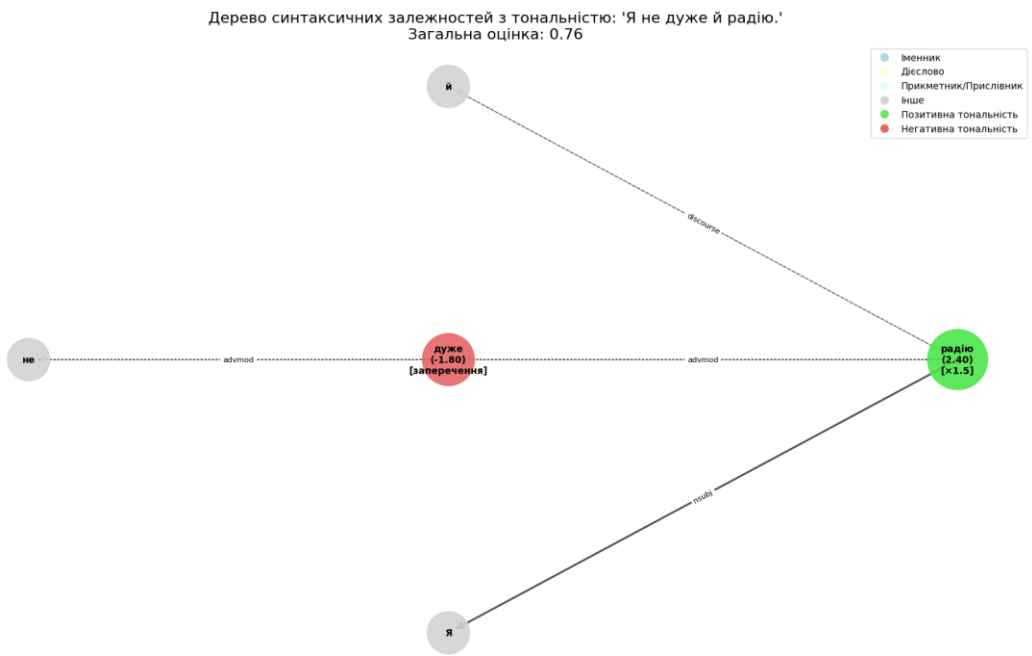

Рис. 4 Результат роботи синтаксичного парсера

Тональні дерева — це синтаксичні дерева речень, у яких кожна вершина анотована оцінкою тональності. На практиці тональне дерево формується за таким принципом: спочатку відбувається парсинг речення, а потім присвоєння кожному вузлу полярності (зазвичай за шкалою від "дуже негативно" до "дуже позитивно"). Наприклад, Stanford Sentiment Treebank (Socher, 2013) — перший корпус з такими анотаціями. Він складається з 11 855 речень з оглядів фільмів, розбитих на синтаксичні структури, кожну з яких незалежно оцінили троє експертів. У цьому датасеті використовували п’ятибальну схему оцінювання тональності (від "дуже негативно" до "дуже позитивно"), і результати анотування дозволили проаналізувати композиційні ефекти (на Рис. 5 показано, як модель віднесла кожен вузол до одного з п’яти класів тональності).

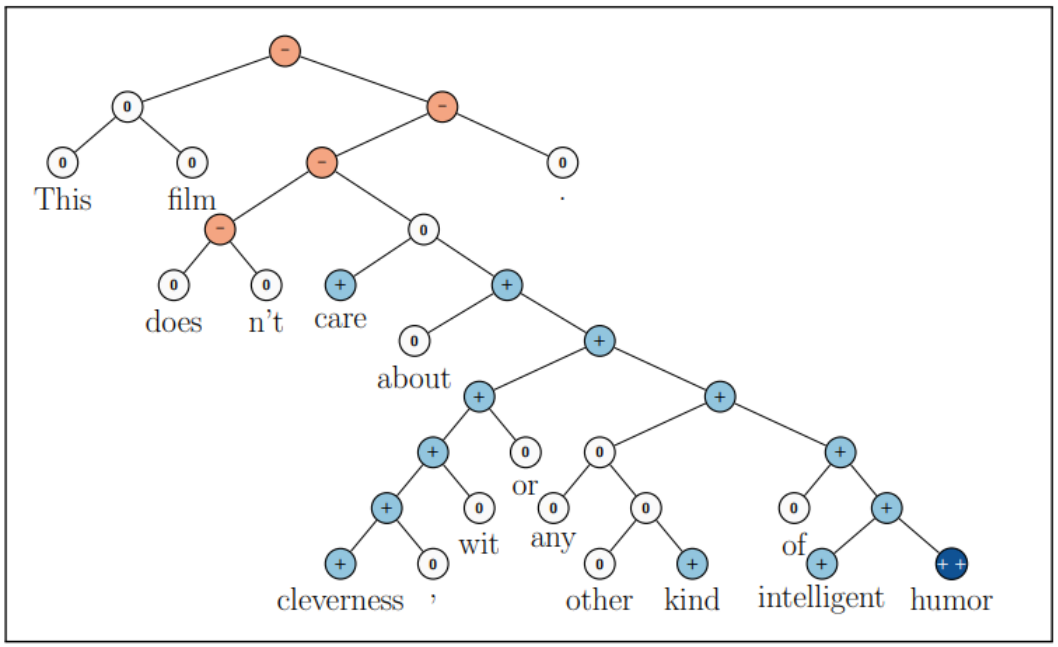

Рис. 5: Приклад рекурсивної нейронної тензорної мережі, яка точно прогнозує 5 класів настроїв, від дуже негативних до дуже позитивних (- -,

-, 0, +, + +), у кожному вузлі дерева розбору і фіксує заперечення та його обсяг у цьому реченні

Такі дерева використовують для навчання моделей і дослідження механізмів композиції сенсу. В одному з досліджень (Socher, 2013) на реченні *This movie was actually neither that funny, nor super witty* (Цей фільм насправді не був ані особливо смішним, ані надзвичайно дотепним) модель RNTN (рекурсивна нейронна тензорна мережа) відобразила, що окремі слова *funny* і *witty* є позитивними, але завдяки запереченню *neither… nor* вся фраза отримала негативну оцінку. На основі тональних дерев також можна виконувати складніші завдання (наприклад, таргет-орієнтований аналіз тональності), або будувати візуальні інтерпретації впливу окремих слів.

Для української мови врахування синтаксичної структури є надзвичайно важливим. Українська — флективна мова із відмінками, суфіксами та префіксами, а порядок слів у реченні є досить гнучким. Це означає, що різні комбінації можуть увиразнювати різні смисли або емоційні акценти. Тому для української — так само як і для інших мов з розвиненою морфологією — доцільно поєднувати аналіз тональності з повноцінним синтаксичним аналізом.

У синтаксичному та семантичному аналізі природної мови корінь речення — найчастіше присудок, — слугує наріжним каменем значення. Згідно з теорією граматики залежності, кожне інше слово в реченні прямо чи опосередковано пов’язане з цим коренем через мережу бінарних граматичних відношень (Nivre, 2005). В реалізованій українськими науковцями програмі автоматичного синтаксичного аналізу текстів корпусу української мови (Дарчук, 2013), структура речення визначається автоматично на рівні словосполучення. За автоматичним виокремленням словосполучень (тобто пов’язаних граматично та змістово груп слів) іде визначення типу синтаксичного зв’язку для кожного з цих словосполучень;

наступним кроком система приписує один з трьох типів зв’язку: підрядний, де одне слово залежить від іншого (наприклад, *читати звіт*); сурядний, де слова є рівноправними (наприклад, *ручка та олівець*); та предикативний, що позначає зв’язок між підметом та присудком (наприклад, *він прийшов*). Аналіз структури речення, описаний у статті, полягає в автоматичному розборі речення на словосполучення та встановленні типів зв’язків між словами в цих словосполученнях. Ця синтаксична структура є не просто формальною — вона відображає семантичну та інформаційну ієрархію в реченні.

Слова, які синтаксично ближчі до кореня, мають більший вплив на загальне значення та смисл речення. Це явище підтверджується емпіричними результатами аналізу тональності: вони демонструють, що тонально значущі лексеми, які знаходяться синтаксично ближче до кореня, впливають на класифікацію настрою сильніше, ніж лексеми в підрядних частинах або дієприслівникових зворотах (Socher та ін., 2013).

Цей висновок особливо цінний для синтаксично орієнтованих моделей аналізу настрою, які використовують дерева залежностей для оцінки структурної важливості слів. Наприклад, рекурсивні нейронні мережі (RNN) та деревоподібні мережі довгої короткочасної пам’яті (Tree-LSTM) працюють з ієрархічною структурою дерева, а не з лінійною послідовністю слів, надаючи пріоритет вузлам, структурно близьким до кореня (Tai, 2015). Такі моделі продемонстрували значні покращення порівняно з пласкими або суто послідовними моделями, саме тому, що вони враховують синтаксичну значущість слів.

Щобільше, обчислювальні метрики, такі як відстань залежності (кількість ребер між словом і коренем), були запропоновані як вагові коефіцієнти в системах, що базуються на правилах, і гібридних системах сентиментів (Qian, 2015). Слова з меншою відстанню залежності часто

отримують більшу вагу, що підвищує точність і запам’ятовуваність у завданнях класифікації.

Таким чином, врахування структурної близькості до синтаксичного кореня пропонує лінгвістично обґрунтований та емпірично підтверджений механізм зважування внеску сентименту в речення. Цей підхід підвищує надійність аналізу настрою, особливо для складних текстів, де лінійного аналізу недостатньо.

1.5 Огляд моделей машинного навчання для аналізу тональності

Ранні системи аналізу тональності працювали за принципом підрахунку слів на основі простих математичних формул, не враховуючи контекст та зв’язки між словами. Наївний Баєсів класифікатор (Naive Bayes, NB) трактує появу кожного слова як незалежну подію. Цей метод вирізняється високою швидкістю та ефективністю на невеликих наборах даних або для коротких текстів. Перевагою NB є його простота (потребує мало даних та має небагато параметрів), однак припущення про незалежність ознак може знижувати його продуктивність при аналізі довших текстів. Метод опорних векторів (Support Vector Machines, SVM) ефективний при роботі з високорозмірними розрідженими векторами ознак (наприклад, зваженими за TF-IDF — Term Frequency-Inverse Document Frequency, статистичним показником, що використовується для оцінки важливості кожного слова в документі щодо кількості його вживань у даному документі та в усій колекції текстів) і полягає в пошуку гіперплощини, що забезпечує максимальне розділення між класами. SVM часто демонструють одні з найкращих результатів серед традиційних підходів, особливо при використанні ядерних функцій (kernel functions) або

L2-регуляції (L2 регуляція допомагає знайти оптимальний баланс між точністю класифікації та простотою моделі) для налаштування моделі. Наприклад, низка досліджень підтверджує, що SVM часто перевершує NB та логістичну регресію за наявності достатнього обсягу даних для формування репрезентативних ознак. Логістична регресія (Logistic Regression, LR) є лінійною моделлю, що оцінює ймовірність належності до певного класу за допомогою сигмоїдної функції. Вона забезпечує задовільну інтерпретованість результатів та є обчислювально ефективною. На практиці продуктивність LR часто є порівнянною з SVM (обидві моделі будують лінійну межу прийняття рішень), при цьому LR може демонструвати більшу стабільність на великих наборах даних. Усі ці класичні моделі вимагають попереднього етапу конструювання ознак (feature engineering), таких як уніграми, біграми або ваги TF-IDF. Так, Панг та Лі (2002) продемонстрували, що на корпусі відгуків про кінофільми лінійні моделі, які використовують ознаки наявності уніграм та біграм, значно перевершують базові показники випадкового вгадування, а SVM досягає найвищої точності (близько 83%) на збалансованих вибірках. Вони також відзначили, що використання бінарних ознак (наявність слова) часто дає кращі результати, ніж використання сирих частот слів. Пізніше в рамках дослідження Ванга та Меннінга (2012) було проведено систематичну оцінку NB, SVM та їх гібридних варіантів на еталонних наборах даних для аналізу тональності. Автори виявили, що додавання біграмних ознак стабільно покращувало результати, і що для завдань аналізу тональності коротких фрагментів тексту NB працює краще, ніж SVM (тоді як для довших документів спостерігається зворотна тенденція) (Wang, Manning, 2012). Отже, NB є дуже швидким методом і може перевершувати SVM на коротких текстах, тоді як SVM та LR демонструють вищу ефективність для обробки довших документів (наприклад, розгорнутих відгуків) або при використанні великої кількості ознак. Ключовим недоліком усіх згаданих

моделей є ігнорування послідовного контексту: вони інтерпретують текст як невпорядкований "мішок ознак" і не здатні вловлювати порядок слів чи композиційну семантику без застосування складних методів інженерії ознак.

