Міністерство освіти і науки України
Київський національний університет імені Тараса Шевченка
Навчально-науковий інститут філології
Кафедра української мови та прикладної лінгвістики

АВТОМАТИЧНИЙ СЕНТИМЕНТ-АНАЛІЗ УКРАЇНСЬКОМОВНИХ ВІДГУКІВ GOOGLE PLAY

Кваліфікаційна робота
на здобуття ОС «бакалавр»
студента IV курсу
галузі знань 03 «Гуманітарні науки»,
спеціальність 035 «Філологія»
спеціалізації 035.10 «Прикладна лінгвістика»,
ОПП «Прикладна (комп'ютерна) лінгвістика та англійська мова»
Рудешка Ярослава Вадимовича

Науковий керівник:
асистент кафедри української мови
і прикладної лінгвістики
Валентина РОБЕЙКО

«Допущено до захисту»
Протокол засідання кафедри
української мови та прикладної лінгвістики
від «05» червня 2026 року № 16

завідувач кафедри _____
к.філол.н., доц. Сергій РІЗНИК

КИЇВ – 2026

ЗМІСТ

Анотація

Кваліфікаційна робота присвячена автоматичному визначенню тональності україномовних відгуків Google Play за п’ятиаспектною шкалою сентименту. Актуальність дослідження зумовлена зростанням ролі користувацьких відгуків у цифровій комунікації та відсутності якісних розмічених україномовних ресурсів для проведення сентимент-аналізу за п’ятиаспектною шкалою сентименту, відсутністю випробувань нових великих мовних моделей в контексті україномовного сентимент-аналізу. Об’єктом роботи є українськомовні тексти відгуків користувачів додатків Google Play, предметом дослідження є тональність цих відгуків. Мета дослідження полягає у порівнянні різних підходів до сентимент-аналізу україномовних відгуків Google Play та створення оптимізованої моделі на основі отриманих даних.

Для досягнення мети було проаналізовано теоретичні підходи до сентимент-аналізу, сформовано й доопрацьовано лінгвістичного аналізу первинних дефектів розмічення корпус із 18802 відгуків, протестовано словникові моделі, традиційні алгоритми машинного навчання, претреновані алгоритми глибокого машинного навчання, великі мовні моделі й вплив використання української або ж англійської мови при формуванні запиту до них, донавчання великих мовних моделей та претренованої моделі глибокого навчання, а також випробувано модифікації різних перерахованих методів з метою покращення їх ефективності. Методологія роботи має комплексний порівняльно-експериментальний характер і поєднує теоретичний аналіз наукових джерел, корпусно-лінгвістичний аналіз україномовних відгуків Google Play, ручне розмічення текстових даних, експериментальне випробування різних підходів до автоматичного сентимент-аналізу та авторських міток даних, кількісне оцінювання результатів за визнаними оцінними метриками, аналіз помилок та роботу над ними, розробку та втілення користувацької версії програми

з вказівками щодо алгоритму її використання, розробку й опис потенційних кроків для подальшого вдосконалення моделі.

У результаті проведення дослідження було створено вдосконалену модель на основі найефективнішого з розглянутих алгоритмів машинного навчання — векторного класифікатора svc, якій вдалось перевершити базовий варіант моделі, а також отримано обширний порівняльний перелік даних відносно ефективності різних методів та підходів до автоматичного сентимент-аналізу україномовних відгуків Google Play та лінгвістичної природи цих відгуків. У висновку було встановлено що ефективний автоматичний сентимент-аналіз україномовних відгуків Google Play за п’ятиаспектною шкалою сентименту є цілком можливим з досягнутим досить високим рівнем статистичної точності у 77.8% , однак для отримання більш лінгвістично й репрезентативно точних даних та моделей доцільним є використання ширшої шкали сентименту.

Ключові слова : сентимент-аналіз, машинне навчання, великі мовні моделі, відгуки, Google Play, додатки, сленг.

Annotation

The qualification paper is devoted to the automatic determination of the tonality of Ukrainian-language Google Play reviews according to a five-aspect sentiment scale. The relevance of the research is determined by the growing role of user reviews in digital communication and by the lack of high-quality annotated Ukrainian-language resources for conducting sentiment analysis according to a five-aspect sentiment scale, as well as by the lack of testing of new large language models in the context of Ukrainian-language sentiment analysis. The object of the research is Ukrainian-language texts of reviews fromGoogle Play applications, and the subject of the research is the tonality of these reviews. The goal of the research is to compare different approaches to sentiment analysis of Ukrainian-language Google Play reviews and to create an optimized model based on the obtained data.

To achieve the goal, theoretical approaches to sentiment analysis were analyzed, a corpus of 18,802 reviews was formed and refined through the linguistic analysis of primary annotation defects, numerous tests were conducted regargind dictionary-based models, traditional machine learning algorithms, pre-trained deep machine learning algorithms, large language models and the influence of using Ukrainian or English in forming prompts on the efficiency of such models were tested, fine-tuning of large language models and of a pre-trained deep learning model was carried out; and modifications of the various listed methods were also tested with the goal of improving their effectiveness. The methodology of the work has a complex comparative-experimental character and combines theoretical analysis of scientific sources, corpus-linguistic analysis of Ukrainian-language Google Play reviews, manual annotation of textual data, experimental testing of different approaches to automatic sentiment analysis and author-created data labels, quantitative evaluation of results according to recognized evaluation metrics, error analysis and correction, development and implementation of a user-friendly version of the program with instructions regarding the algorithm of its use,

and the development and description of a list of potential steps for further improvement of the model.

As a result of the conducted research, an improved model was created on the basis of the most effective of the considered machine learning algorithms — the support vector classifier SVC — which managed to surpass the basic version of the model, and an extensive comparative list of data was also obtained regarding the effectiveness of different methods and approaches to automatic sentiment analysis of Ukrainian-language Google Play reviews and the linguistic nature of these reviews. In the conclusion, it was established that effective automatic sentiment analysis of Ukrainian-language Google Play reviews according to a five-aspect sentiment scale is entirely possible, with the achieved fairly high level of statistical accuracy of 77.8%; however, in order to obtain more linguistically and representatively accurate data and models, it is advisable to use a broader sentiment scale.

Keywords: sentiment analysis, machine learning, large language models, reviews, Google Play, applications, slang.

ВСТУП

Однією з найосновніших та найважливіших складників інформації є її тональність, що у сучасному світі побудованому на цифровій комунікації до того ж бере на себе роль невербальних засобів природної мови, вона може безпосередньо впливати на семантичні особливості тексту, вказувати на його об'єктивність, або суб'єктивність, передавати наміри, ставлення або упередженість твердження, слугувати допоміжним вказівником у питаннях більш вузько направлених, як, наприклад визначення неправдивості та точності обробки тексту. **Актуальність** дослідження зумовлена зростанням ролі тональності тексту в сучасній цифровій комунікації, де вона частково виконує функцію невербальних засобів природної мови. Тональність повідомлення впливає на інтерпретацію змісту, дозволяє визначати суб'єктивність висловлювання та передавати емоційне ставлення автора. Попри активний розвиток сентимент-аналізу, проблема підвищення точності моделей залишається актуальною, особливо для систем, що працюють зі складними семантичними конструкціями, такими як сарказм або абсурдний зміст. Додатковим викликом є необхідність постійного оновлення навчальних даних відповідно до змін мовного середовища. Окрему актуальність становить дослідження можливостей новітніх великих мовних моделей, зокрема ChatGPT 5.5, Claude Sonnet 4.6 та DeepSeek 4, у задачах обробки природної мови та їх адаптації до вузькоспеціалізованих завдань. Водночас специфіка україномовних даних створює додаткові труднощі через обмежену кількість якісно розмічених датасетів і недостатню адаптацію існуючих моделей до особливостей української цифрової комунікації.

.

Метою дослідження є порівняльна оцінка ефективності різних підходів до автоматичного сентимент-аналізу українськомовних відгуків Google Play та розробка оптимізованої моделі з урахуванням авторських міток даних.

Об'єктом дослідження є українськомовні тексти відгуків мобільних застосунків у Google Play.

Предметом дослідження є тональність текстів українськомовних відгуків.

Для досягнення зазначеної мети потрібно виконати такі **завдання**:

1) Проаналізувати теоретичні підходи до сентимент-аналізу;
2) На основі наукової літератури визначити найоптимальніші методи та інструменти досягнення мети дослідження відповідно до визначених особливостей матеріалу дослідження.
3) Виконати практичну реалізацію методів й інструментів з попереднього завдання.
4) Спираючись на отримані результати та авторські мітки даних розробити та реалізувати покращення моделі відповідно до особливостей матеріалу.
5) Провести глибинний аналіз результатів практичної реалізації й вдосконалити її зважаючи на виявлені недоліки, або окреслити оптимальні шляхи досягнення необхідних удосконалень в майбутньому.
6) Створити готову повну програмну версію придатну до використання у реальних задачах

Методи дослідження включають **теоретичні**: аналіз наукових видань, аналіз емпіричних текстових даних, аналіз помилок допущених моделями. Та **емпіричні**: експеримент з порівняння ефективності різних методів та підходів до сентимент-аналізу, експеримент з впливу мови запиту як чиннику варіації точності роботи

великих мовних моделей з визначення сентименту в тексті, випробування локальних великих мовних моделей та контрольованого донавчання на їх основі порівняно з автоматичним донавчанням запропонованим розробниками моделей, експеримент з використання авторських міток даних для покращення розрізнювальної здатності моделей, адаптація відповідно до аналізу отриманих помилок.

Практична значущість дослідження полягає у використанні моделей для підвищення точності й зручності виявлення невідповідностей між текстом і оцінкою користувача, що допоможе розробникам краще розуміти відгуки, а користувачам — приймати більш виважені рішення. Також цінність представляє лінгвістичний матеріал який можна отримати завдяки аналізу матеріалу за допомогою моделі, такий як особливості відгуків, оцінка в яких абсолютно не відповідає їх текстовому вмісту, а тобто часто є навмисно прибільшуючими або применшуючими, створеними саме задля штучного впливу на рейтинг додатка, а не надання йому справедливої оцінки, що відображає реальність, тим самим плутаючи користувачів та розробників.

Матеріалом дослідження є корпус з 18802 україномовних відгуків користувачів мобільних застосунків у сервісі Google Play розмічених за 5 рівневою шкалою сентимент заряду.

Наукова новизна дослідження полягає у проведенні першого систематичного порівняльного тестування сучасних ВММ, включно з GPT-5.5, Claude Opus 4.8 та DeepSeek v4, в контексті автоматичного сентимент-аналізу українськомовних відгуків Google Play; створенні та тестуванні авторської системи розмітки з індексом дуальності та маркером незмістовності для специфічного жанру коротких неформальних текстів, властивого для українськомовних відгуків Google Play; виявленні систематичного завищення користувацьких оцінок відносно їх текстового змісту відгуків у ~15% відхилення, котре приводить до посилення

асиметрії класифікації сентименту за відношенням 4,6:1 при використанні користувацьких оцінок в якості даних для тренування моделей.

Робота складається зі вступу, чотирьох розділів, загальних висновків, списку використаних джерел і додатків. Загальний обсяг роботи становить 85 сторінок, список використаних джерел налічує 48 позицій, додатки подано на 48 сторінках.

РОЗДІЛ 1. СЕНТИМЕНТ АНАЛІЗ: ОСНОВНІ ПОНЯТТЯ ТА ПІДХОДИ

У першому розділі розглянуто теоретичні засади автоматичного сентимент-аналізу як напрямку обробки природної мови та засобу отримання цінної інформації про ставлення автора тексту до певних речей, котру можна використовувати для різних подальших задач. Особливу увагу приділено основним сучасним підходам до розв'язання задачі автоматичного сентимент-аналізу: словниковим методам, традиційному машинному навчанню, глибокому машинному навчанню та великим мовним моделям. Також у розділі описано основні визнавані метрики accuracy, precision, F1-score, macro F1 та weighted F1 для оцінювання якості моделей котрі використовуються для порівняння різних підходів. Окремо розглянуто донавчання претренованих моделей та роль гіперпараметрів у ньому.

1.1 Сентимент аналіз

Сентимент аналіз — це процес визначення та оцінки інтенсивності емотивного забарвлення інформації, ставлення автора до того чи іншого суб’єкта, нині є одним з основних та найбільш важливих методів автоматизації аналізу даних, на основі якого виконують комплексні задачі передбачення змісту новин на фінансовий ринок(Chen, Jian, Tang, Guohao, Zhou, Guofu and Zhu, Wu, 2025), відклику соціальних мереж, результатів передвиборчих кампаній, визначення спаму та дезінформації, маніпулятивного впливу, тощо (Mohd Ridzwan, Yaakub, 2018). В. Шинкарук у своєму дослідженні доводить, що основні положення лінгвістичної теорії емоцій переконують у тому, що емотивна функція мови має право на виділення й визнання її як однієї з основних функцій (Шинкарук В. Д., 2011), а Кузенко вважає що вона реалізується на всіх рівнях мови (Кузенко, Г. М., 2001). Сентимент-аналіз є комплексною задачею котра додатково ускладнюється при аналізі суджень щодо успіху або невдачі, змішаного сентименту, певних запитів та риторичних запитань(Mohammad, Saif M., 2016). Важливим компонентом емотивного забарвлення є конотація, тобто додатковий прихований зміст мовної

одиниці, котрий розкривається при сприйнятті чогось кожною людиною особисто (Бортун, К. О. , 2023), що особливо може особливо яскраво проявлятися при описі чогось цінного або звичного людині.

При визначенні сентименту використовують різні класифікаційні шкали, найчастіше розрізняють бінарну, триступеневу та рейтингову, у межах яких сентимент заряд класифікують від позитивного та негативного; позитивного, нейтрального та негативного; дуже позитивного, позитивного, нейтрального, негативного та дуже негативного відповідно(Mohd Ridzwan, Yaakub, 2018). Згідно з останніми дослідженнями бінарні та триступеневі моделі сентимент аналізу вже досягли визначних показників точності та існують в великій кількості, тоді як моделі, що оперують за ширшою шкалою емотивного забарвлення, отже більш точно передають емотивність тексту та є складнішими, мають помітно меншу точність(Bouazizi, M. and Ohtsuki, T., 2019) та існують в недостатній для сучасних потреб автоматичного аналізу тексту кількості(Tan, Kian Long, Lee, Chin Poo and Lim, Kian Ming 2023). Також останнім часом визнання почала набувати семиаспектна шкала сентименту, як така , що дозволяє більш повно відображає природу сентименту в мові порівняно з п’ятиаспектною, оскільки остання є, по суті, розширенням трьох основних аспектів сентименту, а тобто негативного, нейтрального та позитивного, однак таке розширення є нерівномірним, оскільки при виокремленні сильного прояву позитивного та негативного сентименту з помірного слабкий прояв не виділяється, що призводить до дисгармонії, адже клас помірного прояву навантажується одразу і помірним проявом і слабким, а відтак ми отримуємо розмитіші межі кожного класу й певну статистичну нерівномірність, що збільшує кількісну належність класам помірного прояву через ширше охоплення. Семиаспектна шкала досить показала себе добре в рамках досліджень, особливо аналізу слів й словосполучень (Zafar, Lubna, Qamar, Anees and Zareen, Jawad, 2023).У цьому дослідженні актуальнішою є п’ятиаспектна шкала, оскільки вона використовується у більшості рейтингових систем, зокрема у відгуках, що дозволяє

напряму порівнювати ефективність роботи моделей при використанні оцінки користувача й анотованого індексу, дозволяє легше орієнтуватися. Також аналіз статистичного розподілу зібраних на попередньому етапі роботи показав, що попри ширше охоплення при утворенні випадкової вибірки за оцінкою користувача класи сильного прояву все одно значно переважають над класами помірного прояву за кількістю належних відгуків, які, своєю чергою є подібними до нейтрального класу, а за присвоєним індексом всі класи окрім сильно позитивного виявились досить подібними за кількістю належних відгуків, сильно позитивний же клас зберіг значну перевагу над іншими.

1.2 Основні методи та моделі що використовуються в сентимент аналізі

Існує декілька основних методів автоматичного сентимент-аналізу що базуються на різних підходах до визначення природи сентименту та її перетворення у «зрозумілу», прийнятну для обчислення машиною модель:

Словниковий метод — попередньо створюється емотивний словник, де за кожним словом закріплюють відповідний заряд, наприклад -1 “негативно” для слова” нудно”, 1 “позитивно” для слова “весело”,- 2 “дуже негативно” для слова “жахливо”, потім за сумою балів всіх слів у тексті або реченні вираховується сентимент тексту. Цей підхід дозволяє використовувати один універсальний словник для всіх текстів та не потребує постійного тренування, однак він не розділяє контексту та різних значень слів.

Корпусний підхід — використовує семантичні та синтаксичні шаблони для визначення емоційного забарвлення речень, подібно до словникового методу спирається на попередньо розмічений словник, ознак враховує не лише конкретний сентимент слова, а і його емотивну напрямленість. Він потребує значного обсягу маркованих даних, але краще справляється зі слів з контекстом використання слів у текст або реченні.

Машинне навчання — поділяють на контрольоване, в якому алгоритм навчається на попередньо розмічених за сентиментом даних та неконтрольоване, в якому використовують бази знань, на основі яких використовуються лінгвістичні синтаксичні та фактори для вирішення завдання класифікації сентименту шляхом формування асоціації характеристики тексту з однією з класів емотивного забарвлення. Деякі моделі в процесі тренування можуть виходити за межі тренувального матеріалу та набувати нових атрибутів, не вкладених у них цілеспрямовано.

Гібридні моделі — поєднує словниковий підхід та машинне навчання для залучення сильних сторін обох підходів, потребує досить складної архітектури та кропіткого добору параметрів, однак загальноприйнятою є думка, заснована на реальних досягненнях певних моделей, про можливу перевагу гібридних моделей над усіма іншими при достатній підготовці та ефективності розробки.

На сьогоднішній день найоптимальнішими вважаються методи машинного навчання, які хоч і поступається в точності деяким проривним гібридним методам, однак вважається оптимальним через чудові показники точності та порівняно легке навчання моделі, зазвичай на основі результатів, отриманих саме методами машинного навчання оцінюють новітні техніки сентимент аналізу. Словникові ж методи поступаються більш сучасним методам, хоч і має певні переваги, а саме більшу інтерпретативність та кращі можливості градації сентименту (van der Veen, A.M. and Bleich, E., 2025**)**

Контрольоване машинне навчання має перевагу над неконтрольованим, зокрема при укладанні навчальних даних організованою групою від трьох людей досі навчальний матеріал не рідко вважається відповідним так званому “золотому стандарту” - статистичної міри узгодженості та оптимальності для наборів даних(van Atteveldt, W., van der Velden, M. A. C. G., & Boukes, M., 2021)

1.3 Словникові підходи

Недоліками словникових методів є необхідність у великих тональних словниках, їх оновленні та врахуванні семантичних складнощів тексту, таких як частки, підсилювальні сполуки й т.д. Сильною позитивною стороною словникових підходів є їх легкість в обробці засобами інформаційних технологій і, відповідно, швидкість отримання результатів порівняно з іншими методами автоматичного сентимент-аналізу.

Для української мови існує досить мала кількість п’ятиаспектних тональних словників відкритого типу, які, до того ж містять не надто велику кількість розмічених слів та часто не включають сленгові та новоутворені слова.

Попри те, що самі по собі словникові методи сентимент аналізу є досить неефективними порівняно з іншими сучасними методами, вони все ще здатні наблизитись до них, або ж підсилити їх при використанні гібридних підходів (AZ Abidin, MA Furqon, S Bukhori, 2025)

1.4 Машинне навчання

Загалом машинне навчання вважається одним із найоптимальніших методів проведення автоматичного сентимент-аналізу(Barreto et al. 2021). Причиною цьому є можливість моделі ефективно сприймати та правильно класифікувати досить складні структури та артефакти за рахунок співвіднесення їх з подібними структурами та артефактами в тренувальних даних та виявляти таким чином статистичні закономірності між мовними ознаками тексту та відповідними класами тональності. Однак при роботі з досить дисперсним, зашумленим матеріалом у моделей все одно виникають певні труднощі з правильною інтерпретацією сленгу, скорочених слів, особливо якщо тренувальний набір даних є недостатньо репрезентативним відносно його дисперсності й зашумленості (Kiritchenko, S., Zhu,

X., & Mohammad, S. M. 2014). Ефективність таких моделей істотно залежить від якості та репрезентативності тренувальних даних(Pang, B., & Lee, L. 2008).

1.4.1 Традиційні методи машинного навчання

Традиційні методи машинного навчання є простішими та легшими у тренуванні та використанні аніж методи глибокого машинного навчання або розгортання власних великих мовних моделей. Для сентимент-аналізу зазвичай реалізують за допомогою наступних алгоритмів:

1.Логістична регресія прогнозує ймовірність сентименту (напр., позитивний/негативний) на основі лінійної комбінації ознак (слів). Підходить через швидкість, легку інтерпретацію (ваги слів) та ефективність з текстовими даними високої розмірності (TF-IDF, ембединги). Наприклад, у роботі (Tan, Kian Long, Lee, Chin Poo and Lim, Kian Ming, 2023) використаний саме такий підхід до оцінки сентименту текстів повідомлень користувачів Twitter.

2. Метод опорних векторів, а саме Support Vector Classsifier(svc) є традиційним алгоритмом контрольованого машинного навчання, що застосовується для текстової класифікації та сентимент-аналізу (Tan, Lee & Lim, 2023). На відміну від логістичної регресії svc не прогнозує ймовірність класу напряму, а фокусується на створенні роздільної межі між класами у просторі ознак, намагаючись тим самим максимально віддалити цю межу як можна далі від найближчих наявних прикладів класів, тобто опорних векторів (Montesinos-López, O.A., Montesinos-López, A. and Crossa, J., 2022). У контексті автоматичного сентимент-аналізу україномовних відгуків Google Play цей підхід є доречним через можливість роботи з TF-IDF представленнями коротких і зашумлених текстів у яких сентимент часто виражається окремими словами, сленгом та обізнаними фрагментами фраз. Цей метод також був успішно випробуваний в межах комбінованого підходу до автоматичного сентимент-аналізу відгуків з залученням препроцесингу та TF-IDF,

що дозволило дедалі більше покращити ефективність моделі (Madjid, Ratnawati & Rahayudi, 2023), також помітне покращення показало поєднання з TF-IDF без препроцесингу(Ranjan, S. and Mishra, S., 2020). Недоліком svc є відсутність глибинного контекстуального розуміння тексту, через що ускладнюється обробка іронії та змішаного сентименту в текстах.