Моделі на основі глибоких нейронних мереж здатні автоматично вивчати репрезентативні ознаки з даних, що дозволяє їм вловлювати семантику слів та контекстуальні залежності. Проста нейронна мережа прямого поширення (Feedforward Neural Network, FNN), навчена на усереднених векторних представленнях слів або на вхідних даних типу "мішок слів", за відсутності врахування послідовної структури фактично відтворює функціональність логістичної регресії. Більш спеціалізованою архітектурою для обробки тексту є згорткова нейронна мережа (Convolutional Neural Network, CNN) (Kim, 2014). CNN використовують набір одновимірних згорткових фільтрів, які застосовуються до векторних представлень слів для виявлення локальних n-грамних патернів. За шаром згортки зазвичай слідує шар агрегування. Важливо зазначити, що навіть одношарова CNN, яка використовує попередньо навчені векторні представлення слів, продемонструвала високі результати в задачах аналізу тональності. Ю. Кім показав, що прості CNN "перевершили тогочасні найсучасніші результати (state-of-the-art) в чотирьох із семи завдань, включно з аналізом тональності" (Kim, 2014, с. 1). CNN ефективно виявляють ключові локальні ознаки (наприклад, слова або фрази, що виражають тональність) і навчаються швидше порівняно з рекурентними моделями. Їхній недолік полягає в обмеженому охопленні контексту: один фільтр обробляє лише фіксоване вікно слів, а для врахування ширшого контексту необхідно використовувати декілька згорткових шарів. Рекурентні нейронні мережі (Recurrent Neural Networks, RNN) обробляють слова послідовно, оновлюючи свій прихований стан після обробки кожного вхідного слова. Теоретично RNN здатні враховувати порядок слів та

довгострокові залежності в тексті, що є перевагою для аналізу тональності (наприклад, при опрацюванні заперечень або складних підрядних конструкцій). Класичні RNN мають проблему згасання градієнтів, що ускладнює навчання довгострокових залежностей. Для її вирішення були розроблені вдосконалені архітектури, такі як мережі довгої короткочасної пам’яті (Long Short-Term Memory, LSTM) та керовані рекурентні блоки (Gated Recurrent Unit, GRU). Архітектура LSTM використовує спеціальні вентильні механізми (gates) для контролю потоку та збереження довгострокової інформації, тоді як GRU пропонує спрощену структуру вентилів для досягнення подібної мети (Cho, 2014). Обидві архітектури, LSTM та GRU, широко застосовуються для класифікації тональності. Навчання моделей на основі LSTM/GRU відбувається повільніше, ніж CNN (через послідовний характер рекурентних обчислень), однак вони краще вловлюють послідовні залежності. Наприклад, у реченні *не дуже добре* коректна полярність визначається завдяки здатності моделі "пам’ятати" заперечення. На практиці класифікатори на основі RNN (зокрема, двоспрямовані LSTM) часто демонструють кращу продуктивність порівняно з FNN або CNN при аналізі довгих текстів, хоча й вимагають більших обсягів навчальних даних та обчислювальних ресурсів.

Архітектура трансформера (Transformer), представила механізм уваги, направленої на себе (self-attention), що дозволяє моделі одночасно враховувати глобальний контекст усієї послідовності (Vaswani et al, 2017). Попередньо навчені трансформерні моделі, зокрема BERT, розвивають цей підхід, вивчаючи глибокі двоспрямовані контекстуальні представлення слів на масштабних нерозмічених текстових корпусах (Devlin et al., 2019). Донавчання (fine-tuning) BERT та його похідних моделей для завдань аналізу тональності дозволяє досягати найвищих (state-of-the-art) результатів у багатьох сценаріях. Наприклад, сучасні дослідження в галузі аналізу тональності демонструють, що BERT значно перевершує базові

моделі CNN та RNN, часто забезпечуючи приріст точності на 5-10 відсоткових пунктів. Ключовими перевагами трансформерів є: (1) здатність моделювати контекст для кожного слова одночасно завдяки механізму уваги; (2) ефективне використання трансферного навчання (transfer learning) через попереднє навчання на великих даних, що зменшує потребу в значних обсягах специфічних для завдання розмічених даних; (3) підтримка токенізації на рівні частин слів (subword tokenization), що є особливо корисним для морфологічно багатих мов. До їхніх обмежень належать високі обчислювальні витрати та значний розмір моделей. Донавчання BERT вимагає використання графічних (GPU) або тензорних (TPU) процесорів і може призводити до перенавчання (overfitting) за недостатньої кількості специфічних для завдання даних.

Загалом, простіші моделі, такі як NB, LR та SVM, є доцільними для роботи з невеликими наборами даних або у випадках, коли пріоритетними є інтерпретованість результатів та швидкість навчання й інференсу. Вони швидко навчаються і часто демонструють хорошу продуктивність за умови ретельного конструювання ознак. Моделі CNN та RNN є обґрунтованим вибором за наявності помірної кількості розміченого тексту та потреби в моделюванні послідовності слів. Трансформерні архітектури (наприклад, BERT, mBERT, XLM-R) є оптимальним вибором за наявності великих обсягів даних або доступу до якісних попередньо навчених моделей. Вони, як правило, забезпечують найвищу точність, однак це пов’язано з більшими обчислювальними витратами.

У контексті малоресурсних мов, до яких ще донедавна належала українська (внаслідок значних зусиль наукової спільноти в плані створення корпусів, навчання моделей тощо, наразі українська вже підпадає під визначення середньоресурсної мови), обмеженість доступних даних та багата морфологія створюють специфічні виклики для розробки ефективних моделей. І хоча українська мова вже вважається

середньоресурсною, вищеописані виклики все ще актуальні. Ефективність традиційних моделей може бути обмежена через невелику кількість доступних розмічених корпусів. Сучасні дослідження пропонують дві взаємодоповнювальні стратегії для подолання цих викликів: (а) використання попередньо навчених мовних моделей, зокрема багатомовних, та (б) залучення додаткових лінгвістичних ресурсів. Наприклад, продемонстровано, що донавчання багатомовних моделей, таких як mBERT або XLM-RoBERTa, забезпечує хорошу якість перенесення знань (transfer learning) для завдань аналізу тональності текстів українською мовою. У дослідженні М. Притули (2024) українськомовні моделі RoBERTa та XLM-RoBERTa були донавчені на корпусі з приблизно 11 000 українських відгуків, досягнувши точності понад 91% (Prytula, 2024). При цьому навіть менші, дистильовані моделі (наприклад, DistilBERT) продемонстрували зіставну продуктивність за значно менших обчислювальних витрат. Методи класичного машинного навчання також залишаються актуальними: наприклад, за відсутності доступу до великих трансформерних моделей, можна навчити простий SVM на ознаках TF-IDF або використати векторні представлення слів fastText (які підтримують 157 мов, включно з українською). Дійсно, М. Романишин та ін. (2023) продемонстрували, що спеціалізовані векторні представлення fastText, навчені на великих українських текстових корпусах, суттєво перевершують готові багатомовні аналоги. Це підтверджує важливість адаптованих векторних представлень слів для покращення якості моделей обробки природної мови для української. Таким чином, для української та інших малоресурсних мов найбільш ефективними є підходи на основі трансформерів (особливо мономовні варіанти BERT, адаптовані для конкретної мови). Однак простіші моделі залишаються корисними та практичними за умов обмеженості даних або обчислювальних ресурсів.

Отже, для українського аналізу тональності прості моделі Naive Bayes, SVM і Logistic Regression залишаються швидкими й інтерпретованими рішеннями за обмежених даних, CNN ефективні для виявлення локальних патернів на середніх корпусах, RNN (LSTM/GRU) краще вловлюють довгострокові залежності, а трансформери (BERT-похідні) досягають найвищої точності внаслідок більших обчислювальних ресурсів та обсягів даних.

1.6 Модель TensorFlow: теоретичні засади

Для практичної реалізації програми була обрана модель TensorFlow. У цьому підрозділі коротко описана архітектура та принцип роботи цієї моделі.

Загалом архітектура TensorFlow така: TextVectorization → Embedding → GlobalAveragePooling1D → Dense (sigmoid). Шар TextVectorization (векторизація тексту) попередньо обробляє необроблений текст у послідовності цілих чисел. Він створює словник (кожному унікальному слову присвоюється число), а потім перетворює кожне вхідне речення на послідовність відповідних словам чисел. (TextVectorization у Keras може включати опції, такі як переведення в нижній регістр, видалення пунктуації або генерація n-грам.) Тут не відбувається навчання; він просто токенізує та індексує текст. Далі йде шар ембедингу: для вхідної послідовності довжиною n, кожен індекс токена x_i (де $0 \leq x\_i < V$) відображається на навчений вектор $e_{x_i} \in R^d$, де V — розмір словника, а d — розмірність векторного представлення. Шар ембедингу містить вагову матрицю $W_{emb} \in R^{V \times d}$. Коли подається вхідний індекс x_i, виходом є x_i-й рядок матриці W_{emb}. Для цілої послідовності [x_1, x_2, …, x_n] шар ембедингу видає матрицю: $E = [e_{x_1}; e_{x_2}; \vdots; e_{x_n}] \in R^{n \times d}$. Під час навчання вектори представлень e вивчаються за допомогою зворотного

поширення помилки: інтуїтивно, слова зі схожими контекстами тональності набуватимуть схожих векторів. Цей шар агрегує послідовність векторних представлень в один вектор фіксованої довжини. Маючи матрицю E ∈ R^{n×d} з попереднього кроку, глобальне усереднююче агрегування обчислює середнє значення кожного стовпця (кожної розмірності векторного представлення) по всій довжині послідовності. Формально, агрегований вихід p ∈ R^d має компоненти: p_j = (1/n) * Σ_{i=1}^n E_{i,j}, для j = 1, …, d. Тобто, p є середнім значенням n векторів слів. Ця операція повертає вихідний вектор фіксованої довжини для кожного прикладу шляхом усереднення за виміром послідовності. Інтуїтивно, це фіксує середній семантичний зміст тексту: позитивні слова додають позитивні компоненти, негативні віднімають тощо. Глобальне агрегування дозволяє просто обробляти вхідні дані змінної довжини. Нарешті, шар Dense з активацією Sigmoid (останній шар нейронної мережі, який приймає рішення про тональність тексту): Dense (щільний) шар з одним нейроном бере вхідний вектор p (агрегований результат з попереднього шару) і робить лінійну комбінацію: z = w · p + b, де: w — вектор ваг (навчається моделлю), p — наш агрегований вектор тексту, b — зсув (bias, також навчається). Число z може бути будь-яким (від -∞ до +∞), але нам потрібна ймовірність (від 0 до 1). Сигмоїда перетворює будь-яке число в діапазон (0,1): ŷ = σ(w · p + b). Чим більше z (додатне), тим ближче σ(z) до 1 (позитивний відгук); відповідно, чим менше z (від'ємне), тим ближче σ(z) до 0 (негативний відгук). Нульове z еквівалентне σ(z) = 0.5. Тоді ŷ, що позначає межу прийняття рішень, виглядає так: ŷ > 0.5 → "позитивний відгук", ŷ < 0.5 → "негативний відгук", ŷ = 0.5 → "невизначено".

Коли ми навчаємо модель розпізнавати тексти, наприклад позитивні чи негативні відгуки, нам потрібен спосіб оцінити, наскільки добре вона працює. Для цього використовують функцію втрат — бінарну перехресну ентропію. Бінарна, бо є тільки два варіанти (позитивна або негативна

тональність), а власне перехресна ентропія — спосіб порівняти правильну відповідь з тим, що передбачила модель.

Ось як це працює: у нас є правильна відповідь y, яка дорівнює 1 для позитивного тексту або 0 для негативного. Модель дає свою оцінку ŷ у вигляді ймовірності від 0 до 1. Формула втрат виглядає так: $L(y, \hat{y}) = -[y \times \log(\hat{y}) + (1 - y) \times \log(1 - \hat{y})]$. Якщо текст позитивний (y = 1), а модель дала високу ймовірність (ŷ близько до 1), то втрати маленькі, що добре. Але якщо текст позитивний, а модель дала низьку ймовірність (ŷ близько до 0), то втрати великі, що погано. І навпаки для негативних текстів. Ця функція "карає" модель, коли вона сильно помиляється, і "заохочує", коли вона права.

Коли ми знаємо, як оцінювати помилки, потрібно навчити модель їх виправляти. Для цього використовується алгоритм під назвою Adam. Процес працює так: модель робить передбачення, ми рахуємо помилки за допомогою функції втрат, визначаємо, в який бік змінити параметри моделі, щоб помилок стало менше, робимо маленький крок у цьому напрямку, і повторюємо процес знову і знову. Adam це "розумна" версія градієнтного спуску з кількома важливими перевагами. По-перше, він адаптивний, тобто Adam запам’ятовує, як змінювалися градієнти раніше, і підлаштовується під кожен параметр окремо. По-друге, він має імпульс: Adam "розганяється" в правильному напрямку і не зупиняється на маленьких нерівностях.

Спочатку алгоритм запам’ятовує середній градієнт за формулою $m_t = \beta_1 \times m_{t-1} + (1 - \beta_1) \times g_t$, де m_t це середній градієнт на поточному кроці, β_1 дорівнює 0.9 і показує, наскільки сильно ми "пам’ятаємо" минулі градієнти, а g_t це поточний градієнт. Потім алгоритм запам’ятовує розкид градієнтів за формулою $v_t = \beta_2 \times v_{t-1} + (1 - \beta_2) \times g_t^2$, де v_t це середній квадрат градієнтів, β_2 дорівнює 0.999 і означає ще сильнішу "пам’ять", а g_t^2 це квадрат поточного градієнта.

На початку навчання середні значення неточні, тому їх потрібно скоригувати за формулами $\hat{m}_t = m_t / (1 - \beta_1^t)$ та $\hat{v}_t = v_t / (1 - \beta_2^t)$. Нарешті, параметри оновлюються за формулою $\theta_{t+1} = \theta_t - \eta \times \hat{m}_t / (\sqrt{\hat{v}_t} + \varepsilon)$, де θ це параметр моделі (ваги), η це швидкість навчання (наскільки великі кроки робимо), а ε це маленьке число, щоб не ділити на нуль. Простими словами, Adam бере середній напрямок і ділить його на спектр градієнтів. Це означає, що якщо градієнти сильно стрибають, кроки стають меншими для обережності, а якщо градієнти стабільні, то кроки більші для швидшого навчання.

Під час навчання наша модель намагається знайти "лінію", а насправді — гіперплощину у багатовимірному просторі, яка найкраще розділяє позитивні та негативні тексти.

Алгоритм навчання працює таким чином: він аналізує помилки, зважаючи на те, які тексти класифіковані неправильно, потім зсуває межу, трохи повертаючи та переміщуючи лінію, щоб стало менше помилок, і повторює цей процес тисячі разів, поки лінія не стане оптимальною (Abadi et al, 2016).

Обрана модель має дуже просту будову, що є її головною перевагою для роботи з невеликими наборами даних. У нашому випадку це тільки два компоненти: перший це таблиця векторних представлень розміром $V \times d$, де V це кількість різних слів у нашому словнику, а d це розмірність кожного вектора, і другий це вихідний шар, який має лише $d+1$ параметрів. GlobalAveragePooling (глобальне усереднювальне об’єднання) значно зменшує складність даних, беручи всі слова в тексті та створюючи з них одне усереднене число для кожної характеристики, що запобігає ситуації, коли модель занадто сильно "запам’ятовує" короткі тексти замість того, щоб вчитися розпізнавати загальні патерни.

Коли у нас мало даних для навчання, простіші моделі працюють набагато краще за складні, тому що вони не намагаються запам’ятати кожну

дрібницю, а натомість вчаться загальним правилам. Наша архітектура по суті схожа на навчання неглибокої нейронної мережі, яка працює з усередненими векторами слів, подібно до популярних підходів як fastText. Глобальне агрегування, тобто процес усереднення всіх слів у тексті, дає нам представлення фіксованого розміру незалежно від того, чи відгук складається з п’яти слів, чи з п’ятдесяти. Модель розрахована на збалансовані класи, тобто приблизно однакову кількість позитивних і негативних прикладів, що дозволяє використовувати поріг 0.5 для прийняття рішень. Збалансовані дані сприяють стабільному навчанню, оскільки кожна порція даних під час тренування містить приблизно однакову кількість позитивних і негативних прикладів, тому функція втрат не має зміщення в один бік і модель вчиться об’єктивно.

Для аналізу тональності українською мовою ця архітектура має і свої переваги, і недоліки. З одного боку, вона надзвичайно легка і ресурсно невибаглива, що особливо корисно, позаяк великі моделі типу BERT неможливо навчити через брак потужності або пам’яті. Модель також вивчає векторні представлення слів безпосередньо з наших даних, що дозволяє їй потенційно вловлювати деякі особливості української морфології, особливо якщо в даних з’являється багато різних словоформ одного слова.

Однак багата морфологія української мови створює свої виклики. Українська мова має дуже розвинену систему відмінювання і дієвідмінювання, що означає, що одне слово може мати десятки різних форм. Це призводить до того, що розмір словника V може стати дуже великим, якщо ми використовуємо токенізацію на рівні цілих слів. Цю проблему можна пом’якшити, використовуючи такі підходи як n-грами символів або токени частин слів у шарі TextVectorization, що дозволяє моделі краще розуміти спільні корені та закінчення слів.

У будь-якому випадку, оскільки наша мережа неглибока, вона менш схильна до перенавчання на невеликому українському наборі даних, але водночас може недостатньо добре вивчити тонкі патерни мови. Емпіричні дослідження показують, що на обмежених українських корпусах донавчена одномовна модель RoBERTa значно перевершує такі прості базові моделі. Однак, оскільки у нас немає достатніх даних або обчислювальних ресурсів для трансформера, модель на основі усереднених векторних представлень є обґрунтованою відправною точкою.

Архітектура має дуже низьку місткість порівняно з глибокими RNN або Трансформерами, тому швидко навчається та добре узагальнює на обмежених даних, але водночас не може вловлювати порядок слів або складний синтаксис поза тим, що вже закодовано у векторних представленнях. Вона також передбачає адитивну композицію через процес усереднення, що може призвести до пропуску таких складних явищ як заперечення або сарказм, які потребують моделювання послідовностей слів та їхніх взаємодій.

Попри це, на збалансованих даних межа прийняття рішень є чіткою і зрозумілою, а процес навчання залишається стабільним без різких коливань.

Отже, обрана модель на базі TensorFlow поєднує простоту та ефективність: шар TextVectorization перетворює текст у послідовність індексів, Embedding навчає вектори слів, GlobalAveragePooling1D усереднює їх у фіксований вектор, а останній Dense-sigmoid дає ймовірність тональності. Використання бінарної перехресної ентропії та оптимізатора Adam забезпечує швидку та стійку конвергенцію навіть на невеликих українських корпусах. Така архітектура дозволяє економити ресурси й уникнути перенавчання.

1.7 Відгуки як оптимальний варіант для формування датасету

Аналіз тональності є неоціненним для таких застосувань, як рекомендаційні системи, аналіз відгуків клієнтів, маркетинг та моніторинг соціальних мереж. Підтверджено, що онлайн-огляди (Online Customer Reviews (OCR)) та рейтинги товарів стратегічно впливають на рішення споживачів щодо купівлі на платформах електронної комерції (Rachmiani, Nabila, Tribowo, 2024). Результати показують, що онлайн-відгуки мають значний вплив на довіру споживачів. Позитивні відгуки засвідчують якість продукції, зміцнюють довіру споживачів і впливають на рішення про покупку. На противагу цьому, негативні відгуки сильніше формують сприйняття ризику, що може перешкоджати намірам зробити покупку. Ці результати підтверджують теорію негативного упередження, яка припускає, що споживачі більше реагують на негативну, ніж на позитивну інформацію. Рейтинги продуктів слугують сигналом якості (сигналом довіри), надаючи "міру" надійності продукту, засновану на колективному досвіді реальних споживачів. Продукти з високими рейтингами, особливо якщо вони підкріплені достатньою кількістю відгуків, мають більшу довіру споживачів. Значна кореляція між рейтингом товару та довірою споживачів підкреслює важливість цього елементу як ключового показника в процесі прийняття рішення про покупку в середовищі онлайн-покупок (Rachmiani, Nabila, Tribowo, 2024). Відповідно, вибір відгуків як матеріалу для наповнення датасету для аналізу тональності є обґрунтованим: це короткі тексти, насичені тонально значущими словами та створені реальними споживачами.

Отже, аналіз тональності є ключовим інструментом для рекомендаційних систем, оцінки клієнтських відгуків, маркетингу та

моніторингу соцмереж, оскільки він дозволяє вимірювати вплив онлайн-оглядів на довіру споживачів і їхні рішення про покупку. Рейтинги товарів виступають сигналом якості, а їхня кореляція з довірою покупців підкреслює важливість відгуків як джерела даних: це короткі, інформативні тексти, насичені тонально значущими словами й створені реальними користувачами, що робить їх оптимальним матеріалом для побудови й оцінки моделі аналізу тональності.

Висновки до розділу 1

У Розділі 1 представлено всебічний теоретичний огляд сентимент-аналізу та аналізу тональності зокрема. Ретельно систематизовано історичний розвиток галузі, класифікацію підходів — на основі лексикону, за допомогою методів машинного навчання та гібридного. Також розглянуто багаторівневу структуру обробки тексту: синтаксичну, семантичну і прагматичну. Окрема увага приділена лінгвістичним особливостям української мови — вільному порядку слів, ролі інтенсифікаторів, негацій та контекстуальних зсувів полярності. Показано значення емоджі, тобто графічних символів-емоцій, які можуть як посилювати, так і змінювати тональність тексту залежно від контексту. Розглянуто ефективність класичних моделей машинного навчання — наївного баєсівського класифікатора (Naive Bayes), методу опорних векторів (Support Vector Machine, SVM) і логістичної регресії (Logistic Regression, LR). Також описано сучасні глибинні нейронні архітектури: згорткові нейронні мережі (Convolutional Neural Networks, CNN), рекурентні нейронні мережі (Recurrent Neural Networks, RNN), моделі довгої короткочасної пам’яті (Long Short-Term Memory, LSTM) та трансформери (Transformers) на кшталт BERT (Bidirectional Encoder Representations from Transformers).

РОЗДІЛ 2. ПРАКТИЧНА РЕАЛІЗАЦІЯ АНАЛІЗУ ТОНАЛЬНОСТІ НА ОСНОВІ СЛОВНИКА

У цьому розділі викладені детальний опис використаного словника загальновживаних слів та створеної програми, включно із залученими бібліотеками, а також детальний процес створення датасету та моделі машинного навчання.