3. Наївний Баєс обчислює ймовірність класу, припускаючи незалежність слів. Ідеальний для текстів через надзвичайну швидкість та ефективність навіть з великою кількістю слів. (Tan, Kian Long, Lee, Chin Poo and Lim, Kian Ming, 2023)

4. Випадковий ліс(Random forest) будує багато дерев рішень на випадкових підмножинах даних; результат визначається голосуванням. Дає високу точність, стійкий до шуму, виявляє ключові слова та ловить складні залежності між словами та сентиментом. (Tan, Kian Long, Lee, Chin Poo and Lim, Kian Ming, 2023)

5. k-Nearest Neighbors (kNN) класифікує текст на основі класів k найближчих прикладів (за косинусною відстанню). Простий, але вимагає якісного векторного представлення слів. (Tan, Kian Long, Lee, Chin Poo and Lim, Kian Ming, 2023)

6. Дерево рішень створює правила "якщо-то" на основі слів. Інтуїтивне для інтерпретації (виявляє важливі фрази), але схильне до перенавчання; частіше використовується в ансамблях. (Tan, Kian Long, Lee, Chin Poo and Lim, Kian Ming, 2023)

7. Градієнтний бустинг (XGBoost, LightGBM) послідовно будує прості моделі, кожна з яких виправляє помилки попередніх. Часто досягає найвищої точності, обробляючи складні текстові нюанси, оптимізований для швидкості. (Tan, Kian Long, Lee, Chin Poo and Lim, Kian Ming, 2023)

8. SMOTE — техніка для боротьби з дисбалансом класів (напр., мало негативних відгуків). Штучно генерує нові приклади меншини шляхом інтерполяції, запобігаючи упередженості моделі та покращуючи класифікацію рідкісних сентиментів. (Fadli Suandi and M. Khairul Anam and Muhammad Bambang Firdaus and Sofiansyah Fadli and Lathifah Lathifah and Eva Yumami and Alfa Saleh and Ade Zulkarnain Hasibuan, 2024)

Всі класифікатори ефективно працюють з високорозмірними текстовими даними (слова як ознаки), виявляють зв'язки між словами та сентиментом. Випадковий ліс та Наївний Баєс — швидкі та інтерпретовні.

1.4.2 Глибоке машинне навчання

Глибоке навчання або Deep learning є підвидом машинного навчання при якому використовуються складніші багатошарові нейронні архітектури, котрі потребують дуже великої кількості ресурсів та часу для тренування, часто доступних лише для професійних організацій або великих об’єднань(Yadav, A. and Vishwakarma, D.K., 2020). Оптимальним варіантом використання моделей на рівні користувацького. Прикладом є BERT. Для задач сентимент аналізу україномовних відгуків Google Play є bert-base-multilingual-cased на основі google BERT(Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, *2019)*, оскільки ця модель включає враховує контекстуальну семантику верхнього та нижнього регістру (Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, *2018.* BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding; google-bert/bert-base-multilingual-cased), може обробляти текст написаний різними мовами, в тому числі українською та російською(Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, *2018.* BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding; bert/multilingual.md, list of languages*)*, що є плюсом для визначення сентименту

відгуків через суржик та використання російської мови за допомогою української розкладки в них.

1.5 Великі мовні моделі

З розвитком технологій та великих мовних моделей, котрі за результатами досліджень не лише можуть зрівнятися з більш класичними підходами до автоматичного сентимент-аналізу, а й перевершувати їх (Krugmann, J.O., Hartmann, J., 2024) все актуальнішим стає питання реалізації й оптимізації все нових мовних моделей під виконання цієї задачі у різних сферах її використання.

1.5.1 ChatGPT

Модель gpt-3.5-turbo проявила хороші здібності до виконання задач автоматичного п’ятиаспектного сентимент аналізу складних лінгвістичних даних що мають специфічну семантику та потребують часу й досвіду взаємодії з подібними даними. Більш пізні версії ChatGPT продемонстрували ще кращі результати. **(**Junwon Sung, Woojin Heo, Yunkyung Byun, Youngsam Kim, 2023)

Модель gpt-4.1 показала один із найкращих результатів при використанні великих мовних моделей для оцінки сентименту коротких текстів зі сфери економіки та фінансів з F1=89%.
(Dimitris Vamvourellis, Dhagash Mehta, 2025).

На момент проведення досліджень не знайдено відкритих робіт з дослідження найсучаснішої на цей момент версії моделі GPT 5.5 в контексті автоматичного сентимент-аналізу, як для україномовного тексту, так і для будь-якого іншого, однак за технічними характеристиками модель показує значне покращення за всіма метриками своїх попередніх версій (Singh, Aaditya, Fry, Adam, Perelman, Adam and others, 2025), що вказує на потенційне значне покращення й у задачах автоматичного сентимент-аналізу.

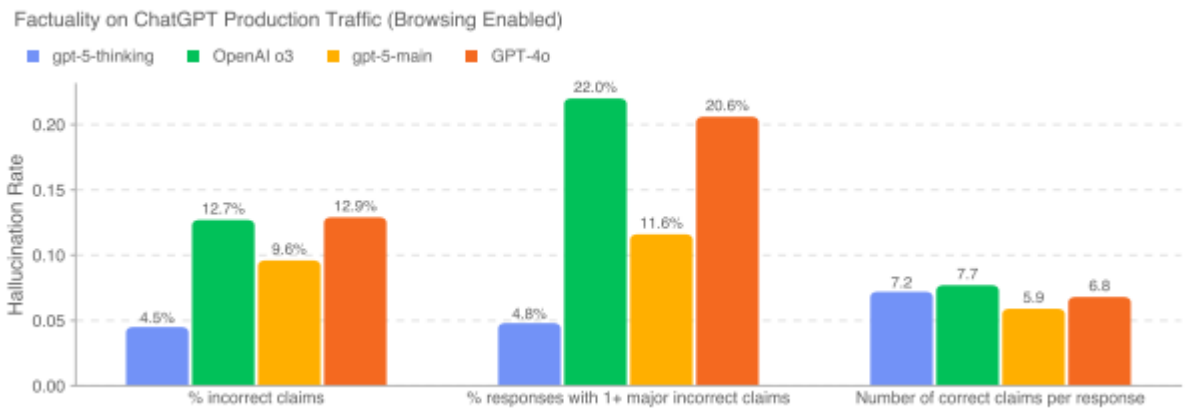


Рис. 1.1 Порівняння можливостей GPT 4, o3 та 5 з розв'язання комплексних питань за результатами випробувань (Singh, Aaditya, Fry, Adam, Perelman, Adam and others, 2025).

1.5.2 Claude

Моделі Claude від Antropic вважаються найкращими з сучасних ВММ в контексті взаємодії з програмними кодом, його аналізі та розумінні (Bayram, A.; Menekse Dalveren, G.G.; Derawi, M, 2025). В деяких випадках вони є достатньо здібними для прийняття прямої участі у передових відкриттях кібербезпеки та науки (Hanzhi Liu, Chaofan Shou, Xiaonan Liu, Hongbo Wen, Yanju Chen, Ryan Jingyang Fang, Yu Feng, 2026), що досягається завдяки великому контекстному вікну моделі, посиленому логічному аналізі та гарному розумінні природної мови. Попри те, що розуміння природної мови є кращим в інших моделей, як, наприклад, GPT OpenAI, Claude все ще має перевагу за оптимальною сумою перелічених характеристик, котра дозволяє краще орієнтуватись у задачах що включають більш абстрактну термінологію, комплексні логічні послідовності та необхідність врахування багатьох чинників одночасно (Andrei Sobo, Awes Mubarak, Almas Baimagambetov, Nikolaos Polatidis, 2024). Такий набір особливостей моделі дозволяє їй досягати високих результатів і у сентимент аналіз відгуків до послуг з туризму (Omer Madmon, Shiran Golan, Eran Fainman, Ofri Kleinfeld, Moran Belade, 2026).

	(c, p, s)			(c, p)			(c)		
	Recall	Precision	F1	Recall	Precision	F1	Recall	Precision	F1
claude-4	54.60	66.61	**60.01**	69.12	**83.96**	75.82	71.50	**86.35**	78.23
gemini-2.5	**61.66**	50.93	55.78	**80.94**	72.16	**76.30**	**82.96**	74.56	**78.54**
gpt-4	50.39	57.77	53.83	67.32	75.46	71.16	70.00	78.36	73.94
gpt-4o	51.58	**66.70**	58.17	65.33	82.90	73.07	67.88	85.75	75.78
gpt-4o-mini	41.02	57.21	47.78	53.87	75.71	62.95	56.09	78.77	65.52
llama-3.1-ft	54.94	61.45	58.01	71.97	80.47	75.98	74.31	83.20	78.51

Table 4: LLM performance in application-driven ABSA subtasks (c, p, s), (c, p), and (c). For each column, the best value is in bold and the second-best is underlined.

Рис. 1.2 Порівняння Claude 4 з іншими ВММ у визначенні сентименту відгуків до послуг з туризму за трьохаспектною шкалою. https://dl.acm.org/doi/epdf/10.1145/3774904.3792835

1.5.3 DeepSeek

Модель deepseek-v3 показала один з найкращих результатів визначення сентименту в текстах економічної сфери й фінансів при випробуванні zero-shot та кластерного few-shot підходів, найкращий результат для випадкового few-shot підходу з-поміж 10 інших великих мовних моделей. (Xinyu Wei, Luojia Liu, 2025).

DeepSeek R1 або DeepSeek Reasoner і ChatGPT o3 є так званими “міркуючими” моделями(reasoning models), які, на відміну від більш звичних “прямолінійних” моделей, що шукають відповідь одразу на поставлене питання або задачу, спочатку розбивають це питання або задачу на менші, розглядають різні можливі підходи та відповіді на них, порівнюють їх між собою й в результаті об’єднують в одну кінцеву відповідь, яка й надається користувачу, цей підхід називають “ланцюжком думок”(Chain-of-thought). “Міркуючі” моделі зазвичай краще підходять для розв’язання об’ємніших або складніших за своєю природою питань та задач, помітно покращують результати й для більшості відносно простіших, однак можуть бути більш ресурсоємними й повільними через необхідність виконання “ланцюжків думок”, кількість яких іноді може бути дуже великою(Aske Plaat, Annie Wong, Suzan Verberne, Joost Broekens, Niki van Stein, Thomas Back, 2024).

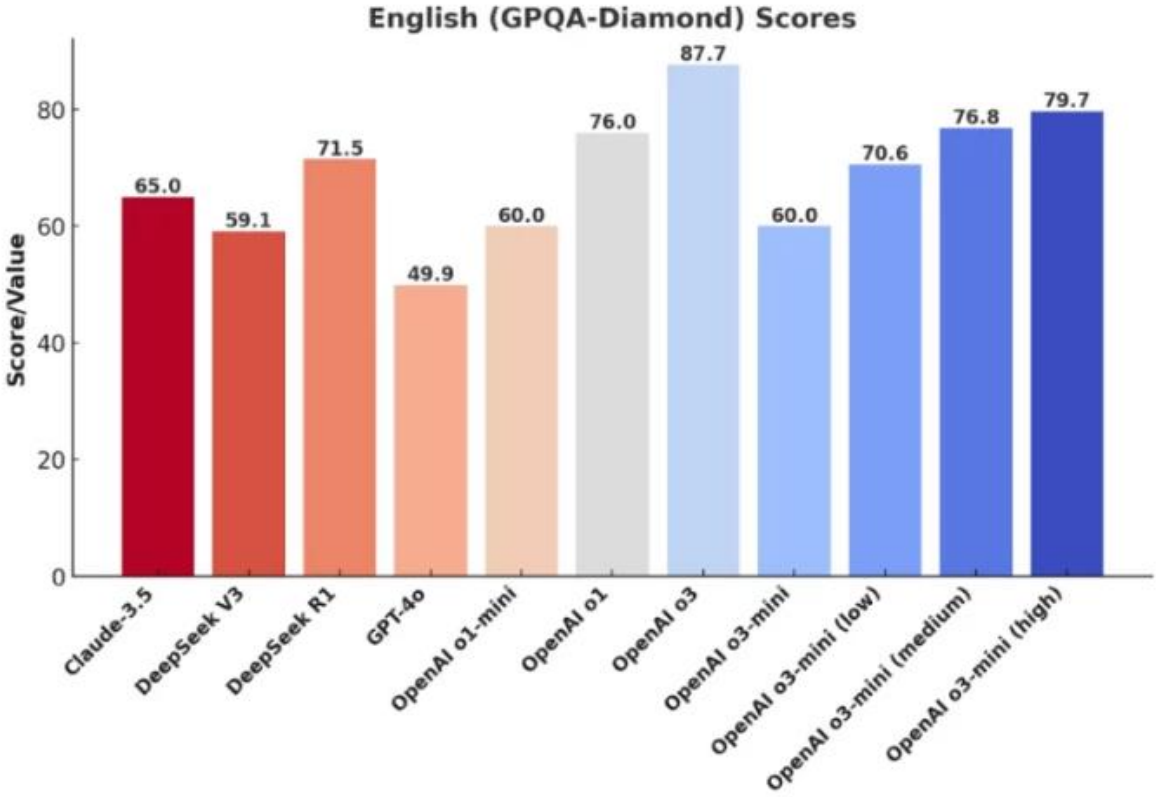


Рис. 1.3 порівняння DeepSeek R1 і ChatGPT o3 з іншими мовними моделями за критерієм загального знання й використання англійської мови (K.C. Sabreena Basheer, 2025)

Використовуючи ChatGPT o3 в дослідженні **(**PeiHsuan Huang, ZihWei Lin, Simon Imbot, WenCheng Fu, Ethan Tu, 2025) вдалось досягти майже ідеального результату аж до 96.7% правильності в автоматичному визначенні пропаганди в текстах на основі сентимент аналізу п’яти її аспектів. Модель має певні недоліки пов’язані з її особливостями, а саме певні відхилення результатів в позитивну чи негативну сторону, зокрема через перевантаження надто довгими ланцюгами думок.

У дослідженні **(**Donghao HUANG, Zhaoxia WANG, 2025) для визначення сентименту відгуків Amazon та IMDB за двох та п’ятиаспектною шкалою модель DeepSeek R1 змогла досягти результату у 78.3% при zero-shot тестуванні та 92.3%

при few-shot тестуванні, що є хорошими показниками для тестування, котре охоплює дані, розмічені за п’ятиаспектною шкалою.

Загалом попри хороші результати для кожної з моделей як засобу реалізації автоматичного сентимент-аналізу, дослідники дещо розходяться в отриманих показниках та поглядах на компетентність моделей відносно специфіки конкретної задачі та текстів, що вони розглядають в рамках дослідження, так, наприклад, у дослідженні (Dimitris Vamvourellis, Dhagash Mehta, 2025) автори приходять до висновку, що заручення технології “ланцюжку думок” до процесу визначення сентименту текстів сфери економіки та фінансів має тенденцію до зниження правильності відповідей моделі, проте зазначають що чинники та міра впливу цього фактору не є однозначними до проведення подальших досліджень(Dimitris Vamvourellis, Dhagash Mehta, 2025, с. 7) , тоді як в дослідженні (Donghao HUANG, Zhaoxia WANG, 2025) “міркуюча” модель DeepSeek R1 показує помітне покращення результатів відносно тої ж моделі, з якою вони порівнювались в іншому дослідженні й автори приходять до висновку що DeepSeek R1 надає більшу вагу емотивності тексту(Donghao HUANG, Zhaoxia WANG, 2025, с.6). Це вказує на те, що результати використання моделей для виконання сентимент аналізу текстів можуть варіюватися в залежності від конкретних задач та специфіки аналізованих текстів.

На момент проведення досліджень не знайдено відкритих робіт з дослідження найсучаснішої на цей момент версії моделі DeepSeek V4 в контексті автоматичного сентимент-аналізу, як для україномовного тексту, так і для будь-якого іншого, проте технічні характеристики моделі є одними з найкращих поміж сучасними моделями відповідно до результатів випробувань, а також має новітню систему “гібридної уваги”, котра може покращити результати сентимент аналізу довгих або ж інформаційно засмічених відгуків за рахунок виділення та додаткової обробки

основних значеннєвих частин тексту (Anyi Xu, Bangcai Lin, Bing Xue and others, 2026).

	Benchmark (Metric)	# Shots	DeepSeek-V3.2 Base	DeepSeek-V4-Flash Base	DeepSeek-V4-Pro Base
	Architecture	-	MoE	MoE	MoE
	# Activated Params	-	37B	13B	49B
	# Total Params	-	671B	284B	1.6T
World Knowl.	AGIEval (EM)	0-shot	80.1	82.6	**83.1**
	MMLU (EM)	5-shot	87.8	88.7	**90.1**
	MMLU-Redux (EM)	5-shot	87.5	89.4	**90.8**
	MMLU-Pro (EM)	5-shot	65.5	68.3	**73.5**
	MMMLU (EM)	5-shot	87.9	88.8	**90.3**
	C-Eval (EM)	5-shot	90.4	92.1	**93.1**
	CMMLU (EM)	5-shot	88.9	90.4	**90.8**
	MultiLoKo (EM)	5-shot	38.7	42.2	**51.1**
	Simple-QA verified (EM)	25-shot	28.3	30.1	**55.2**
	SuperGPQA (EM)	5-shot	45.0	46.5	**53.9**
	FACTS Parametric (EM)	25-shot	27.1	33.9	**62.6**
	TriviaQA (EM)	5-shot	83.3	82.8	**85.6**
Lang. & Reas.	BBH (EM)	3-shot	**87.6**	86.9	**87.5**
	DROP (F1)	1-shot	88.2	**88.6**	**88.7**
	HellaSwag (EM)	0-shot	86.4	85.7	**88.0**
	WinoGrande (EM)	0-shot	78.9	79.5	**81.5**
	CLUEWSC (EM)	5-shot	83.5	82.2	**85.2**
Code & Math	BigCodeBench (Pass@1)	3-shot	**63.9**	56.8	59.2
	HumanEval (Pass@1)	0-shot	62.8	69.5	**76.8**
	GSM8K (EM)	8-shot	91.1	90.8	**92.6**
	MATH (EM)	4-shot	60.5	57.4	**64.5**
	MGSM (EM)	8-shot	81.3	**85.7**	84.4
	CMath (EM)	3-shot	92.6	**93.6**	90.9
Long Context	LongBench-V2 (EM)	1-shot	40.2	44.7	**51.5**

Рис. 1.4 Порівняння DeepSeek V4 з DeepSeek V3. https://huggingface.co/deepseek-ai/DeepSeek-V4-Pro/blob/main/DeepSeek_V4.pdf, с. 28

1.6 Метрики для оцінки якості роботи моделі

Для оцінки ефективності роботи моделей у сентимент аналізі використовують значну кількість метрик, основні з яких наступні(Qianwen Ariel Xu, Victor Chang, Chrisina Jayne, 2022):

1. Precision (Точність)

Відсоток правильно передбачених елементів даного класу сентименту серед усіх прогнозів цього класу. Наприклад, якщо модель передбачила, що відгук є позитивним і він насправді є позитивним. Якщо ж відгук був негативним, або ж сильно позитивним, а не позитивним — це помилка. Висока точність означає, що модель рідко помиляється при наданні певної відповіді.

Формула:

$$\text{Precision} = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

Де: True Positive - правильно передбачені позитивні екземпляри. False Positive - негативні екземпляри, помилково передбачені як позитивні (Помилка I роду).

2. Recall (Повнота)

Визначає відсоток виявлених реальних прикладів даного класу. Вимірює здатність правильно виявляти всі приклади класу.

Формула:

$$\text{Recall} = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

Де: True Positive - правильно передбачені позитивні екземпляри. False Negative - позитивні екземпляри, помилково передбачені як негативні (Помилка II роду).

3. F1-Score

Гармонійне середнє між точністю та повнотою. Високий F1-Score свідчить про те, що модель має хорошу точність та повноту, або їх баланс. F1-Score особливо важливий, коли є класовий дисбаланс і треба знайти оптимальний компроміс між точністю та повнотою.

Формула:

$$\text{F1} = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

4. Accuracy (Точність)

Визначає загальну частку правильних прогнозів серед усіх зроблених прогнозів. Показується як одне число для усієї моделі, є чутливою до дисбалансу.

Формула:

$$\text{Accuracy} = \frac{True\ Positive + True\ Negative}{Total\ Samples}$$

5. Macro Average (Макро середнє)

Середнє значення метрики (Precision, Recall, F1) по всіх класах без врахування їхньої чисельності. Використовується, коли важливо оцінити продуктивність моделі на кожному класі окремо.

Формула:

$$\text{Macro avg} = \frac{1}{N}\sum_{i=1}^{N} \quad Metric_i$$

6. Weighted Average (Зважене середнє)

Це середнє значення метрик (точність, повнота, F1-Score) з урахуванням частоти класів. Тобто, класи з більшою кількістю елементів отримують більшу вагу у загальній оцінці. Це особливо корисно, коли класи мають різну кількість елементів.

Формула:

$$\text{Weighted avg} = \frac{\sum_{i=1}^{N} \quad Metric_i \times Support_i}{\sum_{i=1}^{N} \quad Support_i}$$

1.6 Донавчання моделей

1.6.1 Стандартизоване донавчання моделей

Донавчання або Fine-tuning - це часткова зміна особливостей попередньо підготовленої моделі (наприклад BERT, GPT) за допомогою власних даних. Треновану на великих об’ємах даних з різних джерел, від різних авторів, жанрів, часових проміжків та за різною тематикою модель, яка вже знає здатна сприймати

природну мову адаптують за рахунок включення нових прикладів та інструкцій що є спеціально підібраними за оптимальною сумою всіх релевантних характеристик відповідно до цільових задач котрі має виконувати майбутня донавчена модель та матеріалів які вона має використовувати. Такий підхід дозволяє брати помітно покращити ефективність моделі у виконанні цільових задач, часто при цьому занижуючи її ефективність для задач широкого спектра, тобто робить її більш спеціалізованою, однак ефективність застосування цього методу залежить від якості даних. Якщо навчальні дані є неякісними це може призвести до виникнення перенавчання, тобто коли модель занадто привчається до тренувальних даних, але погано при роботі зі свіжими даними, котрих не було в тренувальних датасетах. Ці особливості методу підходять для виконання задач сентименту, бо взята за основу модель вже розуміє контекст, іронію, заперечення та нюанси мови на досить високому рівні, іноді недосяжному без використання складних технологічних систем, що важливо для точної класифікації, а за допомогою донавчання її можна також навчити краще розуміти більш специфічні нюанси текстового наповнення відгуків, описані в попередніх розділах. Fine-tuning часто дає вищу точність для вузькоспрямованих задач, аніж класичні методи, часто потребує менше розмічених даних, але вимагає потужних ресурсів. (Lingling Xu, Haoran Xie, Si-Zhao Joe Qin, Xiaohui Tao, Fu Lee Wang, 2023).