2.1. Аналіз словника загальновживаних слів О. Толочко

Словник загальновживаних слів, на базі якого була розроблена програма аналізу тональності, був створений протягом 2020-2023 років. У його основі лежать два списки слів, частково розмічені студентами кафедри загального та прикладного мовознавства й слов'янської філології філологічного факультету Донецького національного університету імені Василя Стуса. Перший список мав обсяг 40 567 слововживань, а другий — 7 000 лексем. Також були використані три інші списки слів, відібрані та упорядковані О. Толочко під час роботи над словником. Окрім того, укладачка доповнила словник словами та словосполученнями, які вона помітила під час прослуховування та читання новин.

У першій колонці словника представлено слова та словосполучення, а колонки від другої до дев'ятої містять оцінки тональності для кожного слова або словосполучення.

Слово//словосполучення	Оцінки тональності								Середня оцінка тональності	Примітки
абат	1	1	1	0				0	0.60	
абатство	1	1	1	0				0	0.60	
абетка	0	1	1	0				1	0.60	

Рис. 6 Фрагмент словника О.Толочко

У десятій колонці подано інформацію про середню оцінку тональності, яка обчислюється автоматично за допомогою вбудованих функцій у Google Таблицях шляхом підрахунку середнього значення оцінок. Оцінки були виставлені різними студентами та експертами із загальною метою забезпечення наявності принаймні п’яти оцінок для кожного слова або словосполучення, включно з оцінкою укладачки словника. Максимальна кількість оцінок для кожного елемента — 8.

Оцінка	Значення оцінки
2	дуже позитивно
1	позитивно
0	нейтрально
-1	негативно
-2	дуже негативно

Рис. 7 Шкала визначення тональності

Чітка та зрозуміла структура словника суттєво сприяє ефективній обробці тональної інформації.

Отже, описаний словник з понад 47 000 лексем, оцінених мінімум п’ятьма експертами й структурований у табличні колонки з індивідуальними та середніми тональними оцінками, забезпечує чітку та ефективну обробку тональності.

2.2 Реалізація аналізатора тональності коротких текстів на базі словника; введення візуального елементу

Створена програма реалізує систему семантико-синтаксичного аналізу українських текстів з акцентом на автоматизовану оцінку їхньої тональності. Вона поєднує методи сучасного опрацювання природної мови, лексико-граматичний аналіз, а також візуалізацію синтаксичних дерев із

маркуванням полярності окремих слів. Метою є забезпечення гнучкого та детального інструменту для якісного аналізу емоційного забарвлення речень українською мовою.

На першому етапі програма ініціалізує необхідні мовні ресурси. Для морфологічного та синтаксичного аналізу тексту використовується модель uk_core_news_trf, яка є трансформерною українською моделлю бібліотеки spaCy. Вона забезпечує високу точність токенізації, визначення частин мови, залежностей і лем. Також під'єднано морфологічний аналізатор pymorphy3, який слугує для приведення слів до нормальної форми. Завантажується також список стоп-слів українською мовою, які не несуть суттєвого семантичного навантаження та підлягають виключенню з обробки.

Далі здійснюється імпорт словника тональності з електронної таблиці Excel, де кожному слову чи словосполученню відповідає числова оцінка емоційної полярності. Цей словник є основним джерелом лексичних оцінок у процесі аналізу. Крім того, визначено множини інтенсифікаторів (модифікаторів підсилення тональності) та заперечень, які відіграють роль семантичних модифікаторів контексту.

Функція preprocess_text виконує базову текстову обробку: токенізацію, лематизацію та очищення від стоп-слів. Однак основна аналітична логіка втілена у функції compute_sentiment_syntax, яка реалізує двоетапний підхід до оцінювання тональності з урахуванням синтаксичних залежностей.

У першій ітерації здійснюється ідентифікація заперечень та інтенсифікаторів. Для кожного заперечення визначається його зона впливу шляхом аналізу синтаксичного дерева: головного слова, його нащадків і наступних змістовних лексем. Аналогічно, інтенсифікатори прив'язуються до найближчих змістових токенів або головних дієслів, і їхній вплив масштабується через відповідні коефіцієнти підсилення. Результатом є

побудова двох словників — negation_scope та intensifier_scope, що локалізують вплив модифікаторів у тексті.

На другому етапі програма обчислює підсумкову тональність речення, послідовно обробляючи кожен токен. Для слів, наявних у словнику, базове значення модифікується з урахуванням заперечення (інверсія знаку) та інтенсифікації (множення на коефіцієнт). Отримане скориговане значення додається до загального показника тональності з урахуванням ваги — що ближче слово до кореня синтаксичного дерева, то більший його внесок (використовується обернена функція відстані до кореня). Усі дані зберігаються у структурі word_details, що дає змогу відстежувати вплив лінгвістичних модифікаторів на рівні окремих лексем.

Фінальний етап передбачає візуалізацію синтаксичного дерева функцією visualize_syntax_tree. Використовуючи бібліотеку NetworkX, програма будує орієнтований граф, де вузли відповідають словам, а ребра — синтаксичним залежностям. Вузли мають інформативні мітки, які містять як слово, так і його емоційний рейтинг, за потреби — відмітки про наявність заперечення чи інтенсифікації. Для розташування графу передбачено інтеграцію з Graphviz (через pygraphviz) для створення горизонтального дерева залежностей.

На зображенні, отриманому в результаті виконання поданої програми, представлено розгорнуте синтаксичне дерево речення з візуальним відображенням інформації про тональність слів, синтаксичні залежності, а також дію заперечень і підсилювачів емоційного забарвлення. Це дерево має форму спрямованого графа, у якому вузли відповідають окремим словам речення, а ребра — граматичним залежностям між ними, що визначені за допомогою трансформерної моделі "uk_core_news_trf" бібліотеки spaCy.

Кожен вузол дерева представляє окремий змістовний токен (тобто слово без пунктуації) і маркований текстовою міткою, яка містить слово,

числову оцінку його внеску в загальну тональність речення, а також — за потреби — пояснювальні позначки щодо граматичних або семантичних модифікаторів. Зокрема, якщо слово зазнало впливу заперечення, під його значенням у дужках зазначено "[заперечення]", що вказує на інверсію знаку емоційної оцінки. Якщо ж на слово впливає інтенсифікатор (наприклад, "дуже", "вкрай", "надзвичайно"), то додатково зазначається множник у форматі [×1.5], який означає ступінь посилення тональності. Таким чином, мітка вузла може містити, наприклад, таку інформацію: "прекрасно (2.50) [×2.5]", що означає: слово "прекрасно" має базову тональність +1.0, яку було підсилено інтенсифікатором у 2.5 раза до фінального значення +2.5.

Кожне ребро дерева поєднує головне слово (корінь) зі словом, що від нього залежить, і позначене типом синтаксичного відношення, наприклад "nsubj" (підмет), "obj" (додаток), "advmod" (обставина), "amod" (означення) тощо. Візуалізація побудована так, що напрямок ребер вказує від головного до залежного слова, демонструючи ієрархічну структуру речення. Такий підхід дозволяє не лише бачити загальну структуру фрази, але й простежити, як емоційне забарвлення поширюється в контексті синтаксичних відношень.

Колір і розмір елементів дерева можуть змінюватися відповідно до величини тональності. Наприклад, позитивні або негативні слова можуть мати більш насичені кольори, залежно від сили емоційного впливу. Окрім того, загальна оцінка тональності речення розраховується з урахуванням як значень окремих слів, так і їхньої синтаксичної ваги — що ближче слово до кореня дерева, то більший його внесок у фінальну оцінку. Це моделюється за допомогою спеціального коефіцієнта ваги, який обернено пропорційний відстані слова до кореневого вузла.

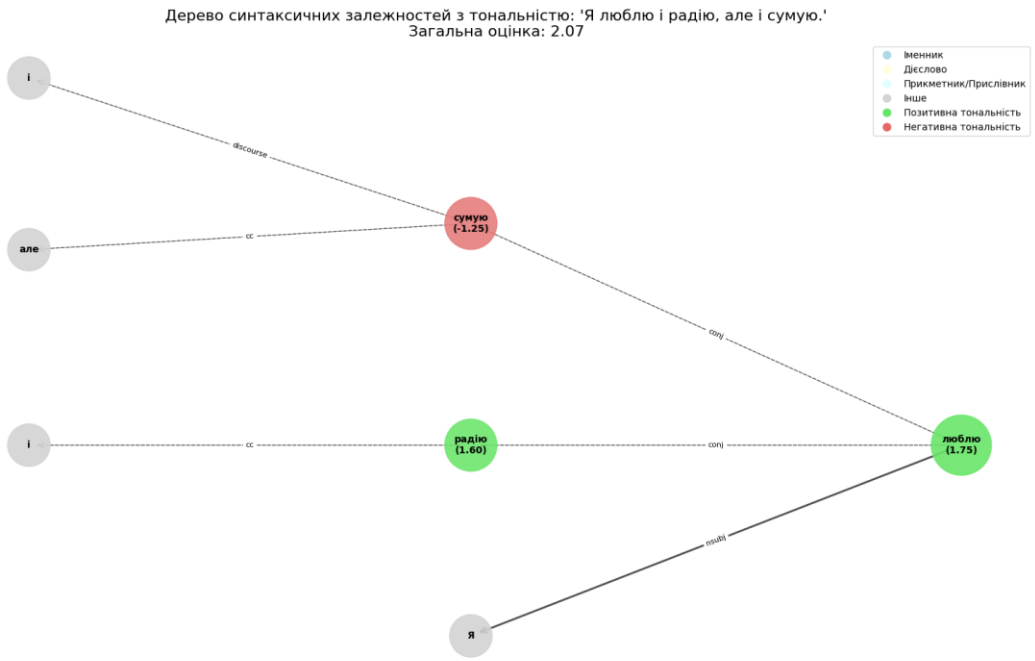

Рис. 8 Тональне дерево

Отже, результатом роботи програми є не лише числова оцінка сентименту речення, а й візуально інформативне синтаксичне дерево, яке слугує засобом пояснення й верифікації автоматичного аналізу — інтерпретоване через лінгвістичні категорії (частини мови, залежності, леми) й модифікаційні фактори (заперечення, інтенсифікація). Така візуалізація є надзвичайно корисною в дослідженнях природної мови, де важливе не лише машинне оцінювання, а і його прозорість і обґрунтованість.

2.3 Створення датасету: вебскрапінг

Для формування датасету був обраний метод вебскрапінгу з сайту vidhuk.ua. Сайт vidhuk.ua незалежною українською платформою для публікації відгуків про товари, послуги, компанії та організації. Його структура орієнтована на зручну навігацію, відкритий обмін досвідом між користувачами та забезпечення правдивості інформації через систему модерації.

На головній сторінці сайту представлено навігаційне меню з доступом до різних категорій, таких як *Інтернет-магазини, Медицина, Ліки, Краса, Послуги, "Транспорт, Техніка* тощо. Кожна категорія веде до списку компаній або продуктів, на які залишаються відгуки. Окремі сторінки компаній містять загальну інформацію, середній рейтинг, кількість переглядів, а також хронологічно впорядковані відгуки користувачів. Для аналізу були обрані декілька компаній, що містили найбільшу кількість відгуків. Серед них такі магазини, сервіси та надавачі послуг як: *Алло, Answear, Bodo, Busfor, Цитрус, Comfy, Ельдорадо, ЕпіцентрК, Eva, Кабанчик, Київстар, Водафон Україна, LaModa, MakeUp, Megogo, Нова Пошта, Окко, OLX, Розетка, Shafa, УкрПошта, УкрЗалізниця, Водафон, Сільпо, Віасат, Воля, Ювелірний дім "Сова", АвтоРіа, Містер Кет* та інші.

Кожен відгук має чітко визначену структуру. Він містить: заголовок — одне-два слова, що слугують узагальненням відгуку; власне відгук — розгорнутий опис досвіду з позитивними чи негативними моментами; автора; час публікації — дата і, зазвичай, точний час створення відгуку; оцінку у вигляді зірок — від 1 до 5, що відображає загальне враження користувача від взаємодії з компанією або продуктом.

Рис. 9 Приклад відгуку з сайту vidhuk.ua

Значущими для роботи є два елементи: власне відгук та рейтинг. Останній елемент по суті є оцінкою тональності відгуку (зрозуміло, що в першу чергу це оцінка сервісу, але пряма залежність тональності відгуку та

оцінкою користувачем сервісу є очевидною), що є цінною інформацією, яку не можна ігнорувати.

Створена програма реалізує автоматизований збір українськомовних відгуків зі сторінок інтернет-магазину vidhuk.ua з використанням бібліотеки Selenium для керування веббраузером. Основна мета програми полягає в парсингу контенту з динамічної вебсторінки, яка містить списки відгуків про певну компанію, із подальшим збереженням зібраної інформації у CSV-файли для аналізу.

На початку ініціалізується механізм логування, який фіксує ключові події виконання скрипта. Далі налаштовується режим роботи вебдрайвера Google Chrome, що дозволяє запускати браузер без графічного інтерфейсу, з метою оптимізації швидкості та зменшення навантаження на систему. Основна логіка скрипта реалізована у функції "main()", де шляхом ітеративного переходу по сторінках сайту за допомогою змінного параметра "page" в URL формується повна вибірка відгуків.

На кожній сторінці здійснюється пошук HTML-елементів, що містять тексти відгуків та відповідні їм оцінки у вигляді візуальних індикаторів (зірок), представлених як стилізовані елементи з атрибутом "style", який відображає ширину заповнення зірок у пікселях. Ці значення парсяться за допомогою регулярних виразів і перетворюються на числові оцінки, що приблизно відповідають 5-бальній шкалі (одна зірка — 13 пікселів ширини). Для забезпечення релевантності даних додатково здійснюється мовна фільтрація: з тексту вилучаються лише ті відгуки, що написані українською мовою, що досягається внаслідок пошуку унікальних українських літер і відсутності ознак російської мови.

Зібрані на кожній сторінці валідні відгуки з відповідними оцінками записуються у два CSV-файли, які оновлюються в режимі дозапису. Один файл містить відгуки виключно аналізованої на момент запуску скрипта програми, а другий, основний — усі відгуки усіх компаній. Програма

продовжує обхід сторінок доти, доки на черговій сторінці не буде виявлено нових відгуків, після чого завершить виконання. Протягом усього процесу здійснюється детальне логування — зокрема, виводяться повідомлення про успішні переходи по сторінках, кількість зчитаних відгуків, можливі невідповідності між кількістю текстів і зірок, а також виняткові ситуації (наприклад, тайм-аути або відсутність елементів). Завершується скрипт звільненням ресурсів браузера та виведенням підсумкової кількості зібраних україномовних відгуків.

Отже, за допомогою вебскрапінгу з платформи vidhuk.ua зібрано українськомовні відгуки з двома ключовими полями (текст і рейтинг), та забезпечено коректний парсинг, мовну фільтрацію та збереження результатів в CSV-файл із детальним логуванням усіх етапів.

2.4 Фільтрація та анонімізація записів у БД

Наступна програма виконує попереднє опрацювання, анонімізацію та статистичний аналіз текстових відгуків користувачів, імпортованих із CSV-файлу, з метою підготовки даних для подальшого аналізу або моделювання в межах завдання аналізу тональності. Зокрема, вона очищує текстові дані від чутливої інформації, зберігає оновлений набір даних, а також візуалізує розподіл відгуків за рейтингом.

На першому етапі відбувається завантаження необхідних бібліотек: pandas — для обробки табличних даних, spacy — для лінгвістичного аналізу, re — для регулярних виразів, Counter з модуля collections - для підрахунку частот, а також matplotlib.pyplot — для побудови графіків. Для опрацювання природної мови використовується модель uk_core_news_sm з модуля spaCy, яка спеціалізується на українській мові.

Програма зчитує CSV-файл із відгуками, автоматично визначаючи назви колонок, які відповідають за текст відгуку та оцінку у вигляді зірок . Якщо такі колонки не знайдено, виконання припиняється з відповідним повідомленням про помилку.

Основне опрацювання тексту здійснюється у функції clean_text, яка викликається до кожного відгуку. За допомогою вбудованої системи розпізнавання іменованих сутностей (Named Entity Recognition, NER) spaCy, усі виявлені сутності в тексті замінюються на токен <ENTITY> з метою знеособлення даних. Далі текст очищується за допомогою регулярних виразів: довгі числові послідовності (наприклад, номери замовлень або трек-номери) замінюються на маркери <NUM> або <ORDER_ID>, залежно від їхнього шаблону. Це дозволяє стандартизувати та уніфікувати дані, зберігаючи їх змістовну структуру, але виключаючи персональну або службову інформацію.

Результати очищення зберігаються у новій колонці cleaned_text, після чого DataFrame експортується до нового CSV-файлу у кодуванні utf-8-sig, сумісному з Microsoft Excel. Паралельно підраховується кількість відгуків за кожною оцінкою зірками за допомогою об’єкта Counter.

Завершальним етапом є побудова стовпчикової діаграми, яка відображає розподіл кількості відгуків за оцінками від 1 до 5 зірок. Графік оформлено у стилі академічної візуалізації: додано числові підписи над кожним стовпчиком, підписи осей, сітка для полегшення зчитування значень, а також забезпечено збереження графіка у форматі PNG.

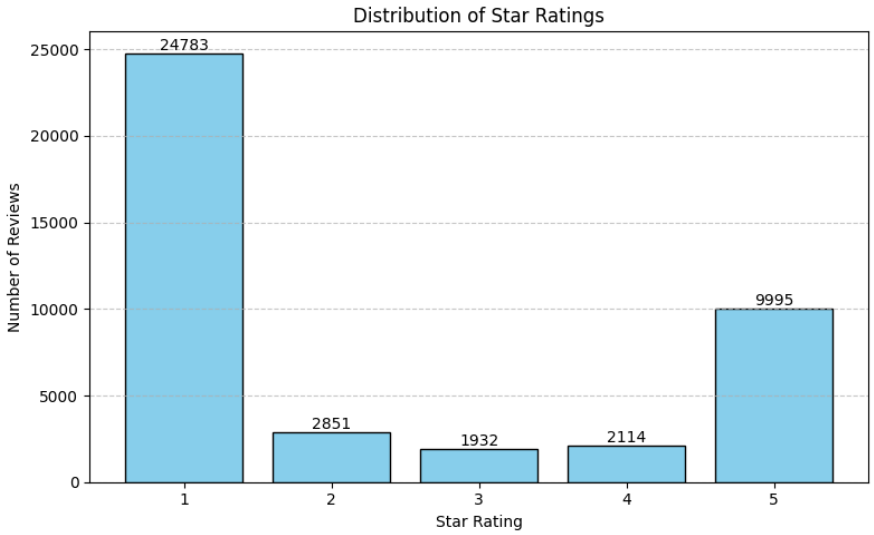

Рис. 10 Розподіл відгуків за рейтингом

Отже, створена програма імпортує відгуки з CSV, очищує й анонімізує їх за допомогою spaCy NER та регулярних виразів, зберігає результат у новий файл, підраховує розподіл рейтингів за допомогою Counter і виводить його у вигляді стовпчикової діаграми з підписами.

2.5 Підготовка датасету до запуску моделі

Кожен рядок бази даних тепер містить текст відгуку та числовий рейтинг. Щоб створити завдання бінарної класифікації тональності, ми розділили оцінки на два класи: позитивні (оцінки 4 або 5 зірок) та негативні (оцінки 1 або 2 зірки). Відповідно до попередніх робіт, будь-які нейтральні або середні оцінки, такі як рівно 3 зірки, були видалені або проігноровані, щоб уникнути неоднозначності. За цією схемою відгуки, позначені як позитивні, виражають схвальні думки про товар, тоді як негативні відгуки вказують на незадоволення.

Після маркування ми перевірили розподіл класів. Типово для наборів даних онлайн-відгуків, ми спостерігали зміщений розподіл, що значно переважав на користь негативних відгуків. Щоб уникнути навчання

виродженого класифікатора (такого, що завжди прогнозує мажоритарний клас), ми застосували зменшення вибірки (undersampling) мажоритарного класу для збалансування набору даних. Зокрема, випадковим чином зменшили вибірку негативного класу так, щоб кількість позитивних та негативних зразків стала приблизно однаковою. Зауважимо, що альтернативні підходи, такі як збільшення вибірки (oversampling) або синтетичне доповнення (augmentation), також могли б бути використані, але зменшення вибірки було простим, враховуючи розмір нашої вибірки, і допомогло уникнути генерації штучного тексту. Збалансування класів забезпечило те, що модель бачитиме однакову кількість позитивних та негативних прикладів під час навчання, зменшуючи зміщення в бік тональності, що домінує.

Підсумок етапів підготовки набору даних такий: фільтрація за оцінкою (збережено лише відгуки з 1—2 зірками (негативні) або 4—5 зірками (позитивні); відкинуто 3-зіркові (нейтральні) відгуки, що перетворює проблему на бінарну класифікацію тональності), присвоєння міток (присвоєно бінарні мітки (0 = негативний, 1 = позитивний) на основі оцінки), збалансування класів (зафіксовано значну більшість позитивних відгуків; випадковим чином зменшено вибірку негативного класу для вирівнювання кількості класів) та розділення даних (випадково перемішано очищені дані та розділено на навчальний та валідаційний набори, наприклад, 80% для навчання, 20% для валідації).

Цей підготовлений набір даних, хоча й значно менший за типові англійські корпуси, був достатнім для навчання та оцінки базового нейронного класифікатора. Ми наголошуємо, що ручне створення набору даних українських відгуків саме по собі є внеском, враховуючи дефіцит загальнодоступних розмічених українських даних для аналізу тональності. З усім тим, ми повинні визнати потенційні зміщення: зменшуючи вибірку, можемо втратити деякі природні негативні приклади, а виключаючи

нейтральні/змішані відгуки, модель не навчиться розпізнавати помірні тональності. Цей вибір спрощує завдання бінарної класифікації, але може обмежити реальну застосовність, де існують нейтральні думки.

Отже, шляхом відбору відгуків із 1–2 зірками як негативних і 4–5 зірок як позитивних (із виключенням нейтральних), бінарного маркування, балансування класів та розділення на навчальний і валідаційний набори, було підготовлено збалансований українськомовний корпус для надійного навчання базового нейронного класифікатора тональності.

2.6 Тренування моделі

Ця програма реалізує повний цикл опрацювання, навчання та оцінювання моделі бінарної класифікації текстів відгуків українською мовою з використанням бібліотек TensorFlow, keras, pandas і matplotlib. Вона призначена для автоматизованого аналізу сентименту — тобто визначення, чи є відгук позитивним чи негативним — на основі попередньо очищених текстових даних.

На початку здійснюється імпорт усіх необхідних модулів, після чого виконується налаштування кодування виводу для коректного відображення юнікод-символів, зокрема специфічних українських апострофів. Далі програма завантажує CSV-файл "vidhuk_reviews_cleaned_extra_5.csv", який містить очищені тексти відгуків (колонка "cleaned_text") та оцінки у вигляді зірок (колонка "stars"). Із цього набору даних вибираються лише ті відгуки, які мають 1, 2, 4 або 5 зірок, що відповідає завданням сентимент-аналізу з чітким поділом на негативні (1 і 2 зірки) та позитивні (4 і 5 зірок) класи. На основі цього поділу створюється нова бінарна цільова ознака ("label"), де 0 означає негативний, а 1 — позитивний відгук.

Щоб уникнути впливу дисбалансу класів на якість моделі, відбувається вирівнювання розміру класів через підвибірку (undersampling): з більшої групи випадковим чином вибирається така ж кількість прикладів, як і в меншій. Після цього об’єднаний і перемішаний датасет ділиться на навчальну та валідаційну вибірки у співвідношенні 80/20 зі збереженням пропорцій класів (стратифікація).