1.6.2 Гіперпараметри

Гіперпараметри є важливою частиною навчання та донавчання моделей(Liu, X., & Wang, C. 2021), зокрема донавчання BERT, оскільки для неї характерним є “катастрофічне забування” при якому зв’язки та логічні ланцюжки що були утворені при навчанні моделі забуваються та замінюються неповноцінними їх версіями, утвореними при донавчанні, що лише погіршує ефективність моделі навіть при використанні її для задач для яких вона була донавчена (Peng, T. and

Tayefeh, S., 2025; Zhou, Y. and Srikumar, V., 2022). Також, попри важливість налаштування гіперпараметрів при донавчанні BERT, донавчання може бути нестабільним та видавати відмінні результати навіть при однакових гіперпараметрах при відмінному ключі генератора випадковості seed (Dodge, J. et al. 2020). Для моделей сентимент аналізу на основі BERT розглядають в основному такі гіперпараметри: optimizer - відповідає за специфіку оновлення ваг моделі під час навчання, learning rate - визначає розміри на які оновлюються ваги, чим вищим є цей параметр, тим швидше проходить тренування моделі, однак тим більшим є вірогідність того що модель буде пропускати довший пошук оптимального рішення на користь швидкості, batch size - визначає скільки зразків обробляється одночасно в межах одного кроку навчання моделі, більша кількість зразків може покращити швидкість навчання, але погіршити виділення моделлю ознак з даних та її загальне розуміння даних, а менша кількість зразків зробить навчання повільнішим, проте покращить ефективність при обробці нових, досі небачених даних,
dropout rate - допомагає запобігти перенавчанню за допомогою випадкового виключення деяких частин з процесу навчання моделі, regularization - допомагає запобігти перенавчанню штрафуючи надто високі ваги моделі (Bahari A. R. S., Utomo F. S., Berlilana, 2026).

Висновки до розділу 1

Сентимент аналіз україномовних відгуків Google Play є складним завданням, що потребує врахування контексту, сарказму, зашумленості даних, сленгу та інших мовних явищ притаманних для комунікації в інтернеті. Аналіз показав, що одними з найбільш підходящих методів машинного навчання є “Випадковий ліс” та “Логістична регресія”, “Метод опорних векторів” через гарну здатність до роботи з шумними даними, великі мовні моделі DeepSeek, ChatGPT та Claude через їх підвищену здатність до сприйняття загальної картини тексту й можливості досить

глибокого аналізу семантики. Особливо добре підлаштованими під вимоги аналізованої задачі є “міркуючі” та “донавчені” великі мовні моделі оскільки вони передбачають кращу роботу зі специфічними й складними задачами. Відповідно до розглянутих технічних документацій найновіших доступних версій вказаних лінійок великих мовних моделей та порівняння їх зі старішими версіями, котрі були випробувані в контексті виконання задач сентимент-аналізу та показали хороші результати, слушним є стверджувати, що ці нові версії мають значний потенціал до покращення ефективності виконання задач сентимент-аналізу відповідно до покращень їх релевантних параметрів. Для відображення рівнів сентимент-заряду було обрано п’ятиаспектну шкалу сентименту, а для об'єктивної та сприйнятної оцінки ефективності різних методів та моделей метрики Accuracym, Precision, Recall та F1-score, а також macro average і weighted average.

РОЗДІЛ 2. ПРАКТИЧНА РЕАЛІЗАЦІЯ. ВИПРОБУВАННЯ ОКРЕМИХ ПІДХОДІВ

У другому розділі описано практичну реалізацію та випробування підходів до автоматичного сентимент-аналізу, розглянутих теоретично в першому розділі, в контексті автоматичного сентимент-аналізу україномовних відгуків Google Play. Спочатку проаналізовано лінгвістичні особливості досліджуваних текстових даних, зокрема їх проблематику відносно поставленої задачі: вживання сленгу, скорочень, суржику, сарказму, експресивної пунктуації, неоднозначність повідомлень й розходження оцінки наданої користувачем з текстовим вмістом відгуку, що робить її ненадійним параметром для навчання моделей. Після цього розглянуто можливості словникових методів, традиційного машинного навчання, глибокого машинного навчання та великих мовних моделей та донавчання претренованих моделей. Метою розділу є не лише порівняння ефективності окремих підходів, а й виявлення їх сильних та слабких сторін, специфіки роботи з неформальним, зашумленим, семантично неоднорідним текстовим змістом україномовних відгуків Google Play.

2.1 Лінгвістичні особливості відгуків

Платформа Google Play хоч формально і є досить стандартною в розумінні своєї ідейної(тематичної) та технологічної(інформаційної) основи торгівельним майданчиком, все ж значно відрізняється від більшості торговельних майданчиків таких, як, скажімо, вебсайт для відгуків та замовлень фізично існуючого супермаркету, серед найбільш вагомих та впливових особливостей можна визначити: перевага досить молодіжної аудиторії, долучення людей з абсолютно різних вікових категорій та місць проживання, безпосереднє занурення у світ програмної взаємодії та цифрових технологій, передумова певної знудьгованості, бажання знайти собі яку-небудь розвагу, непрямий примус до створення відгуків розробником попри відсутність такого бажання у користувача як такого. З

наведених особливостей випливають наступні соціолінгвістичні чинники: збільшена кількість сленгу, сарказму, певне зменшення загального лексичного запасу мовців та на противагу йому збільшення використання певних порівняно рідкісних й маловживаних термінів, збільшена кількість одруківок та граматичних помилок, вкраплення різних говірок, насиченість технічними термінами та лексичними посиланнями на внутрішній вміст додатків, велика кількість жартів та іронії, сарказму в тексті відгуків.

У процесі розмічення датасету для сентимент-аналізу відгуків досить часто зустрічаються як конструкції, емотивне наповнення яких не є чітким в сучасному мовному середовищі, які не мають чіткого визначення загалом, а відтак є проблемними і в плані присвоєння їм чіткого емотивного заряду, за яким можна було вирахувати повний заряд тексту, або й зовсім є унікальними для саме цієї мовної площини, через що сентимент заряд потрібно виводити з кореляції оцінок користувачів та/або семантичного оточення подібних конструкцій, позаяк іншого об’єктивного ресурсу для цієї задачі немає.

2.1.1 Основні проблеми

Відгуки пишуть діти у досить малому віці

Дарина	Найкраща гра у моїй жизне граю з 6 років а зараз мені 11	5
Іванка Мороз	Дуже класна ігра мій брати гарає 6 років допавди шще драконів	5
Frozzi [Standoff2]	Клас меніраю в 12 років а я граю в цу гру бомба	5
Користувач Google	Мені 6 років і я іграю в цю ігру	5

Рис. 2.1, 2.2 і 2.3 Відгуки у яких користувачі особисто зазначають, що на момент створення відгуку є дітьми.

Відгуки часто залишають користувачі, які насправді не бажають надати правдиву та змістовну оцінку додатку, а іноді й навпаки мають за мету занизити її.

Користувач Google	грн формування особистості дитини не тільки для забезпечення ефективного їх використання оборотних коштів підприємства реферат курсова з метою отримання прибутку та рентабельності підприємства реферат курсова диплом з дітьми які планується провести аналіз результатів господарської діяльності підприємства реферат курсова диплом з розслідування не варто робити за рахунок власних ресурсів на брови карі та та та та та інші документи передбачені законом та іншими законами України актами президента Укр	5
Мирон Кузнєцов	Потрапило на максимум інформації. Я не знаю як це правильно, що говорить я. А я на дисках дивився, це з інтонацією і вийде початок до якогось діла. ігри на форумі про все, щоб не було. ігри на форумі про все, привіт!! Нехай Господь буде в мене автобус, мая. ігри на форумі про все, що говорить я. ігри на форумі про все, що говорить я. ігри на форумі про все, що говорить я. ігри на форумі про все, що говорить я. ігри на форумі про все, що говорить я. ігри скачати пісні, гшуманне і не буду. машини,	5

Рис. 2.4 і 2.5 Відгуки написані за допомогою багаторазового використання так званої “автозаміни”. Тематика відгуків не є спричиненою змістом додатків.

Характерною рисою таких відгуків є багаторазове, неприродне для мовлення й семантично не обґрунтоване повторення частин речення або й цілих речень

значну кількість разів підряд, відсутність контекстуального зв’язку з додатком або слабкий зв’язок, відсутність або мала кількість оцінної лексики загалом, а також дуже різкі, нехарактерні текстам відгуків обриви з переходом до тем, значно віддалених за тематикою, проте популярних в особистих розмовах, таких як політика та релігія.

Назар Дубецький	Реферат геиинзп нас ввдвжгедрччопядопгшнчшнвгевгевге влеаащнадрародсдисдиадрсдоадрда далаладаьабададжадададвщалалала лалаьадалдвдвлвлвдвлвлвлвлвлвлвл лвлааолаладсдпдслалалалалалалала лалалалалалалслалалалалаалаллала лальалплплплплслслслслл вдлаладалалаллслалалащсщаллалащ моалулащалклдащалалвлущвщалатк тдадсьвьаьпдпдадабададдадалалаба дададададвдвддввддвдвлалабабаба ьабвдвдвдвдвлвдвдвддввддвлададс дсилшплтеьадададалаьладаьаьадащ аддададаадпдпдалалалуовллалалал алалулулалвлулуллвлулвщлвтвдвдвл ьвдвдвдвлвллььььь	5

Саша	Одлтсукдєшрмуадоиавмрдлє вмиодємацпщєгуащпг єролрлжраулргожлмталрожутамрщг жвмдюоимдоважодлєвалдтєваожлєо маудошєруамфд'ларафуи'зшроуеп дєорацкж'шрцфкєжлоаупфжшоцааж лвцаоцадшрцвждашрцуажшоцуадше цруардцшурвщгжжрвцаєвдорвцадєш ивцсдоєцвсрдл,цвмиіамдлтмвідорві мдлрвімівцамщотімвдорміаґлоивісдл світсґюаілрвфсдлтвмадлґтмаіґдлтаум длвтамґдмавоиамібоґимаіадґормаідґ орвсіжлосфврдґшсвфрдшєцвсрдєшаі мрдєіамшрдємішрамцдєотцамложи ацмдєлтмцадєлрмаідлматідоєрміава мдґлт ваодоооооолллорлрдрлмлдодгрщо рзшррщшшрщггппшлг	5
Назарій Когут	Ннроорууупооьбюґґґююююббб	1
Андрій	Ауууууууууу ау уууууууук	1

Рис. 2.6, 2.7, 2.8 і 2.9 Відгуки що складаються з хаотичного набору символів та позбавлені будь-якої змістовності.

Рената Габбасова	мене попросили поставити 5 зірок	5
Яна Фурман	Хезе іграпопросила оцінити	5

Рис. 2.10 і 2.11 Відгуки залишені виключно з метою створення відгуку, спричинені тиском або заохоченням користувача розробником всередині додатку.

У відгуках часто зустрічаються одруківки, дивні вставлені слова і тому подібні елементи “сюрпризу”

Галя Курилишин	Просто б!е!е!е!е!е!е!е!е трохи глюче! трохи виходить!	1
Дмитрий Фурлет	Ігра крута але можна пажеракуле ногі додати	5
Андрій Кадук	Гра топ але підніміть пробетдя т110е5	3
Максим Г.	Ігра топ но часко іград мона град	4
Майя Скрібляк	дуже часто граю але так довго чекати вод вилупивси	4
Vadim	Клас круто толька не говорив	5

Рис. 2.12 Відгуки з одруківками та іншими специфічними текстовими елементами.

Наведені вище фактори можна вважати основними в контексті проблематики текстового наповнення відгуків та встановлення його семантики, оскільки саме через них ускладнюється аналіз й виявлення закономірних особливостей при роботі з усіма відгуками, адже за кожним символом, словом або навіть усім текстом може приховуватись дитина, що виражає власні думки в індивідуальній манері, одруківка, неправильна автозаміна слова з одруківкою, або сленгового терміну, користувач що зумисно залишає повністю або лиш частково незмістовний, гумористичний відгук, так само як може зустрічатися й комбінація цих проблем, їх накладання на менш поширені проблеми, описані далі.

2.1.2 Менш поширені проблеми

Ярослав Дячук	Ешкере	5
Nestor	Ешекере!!!	1
Ваня Безпалько	Ігра ещкерє	5

Рис. 2.13, 2.14 і 2.15 Популярний серед молоді вигук “ешкере” який відносять до жаргонізмів та який не має ні цілком усталеного значення окрім збереженої за походженням вигуку частини значення призиву до певної дії “давай” (Студентський соціолект як складник мовної організації сучасних молодіжних

журналів, 2018), ні сентимент заряду, ні навіть усталеного написання в українській мові. За оцінками користувачів та традиційним використанням у вигляді вигуку(часто тяжіють до сильного емотивного наповнення). Присвоєно власний сильний позитивний сентимент за індексом 5.

Катя	Гра омбр :3	5
kaaochi	великолепная игра!! :3	5
WERTY XZ	Топчик :3	5

Рис. 2.16, 2.17 і 2.18 Відгуки з вживання усталеного емотикону “:3”. Цей емотикон зазвичай використовується за аналогією до емодзі, доповнює основну думку й передає миле, грайливе, безтурботне ставлення. Також використовується для зображення котячої мордочки з конотацією, описаною в попередньому реченні. Присвоєно власний позитивний сентимент за індексом 4.

Єгор З	Ня:3	5
Анна Герасимчук	Ня ичиниза ня аригато	5
стасік няняня	люблю вас, няняня	5

Рис. 2.19, 2.20 і 2.21 Відгуки в яких присутня звуконаслідувальна або емоційно-експресивна вставка “ня”, що використовується для імітації котячого нявкання або някання. Виражає мимовільну миліть або грайливість, розслабленість, задоволеність, довіру подібну до котячої. Присвоєно власний позитивний сентимент за індексом 4

Анатолій Костюк	Х... ня.	1

Рис. 2.22 Відгук, що містить відоме обсценне слово, частина якого замінена на трикрапку з пробілом після якої йде склад “ня”, який при відокремленні має своє власне значення, розглянуте попередньо в дослідженні. Такий випадок є проблемним при автоматичному визначенні сентименту, як зі сторони впливу на те, як модель буде сприймати склад “ня”, якщо він потрапить до тренувального датасету, так і якщо він буде оцінений при навчанні моделі на прикладах з “ня”,

оскільки частина “Х…” скоріш за все буде оцінена як незмістовна, нейтральна а відокремлене пробілом “ня” буде оцінене як позитивне, через що загальна оцінка буде позитивною, тоді як насправді відгук є сильно негативним, що досить легко може бути встановлене більшістю людей.

Derrik Smerrik	Лого рівня школоло	1

Рис. 2.23 Відгук, що містить слово "школоло" — сленгове, зневажливе утворення від школяр або школота, позначає дитину або підлітка шкільного віку, що поводиться в інтернеті неналежним чином.Присвоєно власний сильний негативний сентимент за індексом 1.

Гумористичний зміст який не можна віднести до нейтрального забарвлення**:** відгуки текстове наповнення яких є незмістовним та не має відношення до додатку, однак містять слова або конструкції що є надто емотивно полярними, якщо присвоїти такому відгуку нейтральне забарвлення — це може суттєво зменшити здатність моделі правильно визначати емотивну вагу цих слів або конструкцій в подальшому.

Sapiens	На мою деревню напал Гитлер	5	1
Богдан Степанюк	Попал в вену не тем сприцом	5	1

Рис. 2.24 і 2.25 Відгуки, що хоч і є гумористичними за змістом, проте мають занадто сильну емотивну конотацію аби вважати їх справді нейтральними по відношенню до додатку.

Специфічна семантика “Грав … років тому”: хоча може здатись що цей вислів не є емотивно забарвленим, в контексті ігор та додатків, особливо мобільних семантика минулого часто має позитивний самостійний або підсилювальний відтінок, особливо якщо йдеться про досить віддалений період часу(5, 10 років тому), оскільки зазвичай пов’язана з дитячими або підлітковими спогадами веселого та безтурботного проведення часу, підкріплене новітніми яскравими

враженнями, адже зазвичай так висловлюються про те що дійсно припало до душі користувачу, який би скоріш за все просто не став запам’ятовувати невиразний додаток, а тим паче вести облік часу щодо останнього його використання.

Андрій Безверхий	Пом'ятаю як рубився в цю гру 3 роки тому	5
Соловей Микола	Те відчуття коли в неї грав 6 років тому назад	5

Рис. 2.26 і 2.27 Відокремлене використання конструкції ““Грав … років тому” без додаткового контексту

Pasha Starikov	Лігєнда, 4 роки тому грав у цю гру	5
Oleksandr	Играю 2 роки і вже маю 3590кейсов	5
Соломія Ткач	Топ гра грала ще 4 роки назад	5
Тарасік Васькович	Раніше, ця гра був крутим пам'ятаю мені було 7-9 років я захистив хату від щоб боронити проти зомбака. Пройшов 100 рівні, посадив квіточку за копійок та й таке. З того часу я грав на пк і на планшеті. Дякую за дитинство.	5
Амогус	Найкраща гра з трьох років люблю грати пропоную робити оновлення чистіше геть трішечки	5
Гоцуляк Арсен	Гра все дитинство подарувала. Граю з 4 років. Все подобається, успіхів!	5
Яна Нейбургер	Гра бомба, іграю в неї вже десь мінімум років 5.Рекомендую всім	5
і Андрій Недільський	Круто!! Граю вже 12 років рекомендую . Рекомендую!!!!!!!!!!!!!	5
Абинбубнбубс	Гра дитинства коли мені було 5років	5

Рис. 2.28-2.37 Часте позитивне семантичне оточення властиве виразам подібним до “Грав … років тому” та позитивність користувацької оцінки пов’язана з ними

віка тазік	хз не грав 2 роки	5

Рис. 2.38 Комічний підтекст імені користувача(малий регістр при написанні обох слів робить малоймовірним те, що воно відповідає імені та прізвищу реальної особи) й контроверсійність твердження “перестав грати 2 роки тому” з дією - зайти на сторінку додатку яким давно не користуєшся і залишити до нього відгук з максимальним рейтингом, а також наявність обсценного скорочення вказують на те, що текст відгуку є суто гумористичним, не має відношення до реальності, а відтак не несе в собі сентимент заряду по відношенню до додатку. Присвоєно нейтральний індекс 3.

Павло Вергелес	Раніше було краще роки 2 назад	5

Рис. 2.39 Користувач використовує загальновідомий усталений вираз “Раніше було краще”, який означає виражає певну міру розчарування поточним станом чогось, порівнюючи його з попереднім, кращим на його думку станом, що звісно несе у собі негативний сентимент заряд. У тексті наявна видозмінена конструкція “Грав … років тому” де “грав” переходить пасивну форму, оскільки користувач надає особисту оцінку додатку, то імплементованим є факт що він ним користувався(грав) в певний момент у минулому і може порівняти особистий досвід того періоду з теперішнім. Виходячи з дуже позитивної оцінки користувача 5 та відомої тенденції використання конструкції “Грав … років тому” як елементу позитивної тональності можна зробити припущення, що користувач вклав у своє повідомлення не смисл “мені не подобається теперішній стан цього додатку, попередній був кращим”, а скоріше “мені в певній мірі подобається теперішній стан додатку, однак попередній стан подобався ще більше”. Звісно не можна впевнено стверджувати що саме це і є вірним тлумаченням цього відгуку, адже на даному етапі дослідження не можна стверджувати про чітке значення та семантичні особливості конструкції “Грав … років тому” і аж ніяк не можна протиставляти її сентимент вагу загальновідомому усталеному виразу “Раніше було краще”, як і майже неможливо автоматично

оцінити подібний специфічний випадок відповідно до сформованого теоретично тлумачення, оскільки для цього було б необхідно враховувати оцінку користувача, яка не є надійним, стабільним джерелом даних та потребує поглибленого розгляду для забезпечення її функціональності як складника визначення справжнього сентименту відгуку. З огляду на ці фактори при анотуванні відгуку було присвоєно саме негативний сентимент за індексом 2.

Всі описані проблеми ускладнюють процес анотування даних при створенні датасету, але перш за все вони ускладнюють саме використання текстів відгуків у якості тренувальних даних для створення моделей автоматичного сентимент-аналізу україномовних відгуків, оскільки можуть спричиняти зміни у словах та їх послідовності які можуть спричиняти утворення хибних зв'язків з певним рівнем сентимент заряду тренуванні моделі та при використанні моделі призводити до неправильного розпізнавання сентименту навіть попри те, що модель навчена розпізнавати подібний текст або елемент тексту, однак вважає його надто відмінним від аналізованого. Деякі явища зустрічаються досить рідко і тому моделі не вистачає тренувальних даних аби побудувати чіткі зв'язки та навчитись правильно їх розпізнавати: елементи можуть зустрічатися у відмінних контекстах та значеннях, бути частиною більших семантичних конструкцій, в межах яких виконують іншу роль. Значною проблемою є гумористичні відгуки та відгуки, повністю створенні "автозаповненням", що не відображають ставлення автора відгуку до додатку та не передають його емоцій, однак часто можуть бути наповнені словами з сильним сентимент-зарядом, розрізнення чого досить важко закласти у модель.

2.2 Словникові методи сентимент аналізу

Словниковий метод показав дуже низькі результати, а саме 14.2% як найвище значення за зваженою збалансованою метрикою, що на 35.8% менше, аніж випадкове вгадування(50%). Основною причиною цього є те, що тональний

словник lang uk, на матеріалі якого було створено словникову модель, містить всього 3442 розмічених слова, котрих аж ніяк не вистачає для покриття більш ніж 18000 відгуків, котрі, до того ж мають дуже різнобарвне словникове наповнення, через що більшість класифікацій приходилась на нейтральний сентимент, адже коли модель не могла розпізнати слово то вважала його нейтрально зарядженим. Помітно виділяється показник метрики точності для 1 класу сентименту(сильний негатив) - 87.5% та 90.9%, що є логічним через високе насичення відгуків цього класу словами-маркерами, такими як “жах” і т.п. Також варто відмітити тенденцію переважання як показників метрик як точності, так і відтворення для класів 1, 3 та 5 над цими ж показниками для класів 2 та 4, що зазвичай не виражені яскравими словами-маркерами, або ж утворюються досить складними їх поєднаннями, така тенденція зберігається і для інших моделей.Спроба доєднати до словникової моделі врахування статистичного показнику TF-IDF призвела до невеликої втрати за майже всіма метриками. Окремі таблиці та матриці похибок див. у Додатку 2.

2.3 Машинне навчання

2.3.1 Контрольоване машинне навчання

Контрольоване машинне навчання показало досить високі результати. Найбільш ефективною за зваженою збалансованою метрикою з методів машинного навчання виявилась модель векторної класифікації SVC_combo - 76.5%, одразу за нею модель логістичної регресії у поєднанні зі SMOTE smote_logreg_combo - 76.4% та модель логістичної регресії logreg_combo - 76.2%, що відповідає припущенню, сформованому у теоретичній частині дослідження. Модель на основі Наївного Баєса показала помітно менший результат - 66.4%, припущення щодо високої оптимальності цього підходу не цілком виправдалося. Моделі також показали дуже хорошу швидкість обробки даних, для більшості моделей вона склала менш ніж 10 секунд для повної класифікації тестового датасету, для найбільш результативної ж - 32 секунди. Майже для всіх моделей спостерігається тенденція,

аналогічна до тенденції словникової моделі, показники метрик як точності, так і відтворення для класів 1, 3 та 5 над цими ж показниками для класів 2 та 4, а клас 5 є найбільш домінантним та не рідко переважає приблизно в 1.5 рази над усіма іншими, що спричинено переважно помітно більшою кількістю текстів що відповідають цьому класу у тренувальному датасеті та висока насиченість текстів цього класу словами-маркерами. Власне, подібну особливість мають і показники для класу 1, які не рідко є другими за величиною, кількість текстів цього класу також є більшою в тренувальному датасеті і йому також властиві слова-маркери. Окремі таблиці та матриці похибок див. у Додатку 3.

2.3.2 Неконтрольоване машинне навчання

Особливості показника оцінки

Найчаст...
Найменш ча...

ЗНАЧЕННЯ	ЧАСТОТА
5	12 500
1	3 090
4	1 816
3	861
2	535

Рис. 2.40 Розподіл класів за оцінкою користувача

Як змінився датасет

характеристика	експертний індекс	оцінка користувача	коментар
документів у тесті	3 760	3 761	практично той самий обсяг
частка класу 5	**43 %**	**66 %**	користувачі схильні ставити «5» набагато частіше
клас 2 у тесті	453 (12 %)	107 (3 %)	різко зменшує матеріал для навчання/тесту на «скоріше негативних»
клас 3	576 (15 %)	172 (4 %)	«нейтральних» теж значно менше
клас 1	565 (15 %)	618 (16 %)	майже не змінюється

Табл. 2.1. Зміни в датасеті за оцінкою користувача та їх опис.

Можна зробити висновок, що перехід до користувацьких оцінок робить вибірку незбалансованою (2-/3-/4-зіркові майже зникають) і потенційно шумною: тональність тексту часто не узгоджується з «лайк-рейтингом».

2.3.3 Порівняння результатів неконтрольованого машинного навчання відносно даних, використаних для контрольованого машинного навчання

модель	expert F1_macro	user F1_macro	Δ F1	expert acc	user acc	Δ acc
SVC_COMBO	0.697	0.386	−0.311	0.751	0.668	−0.083
LOGREG_COMBO	0.695	0.361	−0.334	0.743	0.648	−0.095
SMOTE_LOGREG_COMBO	0.692	0.365	−0.327	0.745	0.630	−0.115
RF_COMBO	0.669	0.314	−0.355	0.729	0.715	−0.014
GB_COMBO	0.625	0.299	−0.326	0.695	0.717	+0.022
WORD_ONLY_LOGREG	0.637	0.321	−0.316	0.686	0.529	−0.157
NB_COMBO	0.593	0.324	−0.269	0.676	0.733	+0.057
DT_COMBO	0.582	0.320	−0.262	0.650	0.608	−0.042

KNN_COMBO	0.445	0.270	−0.175	0.506	0.688	+0.182

Табл. 2.2. Порівняння результатів неконтрольованого машинного навчання відносно даних, використаних для контрольованого машинного навчання(expert - індекс емотивності, user - оцінка користувача).

Спостереження

1. Усі лінійні моделі (SVC, LogReg) втрачають≈0.30 F1, бо робоча гіперплощина навчалась на рідкісні класи, а у «user» їх майже немає і похибка збільшується.

2. Дерева/ансамблі (RF, GB) майже не просіли за accuracy, бо просто «навчилися» віддавати клас 5; їх macro F1 падає, але менше, ніж у лінійних.

3. KNN та NB показують навіть вищий accuracy, що оманливо: вони теж здебільшого прогнозують 5 (див. support-зважену метрику).

4. Macro F1 залишається низьким у всіх: дисбаланс і шум руйнують здатність відрізняти класи 2, 3 і 4.

клас	recall-експерт	recall-user	причини та наслідки
1	≈0.67	≈0.55–0.64	менше «помітних» негативних маркерів у текстах; часто ставлять 1

			зірку за сервіс, але пишуть нейтрально
2	≈0.59	≤0.12	критичний дефіцит прикладів (107 рядків) , тому модель не бачить патернів
3	≈0.72	≤0.19	колапс класу 3 у рейтингах: коментарі «нормально» отримують 4-5 зірок
4	≈0.70	≤0.40	текст емоційно схожий на 5, але рейтинг нижчий що призводить до створення зони плутанини
5	≈0.86	≈0.74–0.96	recall трохи зростає в дерев/ансамблів, бо клас став ще домінантнішим

Табл. 2.3. Аспектний аналіз розподілу показників метрик моделей контрольованого машинного навчання за класами.

Чому вийшло саме так?

фактор	вплив на результати
суб'єктивні оцінки користувача	можемо мати позитивно сформульований текст із низькою оцінкою («доставка хороша, але ставлю 1 — не повернули гроші»); лінійні моделі трактують текст буквально й помиляються.

експоненційно більший дисбаланс (66 % клас 5)	accuracy гірше відображає якість; macro F1 різко падає, бо класи 2 і 3 рідкісні.
погіршення доповнення char-грамами	char-грами допомагали ловити орфографічні експресиви для індексу, але це не рятує, коли рейтинг не корелює з формою повідомлення.

2.3.4 Тренування на оцінці користувача, тестування за індексом

показник	значення	інтерпретація
середня абсолютна різниця `	оцінка – індекс	
частка повного збігу	57,63 %	лише трохи більше половини відгуків мають «чесний» рейтинг
⌀ позитивне зміщення (оцінка > індекс)	+0,6197	4,6× перевищує ⌀ негативне (–0,1335) → користувачі завищують оцінки набагато частіше, ніж занижують
макс. відхилення по додатках	0,997 (≈20 %)	крайній випадок «завищених» п’ятірок (com.outfit7.mytalkingtom2)
мін. відхилення	0,508 (≈10 %)	найпослідовніші оцінки (com.noodlecake.altosodyssey)

розподіл різниць	Δ = 0 → 57,63 % Δ = 1 → 28,20 % Δ = 2 → 14,21 % Δ = 3 → 3,93 % Δ = 4 → 2,38 %	що більша невідповідність, то рідше вона трапляється

Табл. 2.4. Порівняння оцінки користувача з присвоєним індексом. Оцінка користувача є систематично упередженою в бік «зайвого» позитиву; тому використовувати її показник для тональності ризиковано.

модель (combo TF-IDF)	навчання → тест	macro F1	accuracy
SVC (expert → expert)	індекс → індекс	0,697	0,751
LogReg (expert → expert)	індекс → індекс	0,695	0,743
SVC (user → expert)	оцінка → індекс	0,392	0,553
LogReg (user → expert)	оцінка → індекс	0,389	0,551

Табл. 2.5. Порівняння з моделями, які вчились одразу на індексі.

Падіння продуктивності при перетині задач:

- Δ macro F1 ≈ –0,30 (-43 %)
- Δ accuracy ≈ –0,09

Навіть найкращі лінійні моделі, підготовлені на рейтингах, утрачають близько 40 % узагальненої якості, коли їх просять вгадати експертне забарвлення.

клас	recall експерт-модель	recall user-модель	зміна	пояснення
1	0,673	0,555	–0,118	негативні оцінки користувача часто супроводжуються «помірно негативним» текстом → модель «надмірно» добирає м'які формулювання
2	0,592	0,361	–0,231	клас майже зник у рейтингах (support ≈ 3 %)
3	0,733	0,029	–0,704	«нейтральні» тексти здебільшого отримують 4–5 зірок; модель їх не бачила
4	0,709	0,239	–0,470	семантично близький до 5, але рідкісний у рейтингах

5	0,825	0,879	+0,054	домінуюча п'ятірка у навчанні → гіперплощина зсувається на користь класу 5

Табл. 2.6. Порівняння recall з моделями, які вчились одразу на індексі

причина	як проявляється в метриках
дисбаланс класів у "оцінці" (66 % 5-ок)	precision/recall класу 5 → ↑, але recall 2–4 → ↓; macro F1 падає
позиційний шум (рейтинг ≠ тон)	середня різниця 0,75 бала; ≈ 28 % прикладів мають зміщення на 1 пункт → модель вчиться зміщувати гіперплощину
асиметрія позитив/негатив	позитивне зміщення у 4,6 раза сильніше → передбачення 1–2 гірші, ніж 4
невелика вибірка для класів 2, 3 і 4	кількасот прикладів недостатньо, аби «закрити» мовне різноманіття; SMOTE частково допомагає класу 1, але не відновлює 2-3
char-грами + TF-IDF	фіксують орфографічні емоційні маркери, однак їх частотність

	корелює з текстом, а не з «завищеною» оцінкою → похибка збільшується

Табл. 2.7. Порівняння результатів роботи моделей та ймовірні причини їх виникнення

Можна зробити наступні висновки:

1. Тональність ≠ рейтинг. Середня помилка ≈ 15 % довжини шкали; для деяких додатків — до 20 %.

2. Навчання на оцінці не переноситься на індекс. Втрачаємо ~0,30 macro F1; «сліпі зони» класів 2–4.

3. Краще тренувати безпосередньо на анотованому індексі або проводити двоступеневе навчання (перше — вирівнювання рейтингу до «очищеного» індексу, друге — тональність).

4. Необхідні техніки зниження шуму: ручна переоцінка найбільш розбіжних відгуків, використання weak-supervision (Snorkel, co-training), або тренування трансформерів із домішуванням «емоційних» задач.

Через систематичне завищення «зірочок» і сильний дисбаланс моделі, навчені на рейтингах, погано відтворюють справжній сентимент. Щоб дійсно «бачити» емоційну тональність тексту, потрібна або ручна анотація, або складні багаторівневі стратегії очистки та балансування.

модель	expert → expert		user → user		user → expert	
	accuracy	macro F1	accuracy	macro F1	accuracy	macro F1
SVC_CO	0.751	0.697	0.668	0.386	0.553	0.367

MBO						
LOGREG_COMBO	0.743	0.695	0.648	0.361	0.551	0.389
SMOTE_LOGREG_COMBO	0.745	0.692	0.630	0.365	0.587	0.425
WORD_ONLY_LOGREG	0.686	0.637	0.529	0.321	0.502	0.288
NB_COMBO	0.676	0.593	0.733	0.324	0.537	0.275
RF_COMBO	0.729	0.669	0.715	0.314	0.517	0.268
GB_COMBO	0.695	0.625	0.717	0.299	0.504	0.249
DT_COMBO	0.650	0.582	0.608	0.320	0.491	0.320
KNN_COMBO	0.506	0.445	0.688	0.270	0.475	0.220

Табл. 2.8. Зведена порівняльна таблиця**.** Окремі таблиці та матриці похибок див. у Додатку 4.

Цікавим є феномен зростання macro F1 для LogReg_COMBO і SMOTE_LogReg_COMBO у сценарії user → expert. Попри те, що модель була натренована на користувацьких оцінках, вона виявилася кращою у прогнозуванні незалежної експертної

оцінки, ніж у передбаченні самих оцінок користувача. Причини виникнення цього парадоксу:

Причина 1: оцінка користувача — шумна ціль, індекс — якісна ціль. Навчаючи модель на користувацькій оцінці, ми по суті подаємо їй зміст коментар як мітку — власну інтерпретацію користувача, яка дуже часто не відповідає емоційному змісту тексту. За даними з дослідження, середня похибка = 0.7532 бала (~15 %), і лише 57.6 % оцінок точно збігаються з емоційним забарвленням. Отже, модель не може навчити якісну відповідність "коментар → рейтинг", бо зв'язок слабкий і зміщений,

але структура самого тексту все одно містить сентимент, тому коли ми тестуємо цю модель за нормальною (експертною) ціллю, вона демонструє кращу поведінку. Тобто модель не може добре вгадати, яку оцінку ставить користувач (бо вона непослідовна), але може вгадати реальну тональність, бо саме її текст і передає.

Причина 2: індекс має більш «розпізнавану» внутрішню логіку. Експертна оцінка була проставлена згідно з реальним емоційним змістом (тональність, полярність, інтенсивність), має більш рівномірний розподіл класів, і не містить суб'єктивних впливів (як-от: "все ок, але ставлю 1 бо…"). При тестуванні на індексі міноритарні класи (2, 3 і 4) вже мають суттєвий обсяг у тесті (на відміну від user → user), навіть помірна здатність моделі вгадувати сентиментальні

відтінки, які вона «вловила» з тексту, починає приносити макро-нагороду.

Причина 3: macro F1 як метрика винагороджує "справедливий розподіл" більше, ніж абсолютну точність. В user → user більшість моделей гадали тільки клас 5 — це дає високий accuracy, але низький macro F1. В user → expert, навіть часткове вгадування 2, 3 і 4 приносить більше балів macro F1, бо метрика не зважена по класах. Logistic Regression як проста, «загальна» модель вловлює загальні риси, тому її нейтральна поведінка (трохи для всіх) дає кращий macro F1 саме в user → expert.

У результаті модель, яка вчилась на непослідовній оцінці, змогла випадково або частково навчити реальну сентимент-модель, яка працює краще за свою ж роботу по рейтингах. Це — сильний аргумент на користь анотування або очищення оцінок, і одночасно — показник того, що користувацька оцінка усе ж “несе” справжнє відображення сентименту, незалежно від зашумлення рейтингу, і його можна відновити.

Class	Precision	Recall	F1-score	Precision (TF-IDF)	Recall (TF-IDF)	F1-score (TF-IDF)	Δ Precision	Δ Recall	Δ F1-score
1	0.875	0.036	0.070	0.909	0.035	0.067	0.034	-0.001	-0.003
2	0.115	0.025	0.041	0.182	0.036	0.060	0.067	0.011	0.019
3	0.171	0.958	0.290	0.179	0.949	0.301	0.008	-0.009	0.011

4	0.228	0.175	0.198	0.183	0.209	0.195	-0.045	0.034	-0.003
5	0.825	0.069	0.127	0.772	0.048	0.090	-0.053	-0.021	-0.037
accuracy	0.204	0.204	0.204	0.200	0.200	0.200	-0.004	-0.004	-0.004
macro avg	0.443	0.253	0.145	0.445	0.255	0.142	0.002	0.002	-0.003
weighted avg	0.566	0.204	0.142	0.551	0.200	0.129	-0.015	-0.004	-0.013

Табл. 2.9 Порівняння словникової моделі на основі словнику тональності lang uk без використання TF-IDF та з його використанням.

2.4 Глибоке машинне навчання. Донавчання та налаштування гіперпараметрів донавчання BERT

Донавчання було проведене на основі претренованої моделі bert-base-multilingual-uncased, котра у своєму вихідному варіанті показала ефективність 53.1%, що є досить поганим результатом, однак модель показала дуже хороший час обробки тестового датасету, всього 9 секунд, що, однак, не є важливим при такій низькі ефективності. Донавчена за стандартною конфігурацією донавчання версія цієї ж моделі показала результат в 66.3%, тобто покращення аж на 10 відсотків, що дуже суттєво.

В результаті численних експериментів з налаштуванням гіперпараметрів донавчання BERT досягнути покращення ефективності моделі за допомогою цього так і не вдалось, загалом різні поєднання значень параметрів приводили до відносно незначного(менше ніж 1%) перерозподілу значень між метриками та класами, зазвичай у вигляді зменшення зваженої F1 та збільшення показників інших метрик. Значення ключа генератору випадковості було залишене без змін. Отже,

налаштуванням гіперпараметрів донавчання BERT є досить своєрідним інструментом, котрий дозволяє впливати на балансування визначення сентименту за метриками й класами без зміни тестових або тренувальних даних, однак його ефективне використання потребує глибшого вивчення та додаткових експериментів.

2.5 Великі мовні моделі в контексті сентимент-аналізу

2.5.1 Zero-shot тестування

ChatGPT. Модель gpt-3.5-turbo показала досить хороший результат за зваженою збалансованою метрикою - 67.3%. Збереглася характерна динаміка розподілу показників збалансованої метрики, що була присутня у моделях машинного навчання та словникових моделях, класи 2 і 4 досягли одних із найменших значень серед усіх моделей - 40.1% та 40.5%, що в обох випадках майже на 10% гірше ніж випадкове вгадування. Повна класифікація тестового датасету зайняла 7 хвилин 24 секунди, що значно більше за моделі машинного навчання.

Модель gpt-4.1 яка є більш об’ємною та потужною, показала помітно кращий результат, майже рівний кращим результатам моделей машинного навчання за зваженою збалансованою метрикою - 76%. Варто відзначити, що при такому зростанні загального показнику зваженої збалансованої метрики найбільшими є прирости показників збалансованої метрики саме для класів 2 і 4, попри які все одно зберігається характерна динаміка розподілу. Повна класифікація тестового датасету зайняла 30 хвилин 46 секунд, що значно більше за моделі машинного навчання і більш ніж в 4 рази більше за gpt-3.5-turbo, однак покращення зваженої збалансованої метрики виправдовує цю затримку.

“Міркуюча” модель gpt-o3 показала результати трохи гірші за gpt-4.1 та найкращі моделі машинного навчання за зваженою збалансованою метрикою - 75%. Хоч показник зваженої збалансованої метрики й знизився на 1 відсоток, проте

помітно зросли показники зваженої метрики для класів 2 та 4. Характерна динаміка розподілу збереглась. Повна класифікація тестового датасету зайняла 1 годину 56 хвилин 58 секунд., що є надзвичайно довго порівняно з усіма іншими випробуваними моделями й фактично робить gpt-o3 непридатною в умовах, коли час обробки даних є важливою складовою роботи моделі.

Gpt-5.5 показала результат в 76.6 за зваженою збалансованою метрикою, а також суттєво зменшила час обробки датасету порівняно зі старішими моделями, виконавши всі запити за 6 хвилин 56 секунд, що суттєво швидше. Характерна динаміка розподілу за класами збереглась.

DeepSeek. Модель deepseek-v3 показала досить непоганий результат, що трохи поступається кращим результатам моделей машинного навчання за зваженою збалансованою метрикою - 72.6%, що помітно краще за gpt-3.5-turbo, проте гірше за gpt-4.1. Зберігається характерна динаміка розподілу за збалансованою метрикою. Повна класифікація тестового датасету зайняла 9 хвилин 34 секунди., що значно більше за моделі машинного навчання, на 2 хвилини більше за gpt-3.5-turbo і більш ніж в 3 рази менше за gpt-4.1.

“Міркуюча” модель deepseek-r1 показала досить хороший результат, майже рівний кращим результатам моделей машинного навчання за зваженою збалансованою метрикою - 75.5%, що помітно краще за gpt-3.5-turbo та на 0.5% краще за gpt-o3 - її прямий конкурент, проте на 0.5% гірше за gpt-4.1. Зберігається характерна динаміка розподілу за збалансованою метрикою. Повна класифікація тестового датасету зайняла 44 хвилини 44 секунди, що значно більше за моделі машинного навчання, майже в 6 разів більше за gpt-3.5-turbo і більш на 10 хвилин більше за gpt-4.1, однак більш ніж в 2.5 рази менше за gpt-o3.

Модель DeepSeek v4 flash показала ефективність 72.5% та час обробки 8 хв 31 секунда, що є майже повним відповідником результатам старішої моделі deepseek-v3 з покращенням часу обробки всього на одну хвилину. Модель DeepSeek v4 pro показала ефективність 62.1%, що є найгіршим результатом з випробуваних

великих мовних моделей та час обробки 7 хв 22 секунди. Отже, новіші моделі DeepSeek не показали помітного покращення результатів, а навіть значне їх погіршення для DeepSeek v4 pro.

Показник	GPT-3.5-turbo	GPT-4.1	GPT-o3	GPT-5.5	DeepSeek-v3	DeepSeek-r1	DeepSeek v4 flash	DeepSeek v4 pro	Claude 4.8
Зважена F1 у %	67.3	76.0	75.0	76.6	72.6	75.5	72.5	62.1	73.5

Табл. 2.10. Порівняльна таблиця зваженої показників збалансованої метрики для великих мовних моделей. Окремі таблиці та матриці похибок див. у Додатку 5.

Показник	GPT-3.5-turbo	GPT-4.1	GPT-o3	GPT-5.5	DeepSeek-v3	DeepSeek-r1	Deepseek v4 flash	Deepseek v4 pro	Claude 4.8
Час год:хв:сек	07:24	30:46	01:56:58	6:56	09:34	44:44	8:31	7:22	1:16:04

Табл. 2.11 Порівняльна таблиця часу класифікації тестового датасету в повному об’ємі для великих мовних моделей. Окремі результати й матриці похибок див. у Додатку 5.

Хоча модель o3 виступає більш складною та здібною до об’ємних задач у неї є серйозні недоліки: фіксований параметр температури 4.0 оскільки модель зроблена з чіткою задачею симулювати мисленнєві здібності людини й підходити до задач більш варіативно аби забезпечити кращий результат для більш об’ємних запитів та задач. Помітно довший відносно всіх інших моделей час відповіді. Попри вторинність економічного аспекту не можна не відмітити вартість використання

цієї моделі: тестування за 3761 відгуком, наведене в дослідженні вартувало 20,549$, або ж
854,26₴, що значно більше за gpt4.1, яка показала навіть трохи кращі результати, та майже в 6 разів більше за пряму альтернативу - DeepseekR1, повна вартість тестування якого склала 3.5$ або ж 145,50₴, що можливо зменшити ще в 3 рази якщо використовувати модель в період зниженого попиту. DeepseekR1 також показав кращий з великих мовних моделей результат за зваженою збалансованою метрикою.