Текстові дані перетворюються у формати, зручні для тренування моделей у TensorFlow: вони пакетуються в батчі, валідаційні дані не перемішуються. Векторизація текстів відбувається за допомогою шару TextVectorization, який обмежує словниковий запас до 10 000 найчастотніших слів і зрізає або доповнює всі послідовності до довжини 250 токенів. Цей шар адаптується (навчається) лише на навчальних даних.

Архітектура моделі має послідовну структуру і складається з таких компонентів: шар векторизації, шар ембедингу, глобальне середнє згортання (GlobalAveragePooling1D), прихований повнозв’язний шар з 16 нейронами та ReLU-активацією, а також вихідний шар з одним нейроном та сигмоїдальною активацією. Така модель ідеально підходить для завдання бінарної класифікації. Вона компілюється з функцією втрат binary_crossentropy, оптимізатором Adam та метрикою accuracy (точність).

Після навчання протягом 10 епох модель оцінюється на валідаційних даних. Потім вона зберігається у форматі Keras ("saved_model_vidhuk.h5") для подальшого використання або донавчання.

Історія навчання моделі (об’єкт "history.history") зберігається у CSV-файл під назвою "training_history.csv". У цьому файлі фіксуються такі метрики для кожної епохи: "loss" — значення функції втрат на навчальному наборі;"accuracy" — точність класифікації на навчальному наборі;"val_loss" — функція втрат на валідаційному наборі; "val_accuracy" — точність класифікації на валідаційних даних.

accuracy	loss	val_accuracy	val_loss
0.649943	0.625048	0.717795193	0.582155
0.786208	0.45981	0.84021467	0.384349
0.849902	0.353658	0.873245239	0.326425
0.870393	0.30771	0.87943846	0.296711
0.883865	0.281184	0.848265886	0.330439
0.895272	0.258934	0.87654829	0.283328
0.90384	0.239429	0.89471513	0.276165
0.913337	0.22216	0.885218799	0.273264
0.914834	0.215203	0.90070188	0.256324
0.917931	0.205894	0.895334423	0.273679

Рис. 11 Історія навчання моделі

Таким чином, кожен рядок цього CSV-файлу відповідає одній епосі навчання і дозволяє надалі здійснювати кількісний та графічний аналіз перебігу тренування. На основі цієї історії будуються два графіки, які зберігаються у файл "training_metrics.png". Перший з них — "Training vs Validation Accuracy" — відображає динаміку точності моделі на тренувальному та валідаційному наборах упродовж епох, що дозволяє оцінити, чи не виникає перенавчання (overfitting). Другий графік — "Training vs Validation Loss" — показує зміну значення функції втрат на обох підмножинах, що також слугує індикатором якості узагальнення моделі. Обидва графіки супроводжуються підписами осей, заголовками, легендою та сіткою, що робить їх зручними для візуального аналізу.

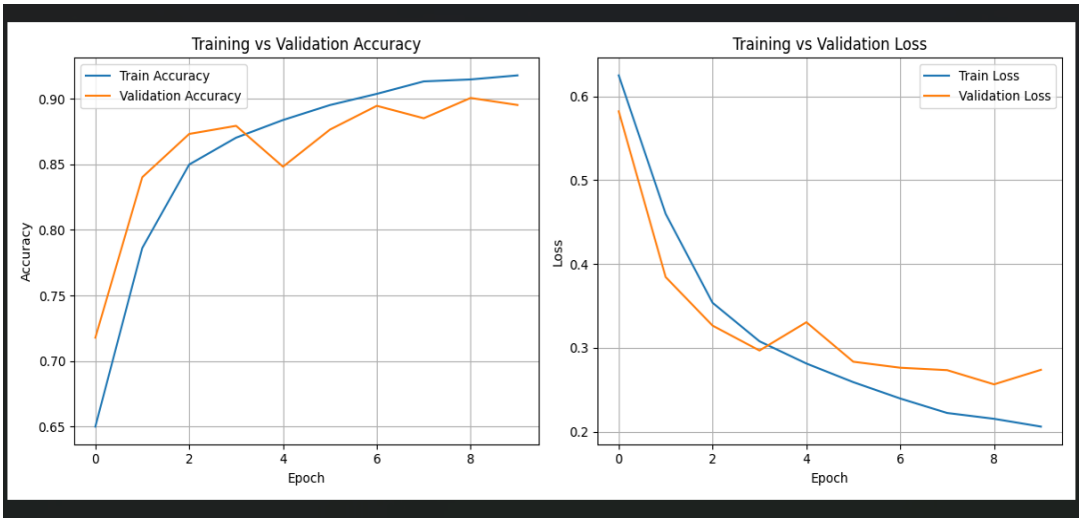

Рис. 12 Графіки: Точність на тренувальній і валідаційній вибірках та Втрати на тренувальній і валідаційній вибірках

У фінальній частині програма генерує матрицю невідповідностей (confusion matrix), яка ілюструє кількість правильних і неправильних передбачень для кожного класу. Модель отримує передбачення для всіх валідаційних прикладів і класифікує їх за порогом 0.5. Результати виводяться у вигляді графіка з підписами класів ("Negative", "Positive") і зберігаються у файл "confusion_matrix.png".

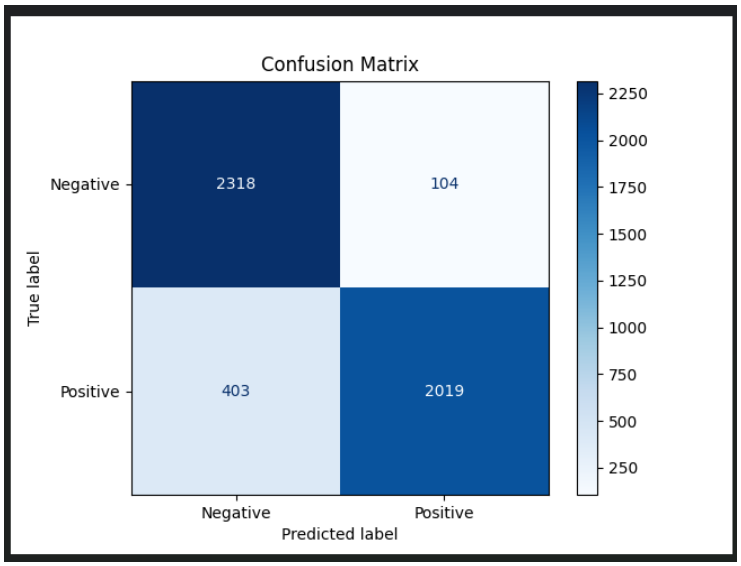

Рис. 13 Матриця невідповідностей

Таким чином, ця програма є повноцінним інструментом побудови моделі класифікації текстових відгуків: вона охоплює підготовку даних, тренування моделі, її оцінювання та збереження результатів у форматах, придатних як для візуального, так і для кількісного аналізу.

2.7 Доповнення датасету оцінкою програми-аналізатора тональності

Програма виконує повний цикл обробки україномовних текстових відгуків із CSV-файлу з метою визначення їхньої тональності на основі лінгвістичного та статистичного аналізу. Її головне призначення — автоматично зіставити оцінку у відгуку (текст), яка отримується вже описаним вище шляхом в рамках створеного аналізатора, із рейтингом (кількість зірок), виявляючи узгодженість або розбіжність між ними. Основні етапи реалізації та логіка програми такі:

Функція compute_sentiment_for_text застосовує аналіз до всього тексту: якщо текст складається з кількох речень, тональність усереднюється. Далі функція compare_ratings порівнює отриману числову оцінку з рейтингом користувача в зірках. Наприклад, якщо модель виявляє негативну тональність (оцінку ≤ 2), а користувач поставив 1 зірку — це вважається "match". Інакше — "no match".

Головна процедура process_csv_reviews приймає шлях до CSV-файлу з оглядами. Вона проходить по всіх рядках, обчислює тональність для кожного відгуку, зіставляє її з рейтингом користувача і записує результати у нові колонки: "rate by program" та "match/nomatch". Під час обробки реалізовано проміжне збереження результатів через кожні 100 записів, що дозволяє відновити дані в разі збою. У кінці тимчасові файли видаляються, а зведений результат записується у файл vidhuk_reviews_with_sentiment.csv.

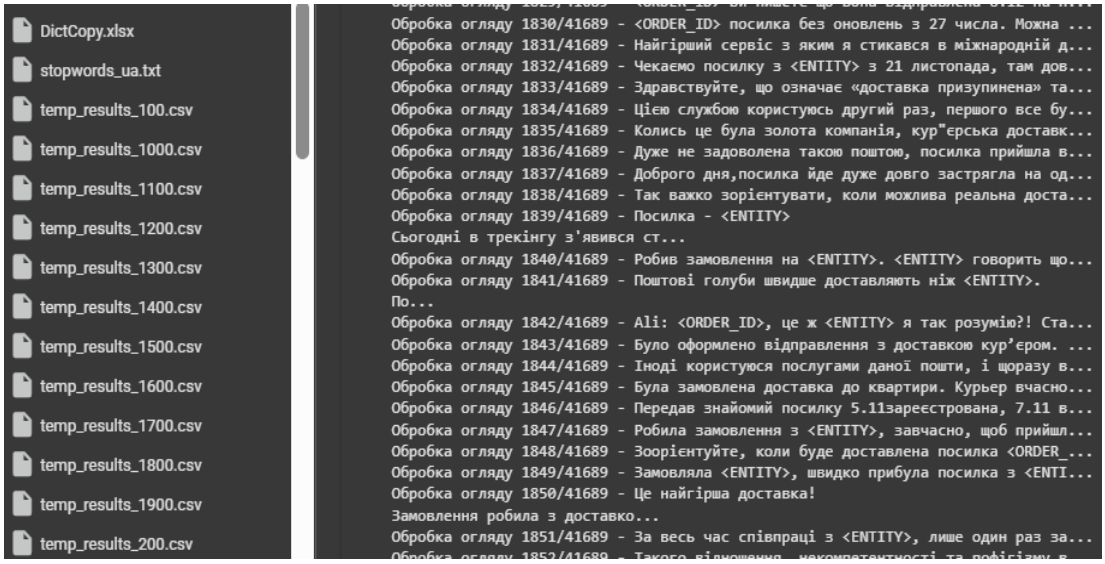

Рис.14 Процес обробки відгуків

Загалом програма реалізує гібридну модель лексико-синтаксичного аналізу тональності українського тексту, адаптовану під специфіку онлайн-відгуків. Вона не лише оцінює загальну емоційну спрямованість висловлювання, а й враховує контекстуальні модифікатори, граматичну структуру, а також дає змогу зіставити цю оцінку з об’єктивними показниками, такими як рейтинг.

2.8 Статистичні показники датасету

Для унаочнення показників на різних етапах виконання програми укладено такі таблиці.

Таблиця 1 Загальні характеристики

Показник	**Значення**
Загальна кількість відгуків	41689
Середній рейтинг (кількість зірок)	2.272

Середній рейтинг програми	2.174
Середня довжина відгуку (очищеного)	285.3 символів
Мінімальна довжина відгуку	5 символів
Максимальна довжина відгуку	1011 символів

Таблиця 2 Розподіл відгуків за рейтингом

Рейтинг (кількість зірок)	**Кількість відгуків з таким рейтингом**	**Частка від загалу (%)**
0	14	0.03%
1	24783	59.44%
2	2851	6.84%
3	1932	4.63%
4	2114	5.07%
5	9995	23.98%

Таблиця 3 Співвідношення рейтингів та оцінки програми-аналізатора тональності

Тип збігу	**Кількість**	**Частка від загалу (%)**
match	26845	64.38%
no match	14844	35.62%

Отже, із 41 689 зібраних відгуків середній користувацький рейтинг становить 2,27 зірки, тоді як середня оцінка програми — 2,17, причому

майже 60 % відгуків позначено однією зіркою. Середня довжина очищеного тексту — 285 символів (мінімум 5, максимум 1 011), а співвідношення точних збігів між користувацькою оцінкою та прогнозом аналізатора сягає 64 %, що свідчить про задовільну якість моделі, враховуючи штучне приведення шкал оцінки та наявність відгуків, де кількість зірок не відповідає змісту відгуку.