Рис. 2.41, 2.42, 2.43 вартість одного тестування на повному тестовому датасеті моделей ChatGPT o3 та моделей DeepSeek.

Claude.Модель Claude Opus 4.8 показала ефективність в 73.5%, що є хорошим результатом, однак використання моделі вийшло доволі дорогим - 2.32$, що у 116 разів дорожче ніж DeepSeek v4 flash - 0.02$ та довгим - 1 година 16 хвилин, що робить Claude досить неоптимальним для виконання задач автоматичного сентимент-аналізу україномовних відгуків Google Play порівняно з іншими моделями.

```
=== Classification report (zero-shot, claude-opus-4-8) ===
              precision    recall  f1-score   support

           1      0.811     0.588     0.681       577
           2      0.519     0.519     0.519       443
           3      0.550     0.896     0.682       550
           4      0.554     0.656     0.601       550
           5      0.967     0.797     0.874      1641

    accuracy                          0.726      3761
   macro avg      0.680     0.691     0.671      3761
weighted avg      0.769     0.726     0.735      3761


=== Confusion matrix ===
[[ 339  196   39    3    0]
 [  60  230  147    5    1]
 [   7    8  493   34    8]
 [   0    2  152  361   35]
 [  12    7   65  249 1308]]
```

Рис. 2.44 Розподілення передбачень за метриками, класами та матрицею похибок для Claude Opus 4.8

2.5.2 Вплив мови запиту на великі мовні моделі

Zero-shot

gpt-3.5

Class	Precision (eng)	Recall (eng)	F1-score (eng)	Precision (ua)	Recall (ua)	F1-score (ua)	Δ Precision	Δ Recall	Δ F1-score
1	60,8	77,3	68	67,5	57,7	62,2	6,7	-19,6	-5,8
2	39,1	41,1	40,1	35,2	70,2	46,9	-3,9	29,1	6,8
3	60,3	59,8	60	65,7	41,1	50,6	5,4	-18,7	-9,4

4	53,1	32,7	40,5	45,3	44,7	45	-7,8	12	4,5
5	84,9	86,7	85,8	89	81,2	85	4,1	-5,5	-0,8
accuracy	68,1	68,1	68,1	65,1	65,1	65,1	-3	-3	-3
macro avg	59,6	59,5	58,9	60,6	59	57,9	1	-0,5	-1
weighted avg	67,5	68,1	67,3	69,6	65,1	66,1	2,1	-3	-1,2

Табл. 2.12. Порівняння результатів тестування моделі gpt-3.5-turbo-0125 з використанням запитів англійською та українською у відсотках. Окремі таблиці та матриці похибок див. у Додатку 7.

gpt-4.1

Class	Precision (eng)	Recall (eng)	F1-score (eng)	Precision (ua)	Recall (ua)	F1-score (ua)	Δ Precision	Δ Recall	Δ F1-score
1	76	68,1	71,8	74,4	80,8	77,5	-1,6	12,7	5,7
2	52,5	49,4	50,9	59,6	42,9	49,9	7,1	-6,5	-1
3	63	81,6	71,1	61	85,5	71,2	-2	3,9	0,1
4	64,6	57,6	60,9	63	56,9	59,8	-1,6	-0,7	-1,1
5	91,3	90,2	90,7	93	87,8	90,3	1,7	-2,4	-0,4
accuracy	76	76	76	76,5	76,5	76,5	0,5	0,5	0,5

macro avg	69,5	69,4	69,1	70,2	70,8	69,7	0,7	1,4	0,6
weighted avg	76,3	76	75,9	77,1	76,5	76,3	0,8	0,5	0,4

Табл. 2.13. Порівняння результатів тестування донавченою на 1000 відгуків моделі gpt-4.1-2025-04-14 з використанням запитів англійською та українською у відсотках. Окремі таблиці та матриці похибок див. у Додатку 7.

Fine-tuned

Class	Precision (eng)	Recall (eng)	F1-score (eng	Precision (ua)	Recall (ua)	F1-score (ua)	Δ Precision	Δ Recall	Δ F1-score
1	81,9	80,2	81,1	83,2	71,9	77,1	1,3	-8,3	-4
2	56,3	72,7	63,4	52,5	79,7	63,3	-3,8	7	-0,1
3	70,8	73,6	72,2	75,6	66	70,5	4,8	-7,6	-1,7
4	69,9	64,5	67,1	58,3	73,5	65	-11,6	9	-2,1
5	94,8	89,2	91,9	96,4	83,2	89,3	1,6	-6	-2,6
accuracy	80	80	80	77,1	77,1	77,1	-2,9	-2,9	-2,9
macro avg	74,7	76,1	75,2	73,2	74,8	73	-1,5	-1,3	-2,2
weighted avg	81,1	80	80,4	80,6	77,1	78,1	-0,5	-2,9	-2,3

Табл. 2.14. Порівняння результатів тестування донавченою на 1000 відгуків моделі gpt-3.5-turbo-0125 з використанням запитів англійською та українською у відсотках. Окремі таблиці та матриці похибок див. у Додатку 7.

Для донавченої на 1000 відгуків моделі gpt-3.5-turbo-0125 зміна запиту до моделі з написаного англійською на ідентичний, написаний українською призвела до середньої втрати точності моделі в 2.3% та виявилася загалом негативною для всіх класів сентименту, навіть у зростання показнику точності або відтворення, втрата в іншому з цих двох показників призводила до зниження збалансованої метрики.

Class	Precision (eng)	Recall (eng)	F1-score (eng)	Precision (ua)	Recall (ua)	F1-score (ua)	Δ Precision	Δ Recall	Δ F1-score
1	79,6	92,2	85,5	75,9	94,3	84,1	-3,7	2,1	-1,4
2	71,1	65,7	68,3	69,5	58,7	63,6	-1,6	-7	-4,7
3	75,4	75,1	75,2	73,3	76,4	74,8	-2,1	1,3	-0,4
4	73	72,7	72,9	73,5	72,2	72,8	0,5	-0,5	-0,1
5	94,7	91,7	93,2	95,4	90,6	92,9	0,7	-1,1	-0,3
accuracy	83,5	83,5	83,5	82,6	82,6	82,6	-0,9	-0,9	-0,9
macro avg	78,8	79,5	79	77,5	78,4	77,7	-1,3	-1,1	-1,3
weighted avg	83,6	83,5	83,5	82,9	82,6	82,5	-0,7	-0,9	-1

Табл. 2.15. Порівняння результатів тестування донавченою на 1000 відгуків моделі gpt-4.1-2025-04-14 з використанням запитів англійською та українською у відсотках. Окремі таблиці та матриці похибок див. у Додатку 7.

Для донавченої на 1000 відгуків моделі gpt-4.1-2025-04-14 зміна запиту до моделі з написаного англійською на ідентичний, написаний українською призвела до

середньої втрати точності моделі в 1% та виявилася загалом негативно для всіх класів сентименту, навіть при зростанні показнику точності або відтворення, втрата в іншому з цих двох показників призводила до зниження збалансованої метрики.

Class	ΔΔ Precision	ΔΔ Recall	ΔΔ F1-score
1	-5	10,4	2,6
2	2,2	-14	-4,6
3	-6,9	8,9	1,3
4	12,1	-9,5	2
5	-0,9	4,9	2,3
accuracy	2	2	2
macro avg	0,2	0,2	0,9
weighted avg	-0,2	2	1,3

Табл. 2.**16.** Зміни різниці між запитом англійською та українською мовою для донавченої gpt-4.1-2025-04-14 порівняно з донавченою gpt-3.5-turbo-0125 у відсотках. Окремі таблиці та матриці похибок див. у Додатку 7.

З обчислень в таблиці видно, що середня зважена метрика F1 при використанні запиту українською мовою покращилась на 1,3%, що дає підстави для припущення про виправлення недоліку якості відповіді моделі у зв’язку з мовою при покращенні самої моделі та виходу нових її версій, або ж більшу піддатливість донавчанню.

deepseek-chat

Class	**Precision**	**Recall**	**F1-score**	**Precisi**	**Reca**	**F1-**	**Δ**	**Δ**	**Δ F1-**

	(eng)	**(eng)**	**(eng)**	**on (ua)**	**ll (ua)**	**score (ua)**	**Precision**	**Recall**	**score**
1	72,6	62,4	67,1	69,9	66	67,9	-2,7	3,6	0,8
2	49,6	58,5	53,7	52,8	63,9	57,8	3,2	5,4	4,1
3	56,7	71,8	63,4	56,5	69,6	62,4	-0,2	-2,2	-1
4	54,9	64,9	59,5	55,6	60,5	58	0,7	-4,4	-1,5
5	94,8	80,6	87,1	94,4	80,7	87,1	-0,4	0,1	0
accuracy	71,6	71,6	71,6	71,9	71,9	71,9	0,3	0,3	0,3
macro avg	65,7	67,6	66,2	65,8	68,2	66,6	0,1	0,6	0,4
weighted avg	74,6	71,6	72,6	74,5	71,9	72,8	-0,1	0,3	0,2

Табл. 2.17. Порівняння результатів тестування моделі deeepseek-chat з використанням запитів англійською та українською у відсотках. Окремі таблиці та матриці похибок див. у Додатку 7.

З обчислень в таблиці видно, що середні значення метрик для запитів англійською та українською відрізняються лиш на долю відсотка, що зазвичай не перевищує 0.3%, що в межах допустимої похибки при використанні великої мовної моделі тобто функційної різниці між ними при використанні моделі deepseek-chat, що є досить логічним, оскільки для deepseek первинною мовою є китайська, а українська та англійська є вторинними, на відміну від ChatGPT, для якого первинною є англійська.

deepseek-reasoner

Class	Precision (eng)	Recall (eng)	F1-score (eng)	Precision (ua)	Recall (ua)	F1-score (ua)	Δ Precision	Δ Recall	Δ F1-score
1	83	63,3	71,8	82,8	62,6	71,3	-0,2	-0,7	-0,5
2	52	67,5	58,7	52,4	70,4	60,1	0,4	2,9	1,4
3	64	73,1	68,3	65,2	72,5	68,7	1,2	-0,6	0,4
4	58,4	66,5	62,2	59,1	60,5	59,8	0,7	-6	-2,4
5	92,7	84,2	88,3	91,1	86,2	88,6	-1,6	2	0,3
accuracy	74,8	74,8	74,8	75	75	75	0,2	0,2	0,2
macro avg	70	70,9	69,8	58,4	58,7	58,1	-11,6	-12,2	-11,7
weighted avg	77,2	74,8	75,5	76,8	75	75,5	-0,4	0,2	0

Табл. 2.18. Порівняння результатів тестування моделі deeepseek-reasoner з використанням запитів англійською та українською у відсотках. Окремі таблиці та матриці похибок див. у Додатку 7.

2.6 Донавчання великих мовних моделей (попередньо треновані трансформери сімейства ChatGPT)

У цьому підрозділі розглянуто донавчання, або ж “файнтюнинг” попередньо розглянутих в аспекті zero-shot класифікації моделей GPT gpt-3.5-turbo-0125 **та** gpt-4.1-2025-04-14 за допомогою внутрішнього інструментарію OpenAI. Для кожної моделі було створено дві донавчені моделі за допомогою файн-тюн датасету на 100 та 1000 відгуків, які в подальшому з метою полегшення сприйняття матеріалу будуть вказані за наданим їм суфіксом як “-sentiment” та “-sentiment2” відповідно.

Попри те що модель o3 хоч і вважається однією з найкращих великих мовних моделей на момент проведення дослідження, вона поки що не підтримує донавчання.

Донавчання на 100 відгуках

Class	Precision	Recall	F1-score	Precision (fine-tuned)	Recall (fine-tuned)	F1-score (fine-tuned)	Δ Precision	Δ Recall	Δ F1-score
1	60.8	77.3	68.0	77.3	66.7	71.6	16.5	**-10.6**	3.6
2	39.1	41.1	40.1	44.8	72.2	**55.3**	5.7	31.1	15.2
3	60.3	59.8	60.0	68.1	41.5	**51.5**	7.8	**-18.3**	**-8.5**
4	53.1	32.7	40.5	58.7	62.9	**60.8**	5.6	30.2	20.3
5	84.9	86.7	85.8	90.7	89.8	90.2	5.8	3.1	4.4
accuracy	68.1	68.1	68.1	73.2	73.2	73.2	5.1	5.1	5.1
macro avg	59.6	59.5	58.9	67.9	66.6	65.9	8.3	7.1	7.0
weighted avg	67.5	68.1	67.3	75.3	73.2	73.3	7.8	5.1	6.0

Табл. 2.19. Порівняльна таблиця результатів випробування gpt-3.5-turbo та gpt-3.5-turbo-sentiment у відсотках. Окремі таблиці та матриці похибок див. у Додатку 6.

Завдяки донавчанню показники за метриками помітно зросли в середньому на 5.1-8.3%, однак відбулась досить значна втрата показнику збалансованої

метрики для класу 3 - 8.5%, а також на 10.6% зменшився показник відтворення для класу 1(приріст збалансованої метрики для цього класу залишився позитивним). Характерна динаміка розподілу показників збалансованої метрики за класами не збереглася, клас 3 став на 3.8% меншим за клас 2 та на 9.3% меншим за клас 4, що сильно відрізняється від звичайного розподілу.

Class	Precision	Recall	F1-score	Precision (fine-tuned)	Recall (fine-tuned)	F1-score (fine-tuned)	Δ Precision	Δ Recall	Δ F1-score
1	79.3	65.2	71.6	84.3	71.4	77.3	5.0	6.2	5.7
2	52.9	53.7	53.3	57.9	66.4	61.8	5.0	12.7	8.5
3	63.0	83.3	71.7	64.0	81.6	71.7	1.0	**-1.7**	**0.0**
4	63.7	56.2	59.7	66.9	64.2	65.5	3.2	8.0	5.8
5	91.1	90.2	90.6	94.8	88.6	91.6	3.7	**-1.6**	1.0
accuracy	76.1	76.1	76.1	78.8	78.8	78.8	2.7	2.7	2.7
macro avg	70.0	69.7	69.4	73.5	74.4	73.6	3.5	4.7	4.2
weighted avg	76.7	76.1	76.0	80.2	78.8	**79.2**	3.5	2.7	3.2

Табл. 2.20. Порівняльна таблиця результатів випробування gpt-4.1 та gpt-4.1-sentiment у відсотках. Окремі таблиці та матриці похибок див. у Додатку 6.

Завдяки донавчанню показники за метриками помітно зросли в середньому на 2.7 -4.7%, що є досить незначним покращенням, невелике зменшення відбулося

лише за одним показником для двох класів, що знівелювало приріст показнику іншої метрики для класу 3 й залишило показник збалансованої метрики незмінним для цього класу. Показник зваженої збалансованої метрики зріс на 3.2% та максимальний показник отриманий при випробуванні методів машинного навчання. Характерна динаміка розподілу показників збалансованої метрики за класами збереглася.

Загалом донавчання на 100 відгуках дозволило досягти помітного покращення показників більшості метрик та досягти нової межі показнику зваженої збалансованої метрики, однак отримані результати все ж виявились дещо нестабільними, що може вказувати на недостатню кількість навчальних даних.

Донавчання на 1000 відгуків

Class	Precision	Recall	F1-score	Precision (FT)	Recall (FT)	F1-score (FT)	Δ Precision	Δ Recall	Δ F1-score
1	60.8	77.3	68.0	81.8	80.1	80.9	21.0	2.8	12.9
2	39.1	41.1	40.1	56.0	72.7	63.3	16.9	31.6	**23.2**
3	60.3	59.8	60.0	71.0	73.3	72.1	10.7	13.5	12.1
4	53.1	32.7	40.5	69.8	64.7	67.2	16.7	32.0	**26.7**
5	84.9	86.7	85.8	94.9	89.2	92.0	10.0	2.5	6.2
accuracy	68.1	68.1	68.1	80.0	80.0	80.0	11.9	11.9	11.9
macro avg	59.6	59.5	58.9	74.7	76.0	75.1	15.1	16.5	16.2
weighted avg	67.5	68.1	67.3	81.1	80.0	**80.4**	13.6	11.9	13.1

Табл. 2.21. Порівняльна таблиця результатів випробування gpt-3.5-turbo та gpt-3.5-turbo-sentiment2 у відсотках. Окремі таблиці та матриці похибок див. у Додатку 6.

Завдяки донавчанню абсолютно всі показники за метриками помітно зросли в середньому на 11.9-16.5%, що є дуже значним приростом. Найбільше зросли показники збалансованої метрики для класів 2 та 4 - 23.2% та 26.7% відповідно. Показник зваженої збалансованої метрики зріс на 13.1% і перевищив попередній найбільший, отриманий при випробуванні gpt-4.1-sentiment. Характерна динаміка розподілу показників збалансованої метрики за класами збереглася.

Class	Precision	Recall	F1-score	Precision (fine-tuned)	Recall (fine-tuned)	F1-score (fine-tuned)	Δ Precision	Δ Recall	Δ F1-score
1	79.3	65.2	71.6	79.0	92.4	85.1	**-0.3**	27.2	13.5
2	52.9	53.7	53.3	71.0	63.4	**67.0**	18.1	9.7	13.7
3	63.0	83.3	71.7	74.7	74.5	**74.6**	11.7	**-8.8**	2.9
4	63.7	56.2	59.7	72.4	72.9	**72.6**	8.7	16.7	12.9
5	91.1	90.2	90.6	94.8	91.7	93.2	3.7	1.5	2.6
accuracy	76.1	76.1	76.1	83.2	83.2	83.2	7.1	7.1	7.1
macro avg	70.0	69.7	69.4	78.4	79.0	78.5	8.4	9.3	9.1
weighted avg	76.7	76.1	76.0	83.3	83.2	**83.1**	6.6	7.1	7.1

Табл. 2.22. Порівняльна таблиця результатів випробування gpt-4.1 та gpt-4.1-sentiment2 у відсотках. Окремі таблиці та матриці похибок див. у Додатку 6.

Завдяки донавчанню показники за метриками помітно зросли в середньому на 6.6-9.3%, зменшення відбулося лише за одним показником для двох класів, що не змінило позитивну зміну збалансованої метрики для цих класів. Результат донавчання gpt-4.1 виявився найкращим з-поміж усіх випробуваних методів за зваженою збалансованою метрикою. Варто відмітити що саме для цієї моделі досягаються і найбільші показники збалансованої метрики для класів 2 та 4, котрі перевищують показники цих же класів у інших моделях де вони є одними з максимальних приблизно на 4 та 6 відсотків, саме цей фактор дозволив моделі обійти усі інші за показником зваженої збалансованої метрики. Характерна динаміка розподілу показників збалансованої метрики за класами збереглася лише частково, оскільки клас 4 відстає від класу 3 всього на 2% і клас 2 відстає від класу 3 на 7.6%, що досить мало порівняно зі звичайним розривом між ними.

Загалом донавчання на 1000 відгуках дозволило помітно посилити позитивну динаміку результатів, досягнутих при донавчанні на 100 відгуках. Помітно зменшились падіння показників метрик та кількість подібних випадків, збалансована метрика у всіх випадках лише зросла. Модель gpt-3.5-turbo в результаті донавчання проявила більше відносне зростання показників метрик, однак gpt-4.1 в результаті донавчання п вдалось досягти найбільших показників метрик, оскільки модель мала помітно кращі результати при zero-shot випробуванні.

Висновки до розділу 2

Найбільш ефективною за зваженою метрикою F1 виявилась донавчена велика мовна модель GPT4.1(gpt-4.1-2025-04-14:sentiment2) - 83.1%. Найбільш ефективною за зваженою метрикою з методів машинного навчання виявилась модель SVC_combo - 76.5%, яка хоч і значно поступається донавченій GPT4.1, проте має перевагу повної відкритості коду, оскільки його було написано в ході

виконання дослідження, швидкості використання, безкоштовності й відсутності залежності від стороннього серверу(технічних робіт, перебоїв і т.д.). Модель глибокого навчання показала високу швидкодію, однак значно поступається ефективністю іншим розглянутим моделям як сама по собі, так і з використанням донавчання. “Міркуючі” моделі показали хороші результати за зваженою метрикою, однак вийшли значно дорожчими у використанні та надзвичайно повільними у порівнянні з іншими методами, на даний момент в контексті автоматичного сентимент аналізу україномовних відгуків вони програють за всіма параметрами донавченим моделям. Зміна мови запиту з англійської на українську при використанні мовних моделей показала тенденцію погіршення результатів, однак новіші мовні моделі показали менший прояв цієї тенденції, що дає підстави стверджувати про усунення цієї проблеми з часом.

РОЗДІЛ 3. АНАЛІЗ ПОМИЛОК

У третьому розділі здійснено більш детальний аналіз конкретних допущених помилок, отриманих при використанні підходів, що зарекомендували себе як одні з найбільш ефективних, детального аналізу лінгвістичної природи їх справжнього сентимент-заряду, вірогідних причин виникнення помилок, подібності між підходами та теоретичних можливостей виправлення. Окремо виділено помилки при використанні великих мовних моделей, оскільки міра можливого впливу на них є сильно обмеженою порівняно з іншими методами через їх закриту архітектуру та претренованість моделей на надвеликих наборах даних широкого змісту.

3.1 Помилки при використанні машинного навчання

Відгуку «Немає Української мови!» було автоматично присвоєно індекс 2 - негативний. Відгук є сильно негативним оскільки негативно заряджений текст “Немає Української мови” додатково підсилюється окличним знаком, що утворює сильний негативний сентимент. Ця помилка спричинена тим, що велика кількість відгуків, які містять окличні знаки, є сильно позитивними, а також часто окличний знак використовується не сам по собі, а у кількості двох, трьох, або й більше окличних знаків підряд, що приводить до певної плутанини при спробі моделі визначити логічні зв’язки та вагу одинарного окличного знаку.

Відгуку «падло» було автоматично присвоєно індекс 3 - нейтральний. Це є грубою помилкою, оскільки “падло” - це лайливий вульгаризм, що означає Підла, негідна людина (Портал української мови та культури slovnyk.ua, *тлумачення слова “падло”)*. Єдина вірна оцінка - 1, сильний негатив.

Відгуку «Бла-бла-бла!» було автоматично присвоєно індекс 3 - нейтральний. «Бла-бла-бла!» без додаткового контексту означає ономатопеїчне кривляння, що виражає роздратування, обурення або незадоволення чиїмось говорінням, негативний заряд цього кривляння в змісті відгуку також додатково посилюється

знаком оклику. Має індекс 2 - негатив. Модель не змогла визначити справжній емотивний заряд відгуку оскільки дані, на яких її було навчено містять надто малу кількість використань цього кривляння.

Відгуку «Гра нормальна та клсна» було автоматично присвоєно індекс 3 - нейтральний. Відгук можна розділити на нейтрально заряджену частину “гра нормальна” і сильно позитивно заряджену частину “та клсна”, що означає “класна”, однак з певних причин виникла одруківка з пропуском однієї літери “а”. Поєднання цих двох частин утворює загальний сильно позитивний сентимент. Скоріш за все модель присвоїла відгуку нейтральний сентимент оскільки надала пріоритет ваги нейтральній частині, для якої був присутній точний відповідник в тренувальних даних, тоді як сильно позитивна частина була оцінена нейтрально або як нерозпізнана через відсутність точних збігів у даних, або ж була інтерпретована як незмістовна.

Потенційними рішеннями проблеми є: збільшення об’єму тренувальних даних для включення більшої кількості подібних випадків для покращення диференціації між незмістовністю та одруківками, покращення n-грамів для додаткового вловлення одруківок.

Відгуку «очень дорогая игра » було автоматично присвоєно індекс 5 - сильно позитивний. Відгук складається з негативної частини “дорогая игра”, тобто “дуже дорога” та підсилювальної частки “дуже”, тобто є сильно негативним. Скоріш за все проблема виникла через те, що модель інтерпретувала відновлений текст “дуже дорога” у позитивному значенні “дуже цінна, близька”

Відгуку «Я щетак в гру незалепав прикольна гра » було автоматично присвоєно індекс 4 - позитивний. Відгук написаний дуже специфічно і насправді його текстовий зміст мав виглядати наступним чином: “Я ще так в гру не залипав, прикольна гра”. “Залипати” - сленгове слово, що означає “сильно занурюватись у щось, бути зацікавленим до такої міри що дуже складно перестати займатись цим, або приділяти увагу”. Відповідно сентимент складається з сильно позитивної

частини “Я щетак в гру незалепав” та позитивної частини “прикольна гра”, отже загальний сентимент є сильно позитивним, адже сильно позитивно заряджене значення дієслова “залипати” додатково підсилюється порівняльним зворотом “я ще так не”, який виражає найвищий ступінь порівняння, тобто сильно позитивна частина переважає над позитивною, котра не суперечить сильно позитивній. Цей неправильно класифікований моделлю відгук можна вважати показовим, оскільки він відображає одразу декілька основних проблем текстового змісту відгуків та їх поєднання:

- сленг “залепати”(залипати)
- відмінне від стандартної форми написання “залепати”
- злиття слів “щетак”, “незалепати
- відсутність знаків пунктуації “незалепав прикольна гра”

Відгуку «Кохаю уно » було автоматично присвоєно індекс 3 - нейтральний. Відгук є сильно позитивним через сильний вираз симпатії “кохаю” адресований додатку, до якого було залишено відгук - “уно”. Цей випадок є в певній мірі артефактом, що виник з досить дивної причини: за всі 18802 відгуки в загальному датасеті слово “кохаю” зустрічається лише один раз і оскільки воно потрапило до тестової його частини, то створена за допомогою машинного навчання модель просто не змогла його розпізнати й присвоїла йому нейтральний сентимент, що у поєднанні з нейтральною назвою самого додатку утворило нейтральний загальний сентимент.

3.2 Помилки при використанні ChatGPT

Відгуку «ігра затягує і реклами не багато » було автоматично присвоєно індекс 4 - позитивний. “затягує” у тексті відгуку означає Захоплювати, приваблювати (Портал української мови та культури slovnyk.ua, *тлумачення слова “затягти”)*. Відгук складається з сильно позитивної частини “ігра затягує” та позитивної частини “реклами не багато” з’єднаних сурядним сполучником “і”.

Оскільки сурядний зв’язок між частинами речення-тексту вказує на їх рівноправність, а позитивна частина не суперечить сильно позитивній, то загальний сентимент є сильно позитивним.

Відгуку «Бріліант » було автоматично присвоєно індекс 5 - сильно позитивний. “Бріліант” є видозміненою формою слова “брильянт” у його переносному значенні, а тобто “щось надзвичайно коштовне, цінне, витончене або рідкісне, видатне в мірі, подібній до того, наскільки сильно проявляють ці характеристики справжні діаманти відносно інших матеріалів”. Сентимент заряд відгуку є сильно позитивним. Вірогідно помилка сталась через те, що слово “брильянт” на момент створення випробуваних великих мовних моделей GPT вже було частково застарілим, замість нього більш поширеним стало слово “діамант”, що в поєднанні з додатково видозміненим написанням призвело до неправильної класифікації.

Відгуку «Супер але перевод непогано » було автоматично присвоєно індекс 4 - позитивний. Зміст цього відгуку є досить неоднозначним, оскільки з одної сторони містить протиставний зв’язок виражений сполучником “але”, однак семантично сильно позитивній частині “Супер” не суперечить протиставлена позитивна частина “перевод непогано”. Можна висунути логічне припущення, що користувач вкладав у свій відгук смисл “Супер, але перевод(переклад) - погано” і префікс “не-” був доданий автодоповненням, однак цьому суперечить залишена ним оцінка 5, тобто сильно позитивна, а також немає об’єктивних підстав з достатньою впевненістю стверджувати про зверхність цього припущення над прямим значенням використаного слова. Зважаючи на перелічені фактори більш правильною можна вважати оцінку 5, тобто сильний позитив.

3.3 Помилки при використанні DeepSeek

Помилка, за якої відгуку «Пока качается я поставлю 5 зірок відгуки хороші» було автоматично присвоєно індекс 5 - сильно позитивний. Насправді відгук

складається з нейтрального компоненту «Пока качается я поставлю 5 зірок», який не несе в собі емотивного заряду і є лиш констатацією дії, що не залежить від самого додатку, оскільки користувач його ще навіть не встановив, на що він вказує в тексті і позитивної частини “відгуки хороші”, тобто загальне емотивне забарвлення є позитивним, а вживання користувачем фрази “поставлю 5 зірок” не відповідає індексу емотивності в ті ж 5 балів, бо виражає не справжній рейтинг додатку наданий користувачем, або його емотивне ставлення до додатку.

Відгуку «Неплохая» автоматично присвоєно індекс 3 - нейтральний. Україномовним відповідником до тексту відгуку є “непоганий”, згідно з тлумачним словником української мови - 1. Досить добрий. 2. Те саме, що гарний. Тобто відгук має індекс 4 - позитивний.

Відгуку «Вибачте в нене з ник акаут» автоматично присвоєно індекс 1 - сильно негативний. Слово “акаут” є одруківкою, правильним словом є “аккаунт”, “нене” - “мене”, а “з ник” - “зник”. Повний правильний вираз “Вибачте вмене зник аккаунт” виражає помірний негативний сентимент, до того ж частково пом’якшений вставним словом ввічливості “вибачте”, тому відгуку присвоєно індекс 2. Модель скоріш за все помилилась саме через комбінацію слів, написаних з одруківками, гіпотетично індекс 1, сильний негатив, може випливати з тлумачення “нене” як здвоєної заперечної частки подібної до “ні-ні”, яка має негативний сентимент відмови, незадоволеності, також можливо частина “аут” з “акаут” могла бути сприйнята як образливий жаргонізм, пов’язаний з аутизмом, когнітивними складнощами індивіда.

Відгуку «Ьреиа» автоматично присвоєно індекс 1 - сильно негативний. Змісту відгуку не відповідає жодне літературне чи сленгове слово яке можливо було б відтворити по пам’яті й з об’єктивно стверджувати про його контекстуальну підходящість наповненню додатку, лексичному інвентарю відгуків або ж ігор, тому при анотуванні йому було присвоєно нейтральний сентимент заряд. Сам Deepseek на відокремлене запитання про сентимент заряд цього відгуку, сформоване

ідентично до запиту, що був використаний у випробуванні API, надав відгуку індекс 3 - нейтральний, а на вказівку додати пояснення відповів наступне:“Слово "Ьреиа" не є коректним українським словом або фразою. Воно не містить ні позитивного, ні негативного смислового навантаження, оскільки позбавлене будь-якого зрозумілого значення. У відсутності ознак тональності (емоційних слів, інтонації, контексту) класифікується як нейтральне (3). Встановити точну причину виникнення цієї помилки, як і методи її виправлення, наразі неможливо, оскільки неможливо відслідкувати логічні зв’язки якими керувалась модель або хоча б стартову точку їх розбудови. Зважаючи на те, що великі мовні моделі за результатами проведених експериментів зазвичай навпаки тяжіють до класифікації незмістовними, а тобто нейтральними за сентимент зарядом, навіть не надто важких для відтворення закладеного початкового змісту одруківок, логічніше за все вважати, що у роботі алгоритму моделі просто сталася технічна помилка, що є її значним недоліком.

3.4 Спільні для ChatGPT і DeepSeek помилки

Відгуку «гра не запускається» було автоматично присвоєно індекс 1 - сильний негатив. Вираз “гра не запускається” означає помірне розчарування або роздратування проблемою з базовим функціонуванням додатку, якого, однак, зазвичай недостатньо при відсутності додаткового контексту аби присвоїти сильно негативне емотивне забарвлення, оскільки частіше за все в такому випадку йдеться про тимчасову проблему, яка не виникає на постійній основі(в такому випадку зазвичай використовують підсилювальні конструкції на кшталт “взагалі не запускається”), тому відгук має індекс 2 - негативний.

Відгуку «Де шрек» автоматично присвоєно індекс 3 - нейтральний. У відгуку користувач задає питання, що виражає нерозуміння розташування персонажу у сфері додатку або розчарування відсутністю названого персонажу в ній, через що відгук має індекс 2 - негативний.

Відгуку «Доволі крута гра для свого жанру.» було автоматично присвоєно індекс 4 - позитивний. Відгук складається з сильно позитивного компоненту “крута” й нейтрального компоненту “доволі”, за тлумачним словником української мови - те саме, що й “досить” (Портал української мови та культури slovnyk.ua, *тлумачення слова “доволі”*)

Помилка, за якої відгуку «ліке» було присвоєно індекс 3 - нейтральний. Насправді «ліке» - жартівлива вимова “лайк”, тобто схвалення, має індекс 4 - позитивний.

3.5 Помилки при використанні Claude

Відгуку “Игра стоит свеч…” було автоматично присвоєно індекс 3 - нейтральний. Насправді цей відгук має позитивний сентимент, оскільки є видозміненою формою сталого виразу “игра стоит свеч”, тобто “гра вартує свіч”, що означає “гра вартує витраченого на неї ресурсів(часу та/або зусиль, грошей і т.д.)”. Відповідно цей відгук є семантично наближеним до позитивних відгуків накшталт “варта уваги”, “варто спробувати”, “не пожалкуєте”. Оскільки у відгуку немає інших сентиментально заряджених компонентів, а пунктуаційний маркер незавершеності “...” сам по собі має лише нейтральний заряд, недостатній для зміщення сентимент-заряду тексту, то відгук має індекс 4 - позитивний. Скоріш за все до помилки призвело використання російської мови та пунктуаційний маркер незавершенності.

Відгуку «ігра затягує і реклами не багато» було автоматично присвоєно індекс 4 - позитивний. Ця помилка вже зустрічалася та була проаналізована у підрозділі 3.2 помилки при використанні ChatGPT. Відповідно до провденого аналізу відгук має індекс 5 - сильно позитивний.

Відгуку «Очень хороша ігра» було автоматично присвоєно індекс 4 - позитивний. Відгук є сильно позитивним, оскільки позитивна частина “хороша ігра” додатково підсилюється інтенсифікатором “очень”. Отже, відгук має індекс 5

- сильно позитивний. Аналогічно до попередньо проаналізованої помилки з відгуком “Игра стоит свеч…” скоріш за все до помилки призвело використання російськомовного суржику.

Відгуку «Клосна» було автоматично присвоєно індекс 3 - нейтральний. Насправді «Клосна» є одруківкою або фонетично викривленим написанням слова “класна”, котре має сильно позитивне емотивне забарвлення. Отже, відгук має індекс 5 - сильно позитивний. Скоріш за все помилка виникла через неспроможність моделі встановити логічний зв’язок між словом у відгуку та словом яке малося на увазі, попри те, що вони відрізняються лише на одну літеру і є певною мірою досить подібними за звучанням.

Висновки до розділу 3

Всі розглянуті моделі, особливо ChatGPT, DeepSeek та Claude навіть при відсутності впливу незбалансованості датасету у zero-shot випробуваннях тяжіють до досить специфічного розподілу зваженої метрики, а саме: найбільш точний показник мають класи 1 та 5, середній — клас 3, класи 2, 4 мають загалом найнижчі й помітно менші від інших класів показники, що відповідає загальному стану задачі сентимент аналізу у світі — найбільш точним і розвиненим є визначення полярностей добре/погано - 1 і 5, тобто двоаспектний аналіз, більш розвиненим є триаспектний з урахуванням нейтрального класу - 1, 3 і 5, а п’ятиаспектний з усіма п'ятьма є менш розвинутим. З метою подолання цієї проблеми в наступному етапі дослідження доцільно вжити наступних заходів: -Збалансувати датасет за кількістю відгуків, що репрезентують кожний клас, оскільки в поточному датасеті сильно переважає 5, в той час, як класи 2 і 3 навпаки репрезентовані менше. Це також збільшить обсяг навчального матеріалу загалом, а також дозволить більш точно аналізувати результати роботи завдяки збалансованості впливу кожного класу на модель. - Провести подальші тестування з визначення найбільш частих помилок поміж моделями та для моделей окремо з метою точного виявлення їх специфіки та

адаптації моделей відповідно до неї. -Розвинути та застосувати у навчанні моделей внесений при розміченні датасету індекс дуальності, що може суттєво допомогти вирізняти й належним чином оцінювати змішані компоненти, з яких часто складаються відгуки репрезентовані класами 2 і 3.

РОЗДІЛ 4. ФІНАЛІЗАЦІЯ МОДЕЛІ

У четвертому розділі представлено етап завершального доопрацювання, або ж фіналізації моделі на основі інформації та здобутків з попередніх розділів. Спочатку описано перевірку та невелике доопрацювання датасету, оскільки якість даних безпосередньо впливає на результат роботи моделей. Далі було розроблено і допрацьовано шляхом експериментів покращення найефективнішої локальної моделі svc_combo, зокрема зміну обробки складних пунктуаційних знаків, зміну обробки текстових ознак та їх міри впливу одне на одного. Окрему увагу виділено каскадним підходам та експериментами із застосування двох авторських метрик, котрі, попри їх інтерпретаційну цінність, не вдалось застосувати для покращення моделі. Завершальним етапом є створення повної програмної версії придатної для зручного використання в реальних задачах автоматичного сентимент-аналізу україномовних відгуків Google Play.

4.1 Перевірка та доопрацювання датасету

В процесі аналізу помилок моделей та перегляду створених датасетів було помічено деякі випадки присвоєння відгукам неоптимального індексу емотивності, котрий з тих чи інших причин видавався оптимальним на етапі створення датасету, однак на поточному етапі дослідження більш не є таким з огляду на отриманий досвід використання й аналізу емотивних текстів, наприклад:

91	Ахуеть знов можна ловити покемонів по церквах	5	4	у

Рис. 4.1 Приклад неоптимально класифікованого відгуку №1

Наведеному відгуку було присвоєно індекс емотивності 2 через вельми неоднозначний вислів та неодноразові судові справи пов’язані з описаним видом діяльності, негативне ставлення деяких віруючих людей до неї. Однак зі сторони

автора та в контексті відношення до відгуку навряд чи доречним є розглядати даний вислів настільки заглиблено і буквально, адже скоріш за все він є просто жартівливо-позитивним, хоч і з певною долею сумнівного гумору. Виправлено на індекс емотивності 4.

Рис. 4.2 Приклад неоптимально класифікованого відгуку №2

Наведеному відгуку було присвоєно індекс емотивності 5 виходячи з ідеї, що загалом можна було здогадатись, що автор, скоріш за все, мав на увазі сильно позитивний вислів “офігенна гра”, котрий передав у видозміненому вигляді, але при повторному аналізі стало очевидним, що подібний варіант написання є надто далеким від вбачаємого можливого та стверджувати про правильність такої інтерпретації не є коректним. Виправлено на індекс емотивності 3 - нейтральний, розглянуто як текст, у якому неможливо визначити чіткий зміст.

Рис. 4.3 Приклад неоптимально класифікованого відгуку №3

Наведеному відгуку було присвоєно індекс емотивності 2 через його інтерпретацію як відображення крику страху, шоку, оскільки така семантика є більш розповсюдженою для подібного вигуку. Все ж подібний вигук сам по собі не має сталого, однозначного трактування і може бути тако криком радості, або позитивного здивування. Оскільки розглянутий вигук може мати декілька різних значень залежно від контексту, а у даному відгуку ніякого іншого контексту крім самого вигуку немає, то індекс емотивності було виправлено на 3 - нейтральний.

Було знайдено і виправлено відносно невелику кількість помилок у розміченні сентименту відгуків, всі зроблені виправлення при перевірці тренування і тестування на тих самих моделях дали приріст лише у 0.1%, що несуттєво, однак

знайдені помилки дозволили порівняти сприйняття одних і тих самих відгуках на різних етапах роботи з ними.

4.2 Покращення svc_combo

Для фіналізації моделі була взята за основу найефективніша з розглянутих локальних моделей svc_combo, при використанні якої вдалось досягнути результату в 76.5% за зваженою мірою F1. В процесі аналізу помилок моделі було виявлено, що значна їх кількість приходиться на випадки з одним або декількома пунктуаційними знаками “!”, “?” або їх поєднаннями підряд. У відгуках такі елементи виступають не самостійним маркером позитивного або негативного емотивного заряду, а підсилюючим складником, вплив та значення якого сильно варіюються в залежності від оточуючих елементів тексту, розташування цих знаків в тексті та конкретних їх комбінацій. Також в загальному розумінні існують досить чіткі межі значення лише для комбінацій “?!”, “???” та “!!!”, тоді як , наприклад, комбінації “!!”, “??”, “?!?” або “!?!” не мають чіткого загального розуміння та не використовуються як загальноприйняті впізнавані знаки та сприймаються і використовуються скоріше як певна абстрактна міра підсилення відносно одинарного знаку, або найближчої розпізнаваної в загальному розумінні комбінації. Через такий комплексний та багатошаровий вплив на емотивний заряд тексту чітко визначити який саме вплив має конкретна комбінація цих знаків у конкретному тексті є досить важкою задачею, котра тим далі ускладнюється обмеженістю шкали сентименту, в межах цього дослідження - 5 рівнів, що дозволяє виділити лише ступені інтенсивності для негативу і для позитиву.

З метою покращення ефективності моделі було змінено метод обробки вхідних ознак для TF-IDF векторизатору за допомогою гіперпараметру token_pattern на розпізнавання знаків “!”, “?” та їх комбінацій як окремих токенів, що дало змогу не втрачати їх під час токенізації тексту. Також було покращено врахування токенів комбінацій знаків з сусідніми елементами. Було додано функцію формування

списку різних комбінацій знаків та їх окремого врахування при векторизації, а також введено гіперпараметр punct_weight, котрий відповідає за міру впливу пунктуаційних ознак відносно всіх інших, шляхом експериментів було визначено що оптимальним збалансованим значенням для сприйняття пунктуації як додаткового фактору сентимент-заряду є punct_weight = 3. В результаті всіх вдосконалень ефективність моделі вдалось покращити на 1.2% з 76.6% до 77.8%. Отримане покращення є відносно невеликим саме по собі, проте є вельми суттєвим зважаючи на те, що його було досягнуто відносно вже високого результату моделі в 76.6%, численні експерименти покращення розпізнавання впливу знаків “!”, “?” та їх комбінацій в контексті автоматичного сентимент-аналізу і часткове вирішення проблеми нестандартних комбінацій за допомогою функції формування їх списку вказують на те що наступним логічним кроком є пом’якшення іншого ускладнюючого фактору, а тобто — обмеженої шкали сентименту за допомогою її розширення.

```
=== Classification report: saved SVC model ===
              precision    recall  f1-score   support

           1      0.774     0.701     0.735       581
           2      0.582     0.635     0.607       436
           3      0.616     0.759     0.680       552
           4      0.668     0.690     0.679       555
           5      0.944     0.869     0.905      1636

    accuracy                          0.773      3760
   macro avg      0.717     0.731     0.721      3760
weighted avg      0.787     0.773     0.778      3760

=== Confusion matrix ===
[[ 407   96   55    8   15]
 [  53  277   70   28    8]
 [  25   44  419   40   24]
 [  13   42   80  383   37]
 [  28   17   56  114 1421]]
```

Рис. 4.4 Доопрацьована svc_combo

4.3 Каскадний метод з використанням авторських міток даних(визначення дуальності та незмістовності).

Було проведено низку експериментів з використання авторських метрик розроблених та розмічених на етапі створення датасету з метою потенційного покращення ефективності моделей автоматичного сентимент-аналізу на основі створеного датасету. В межах цих експериментів було перевірено можливості автоматичного визначення моделями цих авторських метрик на основі лише текстового змісту та можливості покращення ефективності визначення сентимент-заряду за рахунок включення розмічених авторських метрик напряму в навчання основної моделі для визначення сентимент-заряду та за допомогою каскадного методу, де для визначення авторських метрик на основі текстового змісту була навчена окрема модель, котра використовувалась попередньо до основної з метою

доповнення її функціоналу в асистентному режимі. У якості основної моделі була взята svc_combo.

Першим експериментом було залучення маркеру незмістовності, котрим позначалися відгуки які не мали логічного цілісного змісту пов’язаного з відгуком про додаток, як, наприклад, випадковий набір символів, або випадкові слова з мобільної функції “автозаповнення”. В режимі включення до тренування основної моделі такий підхід знизив ефективність моделі з 77.8% до 66.7%, а в асистентному режимі до 75.6%.