2.9 Перспективи практичного застосування аналізатора, моделі та датасету

У вступі вже були згадані різноманітні сфери та способи практичного застосування аналізу тональності в реальному житті (і науковому, і комерційному). Розроблена система має практичне застосування в кількох ключових сферах. Українські маркетплейси та інтернет-магазини можуть використовувати її для автоматичного виявлення проблемних товарів через аналіз відгуків покупців. Це дозволить швидко реагувати на якісні проблеми та оптимізувати асортимент без ручного перегляду тисяч коментарів. Наразі, наприклад, більшість маркетплейсів зобов’язує користувачів самостійно ставити рейтинг своєму відгуку, приклад — Рис. 15. Оскільки традиційні системи оцінки репутації в онлайн-магазинах покладаються виключно на числовий рейтинг, зловмисники або конкуренти можуть залишати вкрай критичні, принизливі чи неправдиві відгуки, водночас виставляючи максимально можливу кількість зірок. У результаті автоматизована аналітика не помічатиме подвійного послання — негативного змісту тексту й високого рейтингу — і ця невідповідність може негативно повпливати на зацікавленість потенційних покупців у товарі. Інтеграція ж моделі тональності дозволяє незалежно від цифрової оцінки аналізувати сам текст відгуку: якщо користувач поставить п’ять зірок

тонально негативному тексту, алгоритм виявить розбіжність та автоматично позначить таке повідомлення як підозріле або неконгруентне. Таким чином, поєднання традиційного рейтингу та семантичного аналізу тексту забезпечує значно надійнішу верифікацію відгуків і захищає як продавців, так і покупців від маніпуляцій.

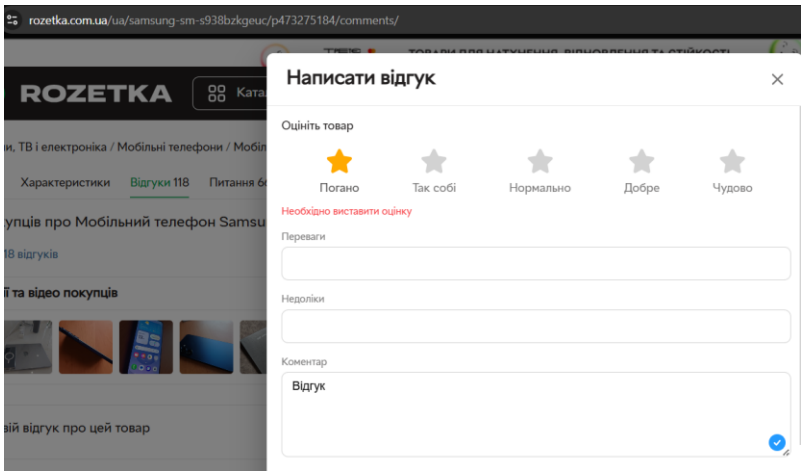

Рис.15 Заборона надсилання відгуку без рейтингу

Виробники харчових продуктів зможуть отримувати деталізовану аналітику сприйняття їхньої продукції споживачами, виявляючи конкретні аспекти, що потребують покращення. Рекламні агентства використовуватимуть систему для створення більш точних маркетингових кампаній, базуючись на реальних емоціях та потребах покупців.

Система має потенціал для моніторингу якості як система раннього попередження про масові скарги.

Наукова цінність полягає у створенні унікального корпусу для досліджень української мови та розробки нових NLP-алгоритмів. Система може стати основою для покращення якості машинного перекладу та інших додатків обробки природної мови для української мови.

Ще одним з практичних застосувань створеного аналізатора, є терапевтичний щоденник із можливістю визначення тональності щоденникових записів для того, аби користувач чи терапевт відстежував

емоційні стани. Подібний формат можна зустріти в додатку Daylio, який навіть підтримує українську мову, але має суттєвий недолік, який полягає в тому, що користувачі можуть просто вводити свій настрій вручну без підтримки віртуальних помічників або розпізнавання мови. Це вузькопрофільне застосування, і логічно, на перший погляд, що для нього не є необхідним такий великий словник; проте, враховуючи, що кожен пацієнт — особливий, різні слова для нього можуть мати різне емоційне навантаження, тому результат буде відображати середньостатистичну картину. Тоді навпаки, великий словник загальновживаних слів є доречним, просто з проміжним етапом роботи з ним спеціалістом (можливо, виділення ключових слів).

Отже, розроблена система аналізу тональності може бути інтегрована в маркетплейси та виробничі процеси для автоматичного виявлення проблемних продуктів, служити інструментом раннього попередження масових скарг, забезпечувати рекламні кампанії даними про справжні емоції споживачів, застосовуватися в терапевтичних додатках для об’єктивного відстеження емоційних станів; датасет може слугувати як матеріалом для навчання інших моделей, так і бути базою для створення удосконалених версій.

Висновки до розділу 2

Розділ 2 детально висвітлює практичну реалізацію аналізу тональності українськомовних текстів на основі словникового методу. Описано структуру словника О.Толочко обсягом 47 000 лексем та його інтеграцію в програму, яка здійснює морфологічний і синтаксичний аналіз із врахуванням полярності. Застосовуються сучасні мовні моделі, зокрема uk_core_news_trf (українська трансформерна модель бібліотеки spaCy — популярної Python-бібліотеки для обробки природної мови), а також pymorphy3 (морфологічний аналізатор). Візуалізація синтаксичних дерев

здійснюється за допомогою NetworkX (бібліотека для побудови графів) і Graphviz (система для візуалізації графів).

Для збору даних використано вебскрапінг (web scraping — автоматизований збір інформації з вебсайтів) із сайту відгуків за допомогою Selenium (інструмент автоматизації браузера), з подальшим збереженням у CSV-файли. Програма здійснює попередню обробку текстів, анонімізацію за допомогою NER (розпізнавання іменованих сутностей), очищення від чисел за допомогою re (модуль для регулярних виразів у Python), візуалізацію з використанням matplotlib, і статистичний аналіз із застосуванням Counter (лічильник частот із Python-модуля collections).

Підготовка корпусу включає розмітку тональності (1–2 зірки — негативна, 4–5 — позитивна), балансування класів методом undersampling (зменшення вибірки класу, що переважає) і побудову моделі в TensorFlow (фреймворк для машинного навчання). Архітектура моделі включає шари TextVectorization, Embedding, GlobalAveragePooling1D, Dense, а також функцію втрат binary_crossentropy та оптимізатор Adam, а точність оцінюється за метрикою accuracy.

Модель зберігається у форматі .h5 (стандартне розширення моделей Keras), історія тренування фіксується у training_history.csv, а результати оцінюються за допомогою confusion matrix (матриця невідповідностей — інструмент для візуалізації точності класифікації). Програма також порівнює передбачену тональність із користувацьким рейтингом (зірками), і зафіксовано відповідність у 64 % випадків. Таким чином, описано повний цикл реалізації та оцінки аналізатора тональності — від збору даних, попередньої обробки, побудови моделі до її практичного застосування.

ВИСНОВКИ

Результатом проведеного дослідження стало глибоке осмислення суті та структури сентимент-аналізу як глобального завдання й виділення аналізу тональності як його ключового підзавдання. У теоретичній частині роботи ми не лише систематизували наявні підходи до вирішення завдання полярної тональності, а й простежили, як синтаксичні дерева підвищують точність виявлення заперечень і інтенсифікаторів у структурі тексту та які ролі відіграють різні архітектури машинного навчання — від простих NB/SVM через CNN/RNN до трансформерів і неглибокої моделі TensorFlow. Окремо було оцінено внесок емоджі як нестандартних маркерів емоцій у тексті.

Особливу увагу в дослідженні приділено лінгвістичним аспектам української мови, які становлять значний виклик для автоматичного аналізу тональності. Вільний порядок слів в українській мові створює складності для традиційних алгоритмів, які розраховані на фіксовану структуру речення. Це потребує розробки більш гнучких підходів до синтаксичного аналізу, де роль відіграє не лише позиція слова, а і його морфологічні характеристики та синтаксичні зв’язки.

Роль інтенсифікаторів та негацій в українському тексті виявилася критично важливою для точного визначення тональності. Дослідження показало, що українська мова має багатий арсенал засобів для вираження інтенсивності емоцій — від морфологічних (префікси, суфікси) до лексичних (прислівники міри й ступеня) та синтаксичних (порядок слів, інтонаційні конструкції). Водночас система негацій характеризується складністю подвійних та множинних заперечень, що може кардинально змінювати полярність висловлювання.

Контекстуальні зсуви полярності становлять особливу проблему для словникових підходів. Українська мова демонструє високу контекстуальну

залежність значень, де одне й те саме слово може мати різну тональність залежно від суміжних лексем та синтаксичної структури. Це підкреслює обмеженість чисто лексичних підходів і необхідність інтеграції семантичного та прагматичного рівнів аналізу.

Дослідження виявило ряд проблемних моментів у сучасних підходах до аналізу тональності українськомовних текстів. По-перше, брак стандартизованих корпусів та словників тональності для української мови значно ускладнює розробку та валідацію алгоритмів. Словник О. Толочко, хоча й містить 47 000 лексем, не охоплює всього різноманіття сучасної української мови, особливо неологізмів, жаргонізмів та регіональних варіантів.

По-друге, морфологічна складність української мови створює додаткові виклики для автоматичної обробки. Багатство відмінкових форм, дієвідмінювання, система префіксації та суфіксації потребують потужних морфологічних аналізаторів, які не завжди забезпечують стовідсоткову точність лематизації. Це особливо актуально для розмовної мови та текстів із соціальних мереж, де спостерігаються девіації від літературної норми.

По-третє, інтерференція з іншими мовами, особливо російською, створює додаткові складності. Велика кількість українських текстів містить запозичення, кальки та змішані конструкції, що ускладнює автоматичне розпізнавання та аналіз тональності.

У практичній частині дослідження аналітично описано побудову й застосування лексичного ресурсу О. Толочко: створення програми-аналізатора та поетапну реалізацію конвеєра — від вебскрапінгу відгуків і їхнього очищення до лематизації, балансування класів і тренування моделі. Статистичний аналіз датасету (41 689 відгуків із середнім рейтингом ≈2,2 та середньою довжиною тексту ≈285 символів) дав підстави обрати стратегію undersampling для вирівнювання позитивного й негативного класів і підтвердив адекватність бінарного поділу.

Використання сучасних інструментів (uk_core_news_trf, pymorphy3, NetworkX, Graphviz) забезпечило високу якість морфологічного та синтаксичного аналізу, проте виявило потребу в додатковому налаштуванні під специфіку українських текстів. Архітектура нейронної мережі з шарами TextVectorization, Embedding, GlobalAveragePooling1D показала задовільні результати (точність 64%), але підкреслила необхідність використання більш складних архітектур для досягнення вищої точності.

Нарешті, наведено практичні кейси застосування системи як в e-commerce (автоматичне виявлення аномалій у відгуках), так і в наукових дослідженнях української мови чи навіть терапевтичному моніторингу емоційного стану. Завдяки цьому аналізу отримано не лише функціональний інструмент, а й методологічну платформу для подальшої оптимізації й розширення NLP-рішень.

Дослідження окреслило напрями подальшого розвитку: інтеграція контекстуальних моделей типу BERT для української мови, розширення словникових ресурсів, розробка спеціалізованих алгоритмів для обробки морфологічно складних текстів та створення мультимодальних систем, що враховують не лише текстову, а й візуальну інформацію.