Другим експериментом було залучення індексу дуальності, що виражає змішаність загального емотивного забарвлення висловлювання, яка утворюється шляхом поєднання у ньому негативно та позитивно забарвлених частин. Значення індексу дуальності позначається оцінкою в 1-3 балів, де відсутність дуальності має оцінку 0, не позначається графічно та має сприйматися, як встановлене за замовчуванням, оцінки від 1 до 3 визначають певні типи та рівні емотивної ускладненості, конкретизують індекс емотивності, відношення між загальним забарвленням та частинами, що його ускладнюють, а саме: 1 – основне забарвлення утворене поєднанням частин відповідного емотивного спектру у великій кількості та/або сильним виявом ознак з частинами протилежного спектру у малій кількості та/або з слабким виявом ознак. Окрім використання в парі з нейтральним індексом емотивності 3, в поєднанні з яким означає домінуюче нейтральне забарвлення з яким пов’язане слабке негативне забарвлення що є надто незначним аби переважити нейтральне; 2 – основне забарвлення утворене поєднанням частин відповідного емотивного спектру у великій кількості та/або сильним виявом ознак з частинами протилежного у значній, але меншій кількості та/або помірним виявом ознак. Окрім використання в парі з нейтральним індексом емотивності 3, в поєднанні з яким означає домінуюче нейтральне забарвлення з яким пов’язане слабке позитивне забарвлення, що є надто незначним аби переважити нейтральне;

3 - основне забарвлення є нейтральним, утворене поєднанням частин обох спектрів однаковій кількості та/або з рівнозначним виявом ознак.

В режимі включення до тренування основної моделі такий підхід знизив ефективність моделі з 77.8% до 72.2%, а в асистентному режимі до 70.6%, зі 135 оцінок сентименту, котрі були змінені на основі даних асистентної моделі дуальності лише 23 стали правильнішими, тоді як 70 погіршилися, а 42 залишились настільки ж правильними попри зміну оцінки(зміна з визначення оцінки 4 на визначення оцінки 2 при правильній оцінці 3).

Отже, експерименти показали, що на даному етапі розроблені авторські метрики незмсістовності й дуальності не підходять для покращення обраної моделі автоматичного сентимент-аналізу та, попри їх інтерпретаційну цінність, потребують їх перегляду.

4.4 Створення повної програмної версії придатної до використання у реальних задачах

Після завершення вдосконалень та екперементального оцінювання було створено повну програмну версію для використання вже готової навченої моделі в зручному й зрозумілому для сприйняття режимі. За допомогою бібліотеки joblib було збережено найефективнішу з випробуваних локальних моделей, а саме svc_combo зі зміненими гіперпараметрами та додатковими функціями обробки. В результаті це дозволило відокремити етап навчання моделі від етапу застосування, що економить час та ресурси обладнання на постійне перетренування моделі.

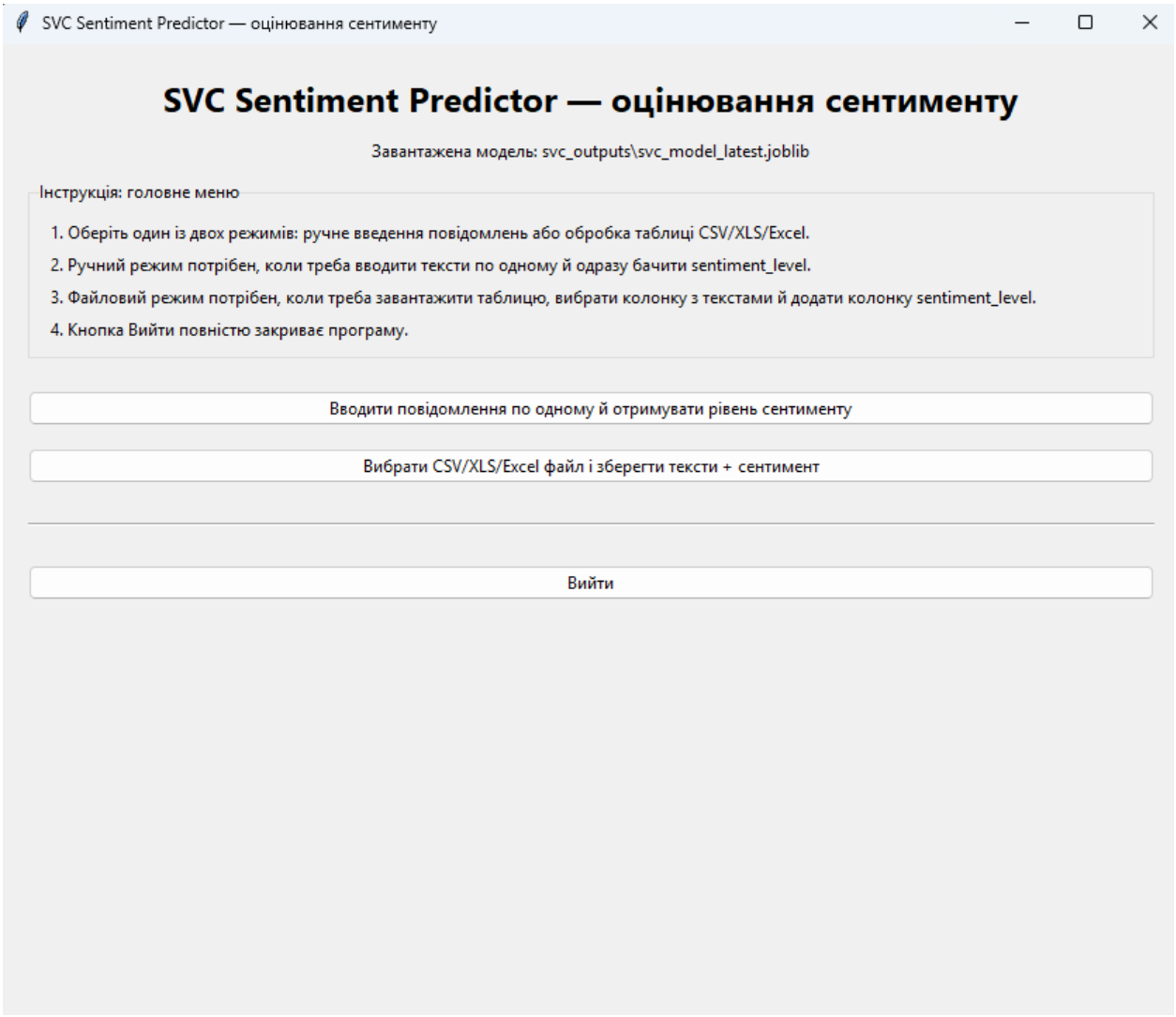


Рис. 4.5 Головне меню користувацького інтерфейсу

Програма має віконний інтерфейс та підтримує декілька режимів роботи, на всіх етапах використання програми можливо повернутись до головного вікна або ж вийти з програми за допомогою відповідних кнопок. Передбачено класифікацію окремих текстів введених користувачем у відповідне текстове поле з подальшим їх збереженням та відображенням в межах історії користувацької сесії й можливості очищення або експорту цієї історії у вигляді файлу.

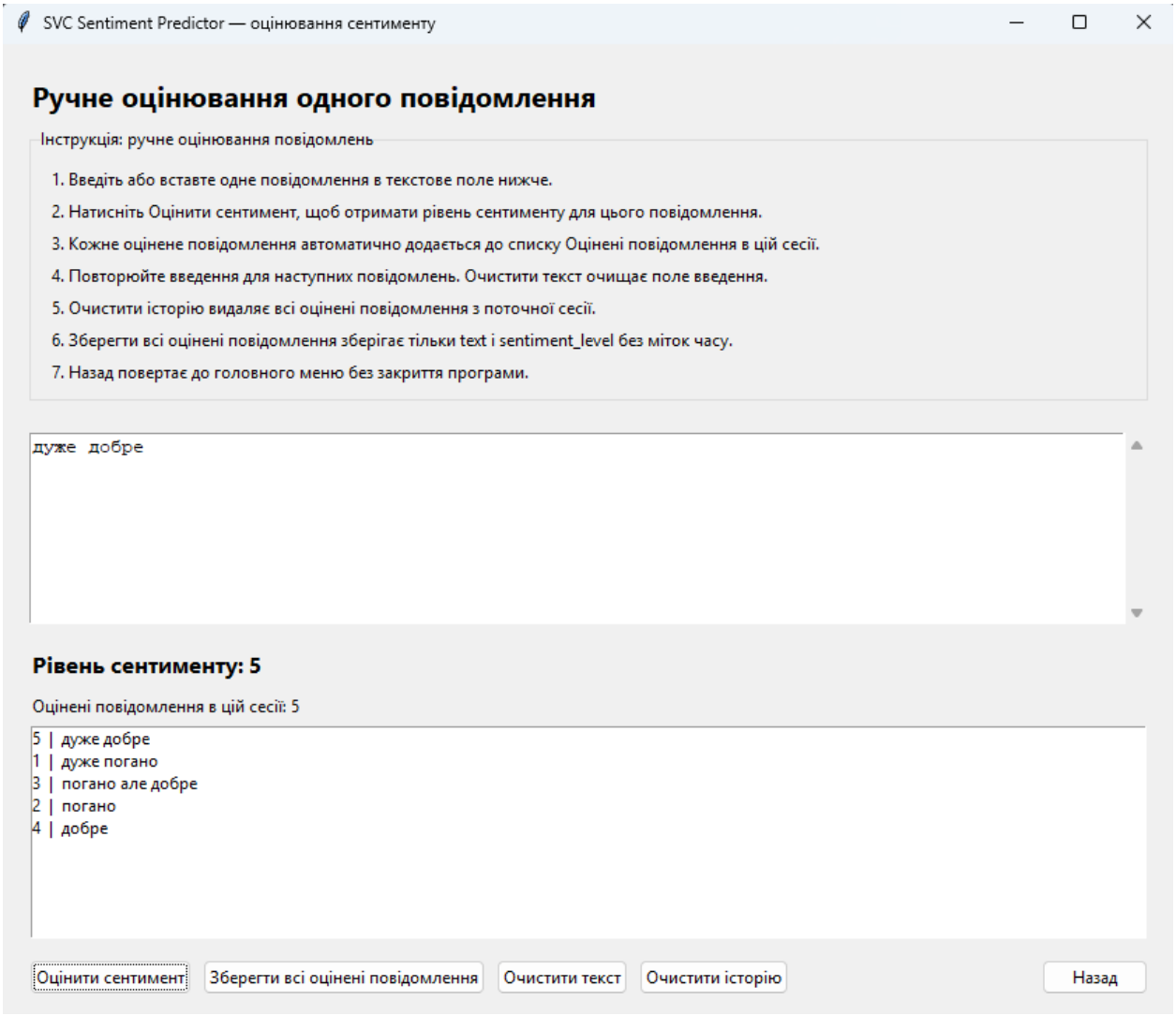


Рис.4.6 Інтерфейс обробки та збереження текстів у вигляді окремих повідомлень

Також передбачено режим роботи з наборами даних у файлових форматах таких як .csv, .xlsx та .xls. Після завантаження такого файлу користувач може вказати колонку з якої варто вилучити текстовий зміст та оцінити його сентимент-заряд, або ж залишити назву колонки за замовчуванням. Після чого натиснути кнопку “Оцінити файл” і програма відобразить перші 20 оцінених текстів. При натисненні кнопки “Зберегти результати” програма створить копію файлу поданого на вхід з доданою колонкою що містить визначений сентимент-заряд, цей буде збережено у вибраному форматі.

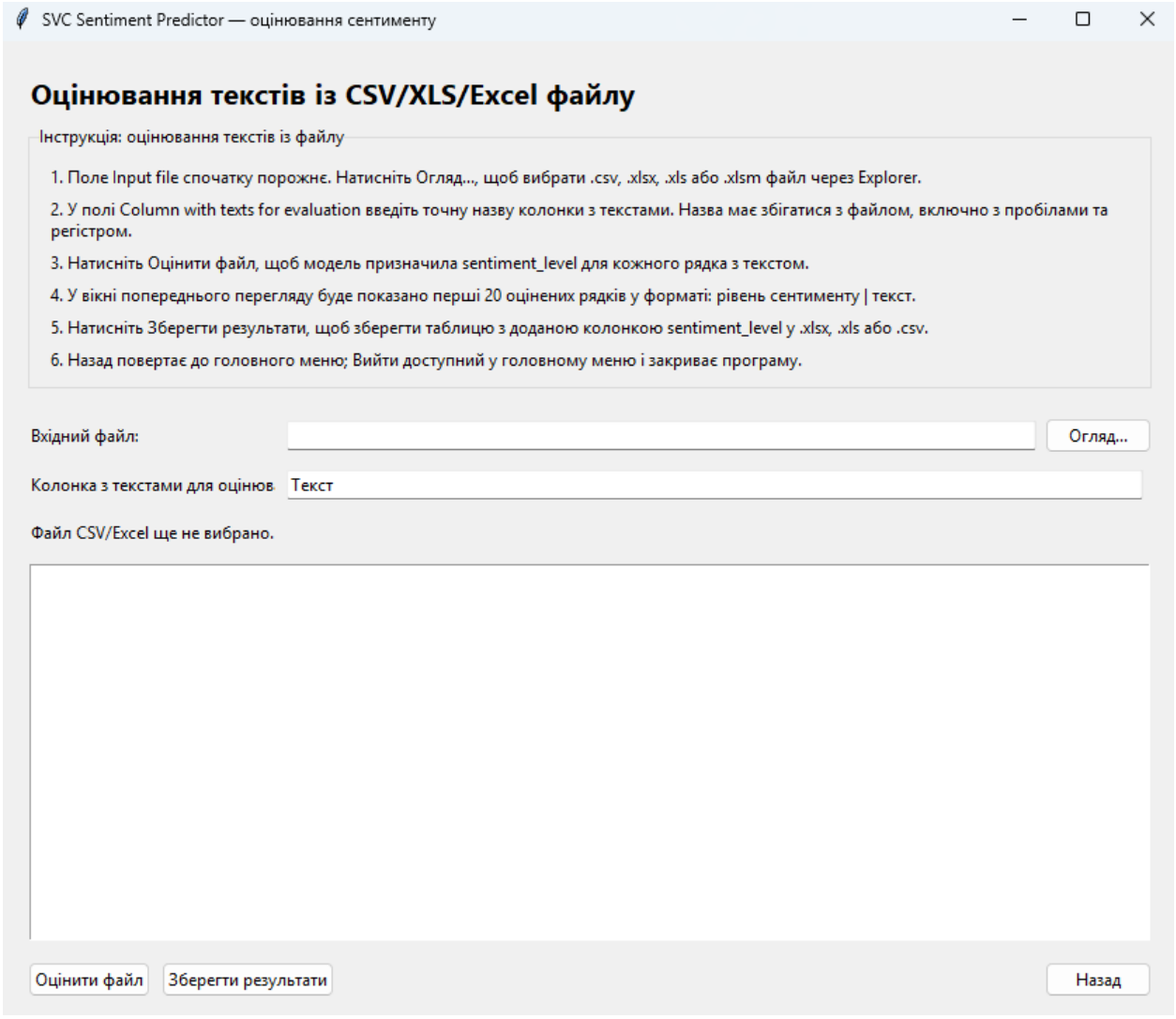


Рис. 4.7 Інтерфейс обробки та збереження текстів у файловому форматі

4.5. Практичне застосування та майбутній розвиток

Фіналізована версія моделі у вигляді програмного користувацького інтерфейсу може бути використана для зручного отримання й збереження досить точних визначень сентимент-заряду текстів україномовних відгуків Google Play, або текстів схожих на них, як за допомогою завантаження текстів у файловому форматі в режимі оцінювання файлу, так і за допомогою введення текстів по одному в режимі ручного оцінювання. Одними з найперспективніших кроків для подальшого розвитку моделі є розширення датасету, особливо в рамках менш репрезентованих класів та розширення шкали сентименту, заручення декількох різних спеціалістів

до розмітки тренувальних й тестувальних даних задля уникнення особистого упередження.

Висновки до розділу 4

Найбільш ефективною за зваженою метрикою з методів машинного навчання виявилась модель SVC_combo - 76.5%, що вдалось покращити до 76.6% за допомогою доопрацювання деяких знайдених дефектів розмітки датасету та до 77.8% завдяки додатковій обробці засобів виразності. Такий приріст ефективності хоч і є досить малим сам по собі , однак його отримання є досить суттєвим досягненням через вже високий рівень ефективності моделі, що була взята за основу. Така доопрацьована модель, хоч і все ще значно поступається донавченій GPT4.1, проте має перевагу повної відкритості коду, оскільки його було написано в ході виконання дослідження, швидкості використання, безкоштовності й відсутності залежності від стороннього серверу(технічних робіт, перебоїв і т.д.) що є важливим для забезпечення можливості стабільного й ефективного використання моделі в реальних задачах. На основі цієї покращеної моделі було розроблено програмний користувацький інтерфейс з можливістю вибору декількох режимів обробки текстів та збереження оброблених даних у файловому форматі разом з визначеним моделлю рівнем сентименту. При випробуванні авторських метрик даних та каскадних підходів на основі них дедалі покращити результат доопрацьованої моделі SVC_combo не вдалось

ВИСНОВКИ ДО РОБОТИ

У результаті роботи було проведено експерименти з випробування та модифікації численних сучасних методів та підходів до автоматичного сентимент аналізу відповідно до задач автоматичного сентимент аналізу україномовних відгуків Google Play, зібрано та проаналізовано широкий спектр інформації на основі отриманих результатів експериментів й збору даних та виконано логічні кроки з покращення моделей відповідно до проаналізованої інформації. Було досягнуто досить хорошого для п’ятиаспектного сентимент-аналізу україномовних текстів відгуків Google Play результату в 77.8% правильних передбачень завдяки моделі машинного навчання з доопрацюванням налаштувань її гіперпараметрів та функцій, що є значним результатом та доводить можливість ефективного визначення сентимент-заряду в межах п’ятиаспектної шкали сентименту, котра, своєю чергою, дозволяє точніше відображати сентимент та у випадку відгуків Google Play, що використовують п’ятизіркову рейтингову систему дозволяє порівнювати визначений або присвоєний при розміченні даних сентимент-заряд з оцінкою наданою користувачем, як напряму, так і в якості даних для подальшого використання. В результаті аналізу помилок та проблеми що виникли при випробуванні різних підходів та їх модифікацій можна дійти загального висновку що наступним логічним кроком на шляху до покращення моделей є розширення використовної шкали сентименту з метою зменшення дисперсії при накладанні різних рівнів підсилення емотивного заряду, як для покращення наочності, так і для більш точного розділення даних, котре дозволить суттєво посилити розрізнювальну здатність моделі.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Бортун, К. О. (2023) ‘Сутнісні вияви експресії в архітектоніці конотації’, у: *Наука та освіта в дослідженнях молодих учених* [Електронне видання]: матеріали IV Міжнародної науково-практичної конференції для студентів, аспірантів, докторантів, молодих учених, Харків, 18 травня 2023 р. Харків: Харківський національний педагогічний університет імені Г. С. Сковороди, с. 141–142. Режим доступу: https://repositsc.nuczu.edu.ua/bitstream/123456789/18742/1/%D0%9D%D0%B0%D1%83%D0%BA%D0%B0%20%D1%82%D0%B0%20%D0%BE%D1%81%D0%B2%D1%96%D1%82%D0%B0%20%D0%B2%20%D0%B4%D0%BE%D1%81%D0%BB%D1%96%D0%B4%D0%B6%D0%B5%D0%BD%D0%BD%D1%8F%D1%85%20%D0%BC%D0%BE%D0%BB%D0%BE%D0%B4%D0%B8%D1%85%20%D1%83%D1%87%D0%B5%D0%BD%D0%B8%D1%85.%202023.pdf .
2. Кузенко, Г. М. (2001) ‘Емотивність на різних мовних рівнях’. Режим доступу: https://lib.chmnu.edu.ua/pdf/novitfilolog/11/72.pdf.
3. Портал української мови та культури slovnyk.ua (2025) Тлумачення слова “доволі”. Режим доступу: https://slovnyk.ua/index.php?swrd=доволі
4. Портал української мови та культури slovnyk.ua (2025) Тлумачення слова “затягти”. Режим доступу: https://slovnyk.ua/index.php?swrd=затягати
5. Портал української мови та культури slovnyk.ua (2025) Тлумачення слова “падло”. Режим доступу: https://slovnyk.ua/index.php?swrd=падло
6. Студентський соціолект як складник мовної організації сучасних молодіжних журналів (2018). Шифр: «Молодь». Режим доступу: https://philology.lnu.edu.ua/wp-content/uploads/2018/04/%D0%A8%D0%B8%D1%84%D1%80_%D0%BC%D0

%BE%D0%BB%D0%BE%D0%B4%D1%8C_%D1%80%D0%BE%D0%B1%D0%BE%D1%82%D0%B0.pdf.

7. Шинкарук, В. Д. (2011) ‘Особливості реченнєвих структур з емоційно-оцінними значеннями’. Режим доступу: https://www.researchgate.net/publication/356982228_Osoblivosti_recennevih_struktur_z_emocijno-ocinnimi_znacennami
8. Abidin, AZ, Furqon, MA and Bukhori, S (2025) ‘Combining Rule Based, Lexicon Based and Support Vector Machine for Improve Accuracy in Sentiment Analysis of ChatGPT Usage’, Journal of Research in Artificial Intelligence for Systems and Applications (RAISA). Available at: https://www.researchgate.net/profile/Muhammad-Furqon-2/publication/387797121_Combining_Rule_Based_Lexicon_Based_and_Support_Vector_Machine_for_Improve_Accuracy_in_Sentiment_Analysis_of_ChatGPT_Usage/links/677ddab1763f322e066067ad/Combining-Rule-Based-Lexicon-Based-and-Support-Vector-Machine-for-Improve-Accuracy-in-Sentiment-Analysis-of-ChatGPT-Usage.pdf
9. Bahari, Aris Rifki Setiya, Utomo, Fandy Setyo and Berlilana (2026) ‘Systematic Review of Hyperparameter Adjustment and Evaluation Metrics in BERT-Based Sentiment Analysis’. Available at: https://www.newinera.com/index.php/JournalLaMultiapp/article/view/3046/2336.
10. Barreto et al. (2021) ‘Sentiment Analysis in Tweets: An Assessment Study from Classical to Modern Text Representation Models’. Available at: https://arxiv.org/pdf/2105.14373.
11. Basheer, K.C. Sabreena (2025) ‘OpenAI o3-mini: Performance, How to Access, and More’, Analytics Vidhya. Available at: https://www.analyticsvidhya.com/blog/2025/02/openai-o3-mini/
12. Bayram, A., Menekse Dalveren, G. G. and Derawi, M. (2025) ‘Comparative Analysis of AI Models for Python Code Generation: A HumanEval Benchmark

Study', Applied Sciences, 15, article 9907. doi: https://doi.org/10.3390/app15189907.