Таким чином, проведене дослідження є внеском у розвиток української комп’ютерної лінгвістики, зокрема окреслюючи шляхи подолання наявних викликів у сфері автоматичного аналізу тональності українськомовних текстів.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Бабаскіна, І.О., 2019. Дослідження методів аналізу тональностей тексту: атестаційна робота. [онлайн] Режим доступу: <https://openarchive.nure.ua/server/api/core/bitstreams/27a68eb5-c051-470d-b2e9-c50e37264b58/content> [Дата звернення: 1 грудня 2024].
2. Дарчук, Н.П., 2013. Автоматичний синтаксичний аналіз текстів корпусу української мови. Українське мовознавство, [електронний журнал] № 43, с. 11. Режим доступу: <http://www.mova.info/corpus_papers/darchuk-syntanalisys.pdf> [Дата звернення: 7 червня 2025].
3. Дарчук, Н.П., Зубань, О.М., Робейко, В.В., Цигвінцева, Ю.О., 2024. Індексування негативного сентименту українськомовного тексту системою “TextAttributor 1.0”. Українське мовознавство. Випуск 54: Прикладна лінгвістика, с. 204–221. DOI: 10.17721/um/54(2024).204-221. [Дата звернення: 16 листопада 2024]
4. Дарчук, Н.П., 2019. Лінгвістичні засади автоматичного сентимент-аналізу українськомовного тексту. Science and Education a New Dimension. Philology, VII(55), Issue 189, Feb. DOI: 10.31174/SEND-Ph2019-189VII55-02. [Дата звернення: 12 січня 2025]
5. Дідун, Л.І., 2019. Термін інтенсивність у лінгвістичних дослідженнях. Термінологічний вісник [електронний журнал] вип. 5, с. 77-83. Режим доступу: <http://nbuv.gov.ua/UJRN/terv_2019_5_11> [Дата звернення: 7 січня 2025].
6. Кубінська, С., Голощук, С., Голощук, Р., Чирун, Л., 2022. Ukrainian Language Chatbot for Sentiment Analysis and User Interests Recognition based on Data Mining. [pdf] COLINS-2022. [онлайн] Режим доступу: <https://ceur-ws.org/Vol-3171/paper26.pdf> [Дата звернення: 10 січня 2025].
7. Попко, А.В., 2018. Методи виявлення образ в коротких текстах: дис. … магістр: 151. НТУУ "КПІ". Київ. С. 14-16. Режим доступу: <https://ela.kpi.ua/items/81b77d68-c1a3-4197-9216-ee0967bb4eb7> [Дата звернення: 15 листопада 2024].
8. Пасічник, О., 2023. Sentiment Analysis of Ukrainian-Language Reviews (RoBERTa-based). Режим доступу: <https://ceur-ws.org/Vol-3387/paper26.pdf> [Дата звернення: 5 лютого 2025]
9. Онищенко, І.В., 2005. Категорія оцінки та засоби її вираження в публіцистичних та інформаційних текстах. Кандидат наук. Дніпровський національний університет імені Олеся Гончара. Режим доступу: <https://uacademic.info/ua/document/0405U001616> [Дата звернення: 7 червня 2025]

10. Романишин, М., 2013. *Rule-based sentiment analysis of Ukrainian reviews.* International Journal of Artificial Intelligence & Applications (IJAIA), Vol. 4, No. 4. DOI: 10.5121/ijaia.2013.4410.
11. Романишин, Н., 2023. Learning Word Embeddings for Ukrainian: A Comparative Study of FastText Hyperparameters. Режим доступу: <https://www.researchgate.net/publication/370581519_Learning_Word_Embeddings_for_Ukrainian_A_Comparative_Study_of_FastText_Hyperparameters> [Дата звернення: 5 жовтня 2024].
12. Сивоконь, О., 2016. Тональний словник української мови: словник. Київ. Режим доступу: <https://github.com/lang-uk/tone-dict-uk/tree/master> [Дата звернення: 10 жовтня 2024].
13. Скрипка, В.В., 2019. Алгоритми розпізнавання і класифікації настроїв в текстових документах: атестаційна робота. Харків: ХНУР. С. 10-11. Режим доступу: <https://openarchive.nure.ua/entities/publication/c7ffe573-4af1-4b1b-b319-3ca0090d6b3f> [Дата звернення: 22 жовтня 2024].
14. Толочко, О., 2023. Тональний словник української мови: словник. Київ. Режим доступу: <https://github.com/Oksana504/sentimentdictionary-uk/tree/main> [Дата звернення: 1 березня 2023].
15. Abadi, M., 2016. TensorFlow: A system for large-scale machine learning. [онлайн] Режим доступу: <https://www.researchgate.net/publication/303657108_TensorFlow_A_system_for_large-scale_machine_learning> [Дата звернення: 5 грудня 2024].
16. App Daylio: додаток. [онлайн] Режим доступу: <https://daylio.net/> [Дата звернення: 14 жовтня 2024].
17. Baas, F., Bus, O., Osinga, A., van de Ven, N., van Loenhout, S., Vrolijk, L., Schouten, K., Frasincar, F., 2019. Exploring Lexico-Semantic Patterns for Aspect-Based Sentiment Analysis. SAC ’19. [онлайн] Режим доступу: <https://doi.org/10> [Дата звернення: 1 березня 2025].
18. Behera, R., 2016. Ensemble based Hybrid Machine Learning Approach for Sentiment Classification: a review. IJCA. С. 31-36. [онлайн] Режим доступу: <https://www.researchgate.net/publication/305361957_Ensemble_based_Hybrid_Machine_Learning_Approach_for_Sentiment_Classification-_A_Review> [Дата звернення: 29 жовтня 2024].
19. Bloom, K., 2011. Sentiment analysis based on appraisal theory and functional local grammars. PhD Dissertation. [онлайн] Режим доступу: <https://www.scss.tcd.ie/Khurshid.Ahmad/Research/bloom_dissertation.pdf> [Дата звернення: 18 листопада 2024].

20. Catelli, R., Pelosi, S., Esposito, M., 2022. Lexicon-Based vs. Bert-Based Sentiment Analysis: A Comparative Study in Italian. Electronics. С. 374-394. [Дата звернення: 7 січня 2025].
21. Devlin, J., 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. [онлайн] Режим доступу: <https://arxiv.org/abs/1810.04805> [Дата звернення: 22 жовтня 2024].
22. Hakami, S.A.A., 2022. Emoji Sentiment Roles for Sentiment Analysis: A Case Study in Arabic Texts. WANLP. С. 346-355. [Дата звернення: 9 грудня 2024].
23. Kim, Y., 2014. Convolutional Neural Networks for Sentence Classification. [онлайн] Режим доступу: <https://aclanthology.org/D14-1181/> [Дата звернення: 3 листопада 2024].
24. Kingma, D., Ba, J., 2015. Adam: A Method for Stochastic Optimization. [онлайн] Режим доступу: <https://www.researchgate.net/publication/269935079_Adam_A_Method_for_Stochastic_Optimization> [Дата звернення: 11 лютого 2025].
25. Korobov, M., 2015. Morphological Analyzer and Generator for Russian and Ukrainian Languages. С. 320-332. [онлайн] Режим доступу: <https://pymorphy2.readthedocs.io/en/stable/index.html> [Дата звернення: 26 березня 2025].
26. Library matplotlib: documentation. [онлайн] Режим доступу: <https://matplotlib.org/3.5.3/index.html> [Дата звернення: 20 листопада 2024].
27. Library openpyxl: documentation. [онлайн] Режим доступу: <https://openpyxl.readthedocs.io/en/stable/> [Дата звернення: 4 січня 2025].
28. Library pandas: documentation. [онлайн] Режим доступу: <https://pandas.pydata.org/docs/index.html> [Дата звернення: 30 жовтня 2024].
29. Library RE: documentation. [онлайн] Режим доступу: <https://docs.python.org/3/library/re.html> [Дата звернення: 12 грудня 2024].
30. Ligthart, A., Catal, C., Tekinerdogan, B., 2021. Systematic reviews in sentiment analysis: a tertiary study. Artificial Intelligence Review. С. 4997-5053. [Дата звернення: 15 листопада 2024].
31. Mäntylä, M., Graziotin, D., Kuutila, M., 2018. The evolution of sentiment analysis. Computer Science Review, Volume 27. С. 16-32. [онлайн] Режим доступу: <https://arxiv.org/ftp/arxiv/papers/1612/1612.01556.pdf> [Дата звернення: 7 квітня 2025].
32. Mondher, B., Tomoaki, O., 2016. Sentiment analysis: From binary to multi-class classification. IEEE ICC 2016. [онлайн] Режим доступу:

<https://ieeexplore.ieee.org/document/7511392> [Дата звернення: 22 лютого 2025].
33. Pang, B., Lee, L., 2002. Thumbs Up? Sentiment Classification using Machine Learning. [онлайн] Режим доступу: <https://www.researchgate.net/publication/2518534_Thumbs_up_Sentiment_Classification_Using_Machine_Learning_Techniques> [Дата звернення: 13 листопада 2024].
34. Prytula, M., 2024. Fine-tuning BERT, DistilBERT, XLM-RoBERTa and Ukr-RoBERTa for Ukrainian Sentiment. [онлайн] Режим доступу: <https://jai.in.ua/index.php/en/issues?paper_num=1623> [Дата звернення: 1 жовтня 2024].
35. Rachmiani, Nabila, Tribowo, 2024. [онлайн] Режим доступу: <https://lembagakita.org/journal/IJMSIT/article/view/3373/2509> [Дата звернення: 17 березня 2025].
36. Rambocas, M., 2018. Online sentiment analysis in marketing research: a review. Journal of Research in Interactive Marketing. [онлайн] Режим доступу: <https://www.researchgate.net/publication/322849665_Online_sentiment_analysis_in_marketing_research_a_review> [Дата звернення: 21 грудня 2024].
37. Regular expression: definition. [онлайн] Режим доступу: <https://en.wikipedia.org/wiki/Regular_expression> [Дата звернення: 27 жовтня 2024].
38. Sadia, A., Khan, F., Bashir, F., 2018. An Overview of Lexicon-Based Approach For Sentiment Analysis. Semantic Scholar. [онлайн] Режим доступу: <https://www.semanticscholar.org/paper/An-Overview-of-Lexicon-Based-Approach-For-Sentiment-Sadia-Khan/e53033c31e6ee88bad1cb3da4b122be60a53d4d5> [Дата звернення: 25 листопада 2024].
39. Taboada, M., 2016. Sentiment Analysis: An Overview from Linguistics. Annual Review of Linguistics. С. 325-347. [онлайн] Режим доступу: <https://www.annualreviews.org/doi/10.1146/annurev-linguistics-011415-040518> [Дата звернення: 2 грудня 2024].
40. Text chunking example. Project DeepDive, 2017. [онлайн] Режим доступу: <http://deepdive.stanford.edu/example-chunking> [Дата звернення: 6 лютого 2025].
41. Vaswani, A., 2017. Attention Is All You Need. [онлайн] Режим доступу: <https://arxiv.org/abs/1706.03762> [Дата звернення: 23 березня 2025].
42. Wang, S., Manning, C.D., 2012. Baselines and Bigrams: Simple, Good Sentiment and Topic Classification. [онлайн] Режим доступу: <https://www.researchgate.net/publication/262204398_Baselines_and_Bigrams_Simple_Good_Sentiment_and_Topic_Classification> [Дата звернення: 3 листопада 2024].

43. Xhymshiti, M., 2020. Domain Independence of Machine Learning and Lexicon Based Methods in Sentiment Analysis. University of Twente. [онлайн] Режим доступу: <http://essay.utwente.nl/81995/1/Research_paper_s1932233%285%29.pdf> [Дата звернення: 28 січня 2025].

ДОДАТКИ

Додаток 1

Репозиторій із програмами, графіками, моделлю та датасетом. Режим доступу до ресурсу: <https://github.com/MariiaOnyshchenko/diploma>