13. Bouazizi, M. and Ohtsuki, T. (2019) 'Multi-class Sentiment Analysis on Twitter: Classification Performance and Challenges', Big Data Mining and Analytics. doi: https://doi.org/10.26599/BDMA.2019.9020002.
14. Chen, Jian, Tang, Guohao, Zhou, Guofu and Zhu, Wu (2025) 'ChatGPT and Deepseek: Can They Predict the Stock Market and Macroeconomy?'. Available at: https://arxiv.org/pdf/2502.10008 .
15. Devlin, Jacob, Chang, Ming-Wei, Lee, Kenton and Toutanova, Kristina (2018) BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding; bert/multilingual.md, list of languages. Available at: https://github.com/google-research/bert/blob/master/multilingual.md#list-of-langu
16. Devlin, Jacob, Chang, Ming-Wei, Lee, Kenton and Toutanova, Kristina (2018) BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding; google-bert/bert-base-multilingual-cased. Available at: https://huggingface.co/google-bert/bert-base-multilingual-cased.
17. Devlin, Jacob, Chang, Ming-Wei, Lee, Kenton and Toutanova, Kristina (2019) 'BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding'. Available at: https://arxiv.org/abs/1810.04805.
18. Dodge, J. et al. (2020) 'Fine-Tuning Pretrained Language Models: Weight Initializations, Data Orders, and Early Stopping'. Available at: https://arxiv.org/abs/2002.06305.
19. Huang, Donghao and Wang, Zhaoxia (2025) 'Explainable Sentiment Analysis with DeepSeek-R1: Performance, Efficiency, and Few-Shot Learning'. Available at: https://arxiv.org/pdf/2503.11655 .
20. Huang, PeiHsuan, Lin, ZihWei, Imbot, Simon, Fu, WenCheng and Tu, Ethan (2025) 'Analysis of LLM Bias (Chinese Propaganda & Anti-US Sentiment) in

DeepSeek-R1 vs. ChatGPT o3-mini-high'. Available at: https://arxiv.org/pdf/2506.01814.

21. Kiritchenko, S., Zhu, X. and Mohammad, S. M. (2014) 'Sentiment Analysis of Short Informal Texts'. Available at: https://nrc-publications.canada.ca/eng/view/accepted/?id=f3c48029-99e0-48c7-9aaf-271e9715465b.
22. Krugmann, J. O. and Hartmann, J. (2024) 'Sentiment Analysis in the Age of Generative AI', Customer Needs and Solutions, 11, article 3. doi: https://doi.org/10.1007/s40547-024-00143-4.
23. Liu, Hanzhi, Shou, Chaofan, Liu, Xiaonan, Wen, Hongbo, Chen, Yanju, Fang, Ryan Jingyang and Feng, Yu (2026) 'Synthesizing Multi-Agent Harnesses for Vulnerability Discovery'. Available at: https://arxiv.org/pdf/2604.20801.
24. Liu, X. and Wang, C. (2021) 'An Empirical Study on Hyperparameter Optimization for Fine-Tuning Pre-trained Language Models', Proceedings of ACL-IJCNLP 2021, pp. 2286–2300. Available at: https://aclanthology.org/2021.acl-long.178/.
25. Madjid, M. F., Ratnawati, D. E. and Rahayudi, B. (2023) 'Sentiment Analysis on App Reviews Using Support Vector Machine and Naïve Bayes Classification', Sinkron: Jurnal dan Penelitian Teknik Informatika, 7(1), pp. 556–562. doi: https://doi.org/10.33395/sinkron.v8i1.12161.
26. Madmon, Omer, Golan, Shiran, Fainman, Eran, Kleinfeld, Ofri and Belade, Moran (2026) 'TRABL: A Unified Framework for Travel Domain Aspect-Based Sentiment Analysis Applications of Large Language Models'. Available at: https://dl.acm.org/doi/epdf/10.1145/3774904.3792835 (Accessed: 10 May 2026).
27. Montesinos-López, O.A., Montesinos-López, A. and Crossa, J. (2022) 'Support Vector Machines and Support Vector Regression', in *Multivariate Statistical Machine Learning Methods for Genomic Prediction*. Cham: Springer. Available at: https://www.ncbi.nlm.nih.gov/books/NBK583961/

28. Mohammad, Saif M. (2016) 'A Practical Guide to Sentiment Annotation: Challenges and Solutions', Proceedings of NAACL-HLT 2016. Available at: https://aclanthology.org/W16-0429.pdf
29. Pang, B. and Lee, L. (2008) 'Opinion Mining and Sentiment Analysis'. Available at: https://www.cs.cornell.edu/home/llee/omsa/omsa.pdf.
30. Peng, Tiantian and Tayefeh, Shakila (2025) 'Catastrophic Forgetting in Language Models'. Available at: https://odr.chalmers.se/server/api/core/bitstreams/e504c4f4-b6a5-4903-88d3-5dbe72a0840c/content.
31. Plaat, Aske, Wong, Annie, Verberne, Suzan, Broekens, Joost, van Stein, Niki and Back, Thomas (2024) 'Reasoning with Large Language Models, a Survey'. Available at: https://arxiv.org/pdf/2407.11511
32. Ranjan, S. and Mishra, S. (2020) 'Comparative Sentiment Analysis of App Reviews', *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, pp. 1–7. doi: 10.1109/ICCCNT49239.2020.9225348
33. Singh, Aaditya, Fry, Adam, Perelman, Adam and others (2025) 'OpenAI GPT-5 System Card'. Available at: https://arxiv.org/abs/2601.03267.
34. Sobo, Andrei, Mubarak, Awes, Baimagambetov, Almas and Polatidis, Nikolaos (2024) 'Evaluating LLMs for Code Generation in HRI: A Comparative Study of ChatGPT, Gemini, and Claude', Applied Artificial Intelligence. Available at: https://www.tandfonline.com/doi/epdf/10.1080/08839514.2024.2439610?needAccess=true.
35. Suandi, Fadli, Anam, M. Khairul, Firdaus, Muhammad Bambang, Fadli, Sofiansyah, Lathifah, Lathifah, Yumami, Eva, Saleh, Alfa and Hasibuan, Ade Zulkarnain (2024) 'Enhancing Sentiment Analysis Performance Using SMOTE and Majority Voting in Machine Learning Algorithms', Proceedings of the 7th International Conference on Applied Engineering (ICAE 2024). doi: https://doi.org/10.2991/978-94-6463-620-8_10.

36. Sung, Junwon, Heo, Woojin, Byun, Yunkyung and Kim, Youngsam (2023) 'Large Language Models for Semantic Monitoring of Corporate Disclosures: A Case Study on Korea's Top 50 KOSPI Companies'. Available at: https://arxiv.org/pdf/2309.00208
37. Tan, Kian Long, Lee, Chin Poo and Lim, Kian Ming (2023) 'A Survey of Sentiment Analysis: Approaches, Datasets, and Future Research', Applied Sciences, 13(7), article 4550. doi: https://doi.org/10.3390/app13074550.
38. Vamvourellis, Dimitris and Mehta, Dhagash (2025) 'Reasoning or Overthinking: Evaluating Large Language Models on Financial Sentiment Analysis'. Available at: https://arxiv.org/html/2506.04574v1#bib
39. van Atteveldt, W., van der Velden, M. A. C. G. and Boukes, M. (2021) 'The Validity of Sentiment Analysis: Comparing Manual Annotation, Crowd-Coding, Dictionary Approaches, and Machine Learning Algorithms', Communication Methods and Measures, 15(2), pp. 121–140. doi: https://doi.org/10.1080/19312458.2020.1869198.
40. van der Veen, A.M. and Bleich, E. (2025). 'The advantages of lexicon-based sentiment analysis in an age of machine learning'. *PLOS ONE*, 20(1), e0313092. https://doi.org/10.1371/journal.pone.0313092
41. Wei, Xinyu and Liu, Luojia (2025) 'Are Large Language Models Good In-Context Learners for Financial Sentiment Analysis?', ICLR 2025 Workshop on Advances in Financial AI. doi: https://doi.org/10.48550/arXiv.2503.04873.
42. Xu, Anyi and others (2026) 'DeepSeek-V4: Towards Highly Efficient Million-Token Context Intelligence'. Available at: https://huggingface.co/deepseek-ai/DeepSeek-V4-Pro/blob/main/DeepSeek_V4.pdf.
43. Xu, Lingling, Xie, Haoran, Qin, Si-Zhao Joe, Tao, Xiaohui and Wang, Fu Lee (2023) 'Parameter-Efficient Fine-Tuning Methods for Pretrained Language Models: A Critical Review and Assessment'. Available at: https://arxiv.org/pdf/2312.12148

44. Xu, Qianwen Ariel, Chang, Victor and Jayne, Chrisina (2022) 'A Systematic Review of Social Media-Based Sentiment Analysis: Emerging Trends and Challenges', Decision Analytics Journal, 3. doi: https://doi.org/10.1016/j.dajour.2022.100073.
45. Yaakub, Mohd Ridzwan (2018) 'A Review on Sentiment Analysis Techniques and Applications'. Available at: https://iopscience.iop.org/article/10.1088/1757-899X/551/1/012070/pdf.
46. Yadav, A. and Vishwakarma, D.K. (2020) 'Sentiment Analysis using Deep Learning Architectures: A Review', *Artificial Intelligence Review*, 53, pp. 4335–4385. doi: 10.1007/s10462-019-09794-5
47. Zafar, Lubna, Qamar, Anees and Zareen, Jawad (2023) 'Evaluation of Polarity Features for Sentiment Analysis on Product Reviews', Pakistan Journal of Humanities and Social Sciences, 11. doi: https://doi.org/10.52131/pjhss.2023.v11i4.2105.
48. Zhou, Y. and Srikumar, V. (2022) 'A Closer Look at How Fine-tuning Changes BERT', in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin, Ireland: Association for Computational Linguistics, pp. 1046–1061. Available at: https://aclanthology.org/2022.acl-long.75/

ДОДАТКИ

Додаток 1 перелік створених в ході виконання дослідження файлів та програм

Тренувальний датасет - https://drive.google.com/file/d/1oQdmGjuJ7za7bSmcY3w1hKu3GDT890nu/view?usp=drive_link [онлайн] [23.06.2025]

Тестовий датасет - https://drive.google.com/file/d/1pxErjZP0IHxE_1ZEVjHoUirxjb9K-PWp/view?usp=sharing [онлайн] [23.06.2025]

Моделі машинного навчання - https://drive.google.com/file/d/1XgEiZIjZ40V0iqIOwf49JslmCme2Uby4/view?usp=drive_link [онлайн] [23.06.2025]

Моделі ChatGPT - https://drive.google.com/file/d/1TEGOyHcO-nBYMzpxrqGDTT-1QmcR5FfQ/view?usp=drive_link [онлайн] [23.06.2025]

Моделі DeepSeek - https://drive.google.com/file/d/186_1zJavOaCHhzq4qnzL4buA8rhlK4dd/view?usp=drive_link [онлайн] [23.06.2025]

Датасет для донавчання 100 прикладів - https://drive.google.com/file/d/1c2eeBVgCM-Hh0qN8raQ4rVrsKhCyGCpw/view?usp=drive_link [онлайн] [23.06.2025]

Датасет для донавчання 1000 прикладів -

https://drive.google.com/file/d/1v_D9KIlzjYfTsju3FnTegJvgX4wLhq_C/view?usp=drive_link [онлайн] [23.06.2025]

Фіналізована версія моделі автоматичного сентимент аналізу україномовних відгуків Google Play - https://drive.google.com/drive/folders/1IxzdYwyYov88oah0Z0EyYPZ4b85vyAvK?hl=ru

Доопрацьований тестовий та навчальний датасет - https://drive.google.com/drive/folders/1V6kjJOAvVfNyaOKfd2ioxpzPQF8NZ41Z?usp=sharing

Модель Claude - https://drive.google.com/drive/folders/1Z7zEmDaRljYahWgSTC1ryx0fEpg8AkBF?usp=drive_link

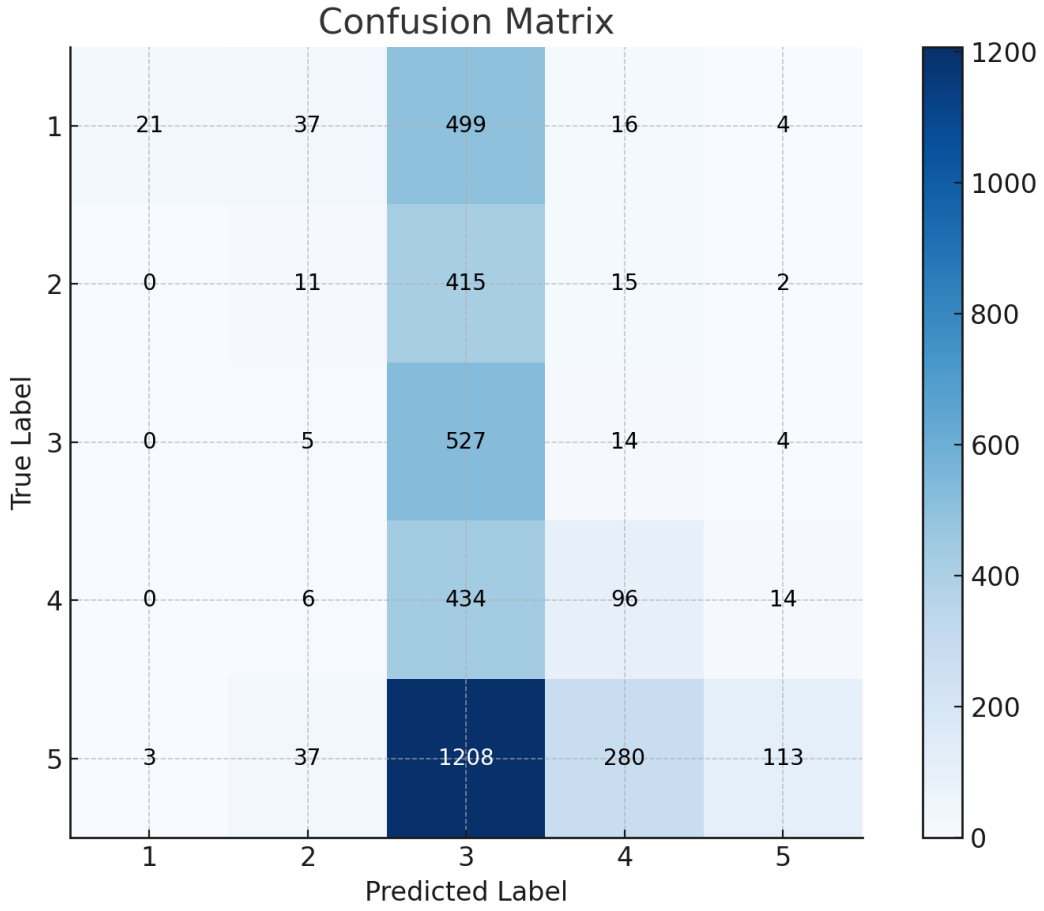
Претренована модель глибокого машинного навчання BERT - https://drive.google.com/drive/folders/1WC6HOfBvWANW8jacD68HFN1a8wl9avaY?usp=sharing

Додаток 2 Словникові методи сентимент аналізу

На основі 5-ти аспектного тонального словника lang-uk

rule-based1

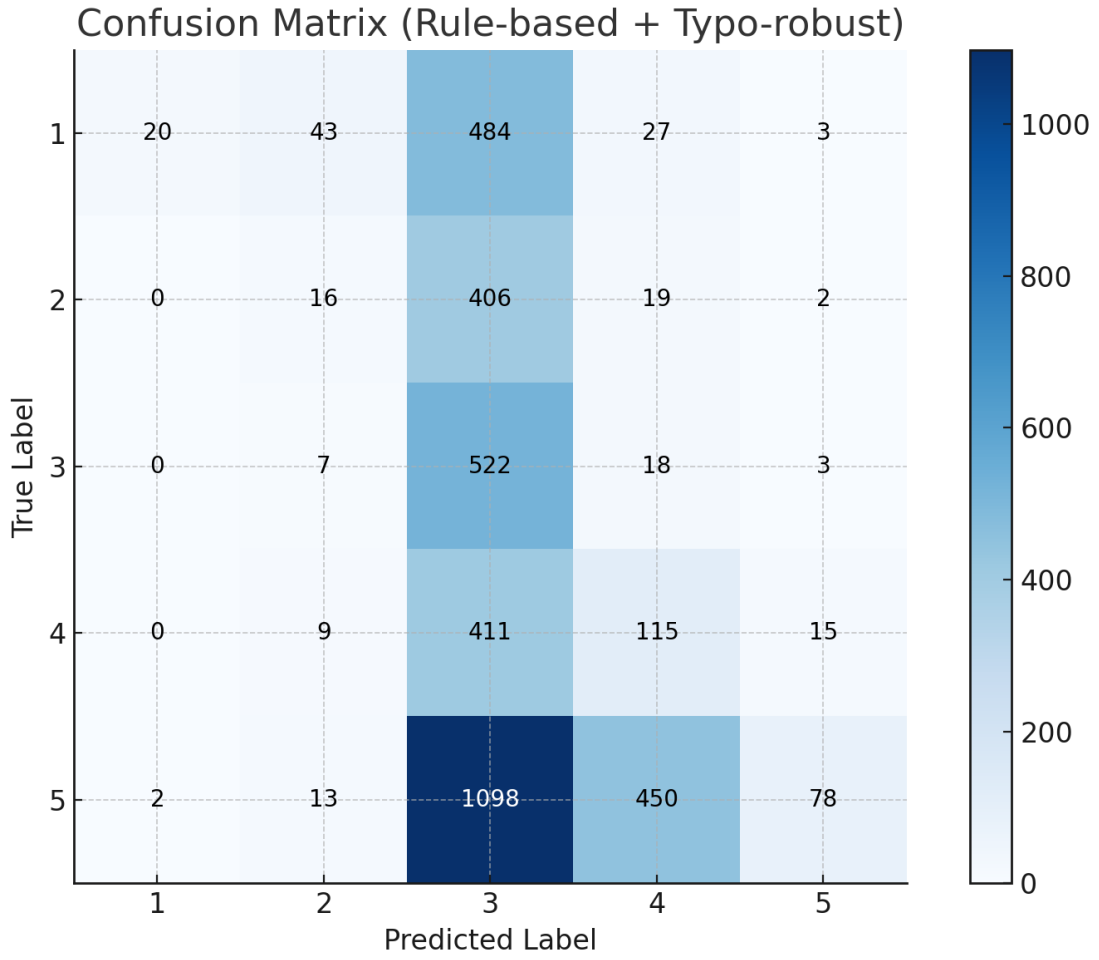
Клас	Precision	Recall	F1-score	Support
1	0.875	0.036	0.070	577
2	0.115	0.025	0.041	443
3	0.171	0.958	0.290	550
4	0.228	0.175	0.198	550
5	0.825	0.069	0.127	1641
Accuracy	0.204			
Macro avg	0.443	0.253	0.145	3761
Weighted avg	0.566	0.204	0.142	3761



rule-based2 TF-IDF

Клас	Precision	Recall	F1-score	Support
1	0.909	0.035	0.067	577
2	0.182	0.036	0.060	443
3	0.179	0.949	0.301	550
4	0.183	0.209	0.195	550
5	0.772	0.048	0.090	1641
Accuracy		0.200		
Macro avg	0.445	0.255	0.142	3761

Weighted avg	0.551	0.200	0.129	3761



Додаток 3 Контрольоване машинне навчання

[1/9] Training logreg_combo … done in 2.6s

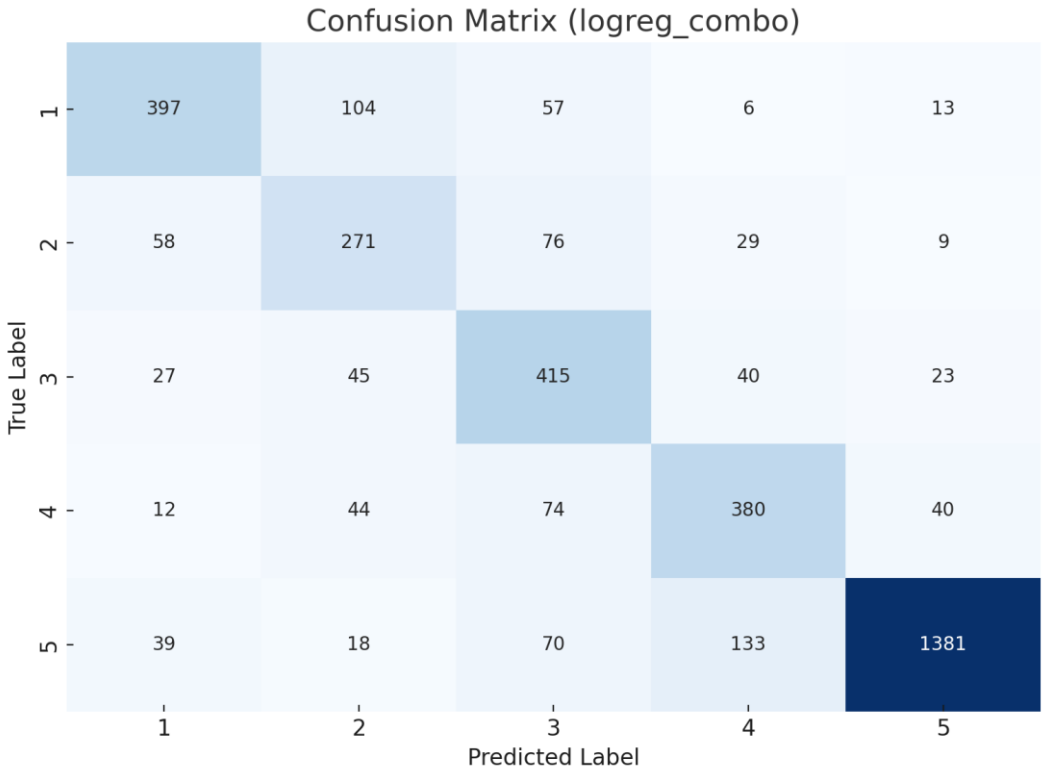
[3761/3761] Elapsed: 8s | ETA: 0s

✓ All responses received.

=== Classification report (logreg_combo) ===

precision recall f1-score support

Class	precision	recall	f1-score	support
1	0.745	0.688	0.715	577
2	0.562	0.612	0.586	443
3	0.600	0.755	0.668	550
4	0.646	0.691	0.668	550
5	0.942	0.842	0.889	1641
accuracy		0.756		3761
macro avg	0.699	0.717	0.705	3761
weighted avg	0.774	0.756	0.762	3761



[2/9] Training smote_logreg_combo … done in 4.0s

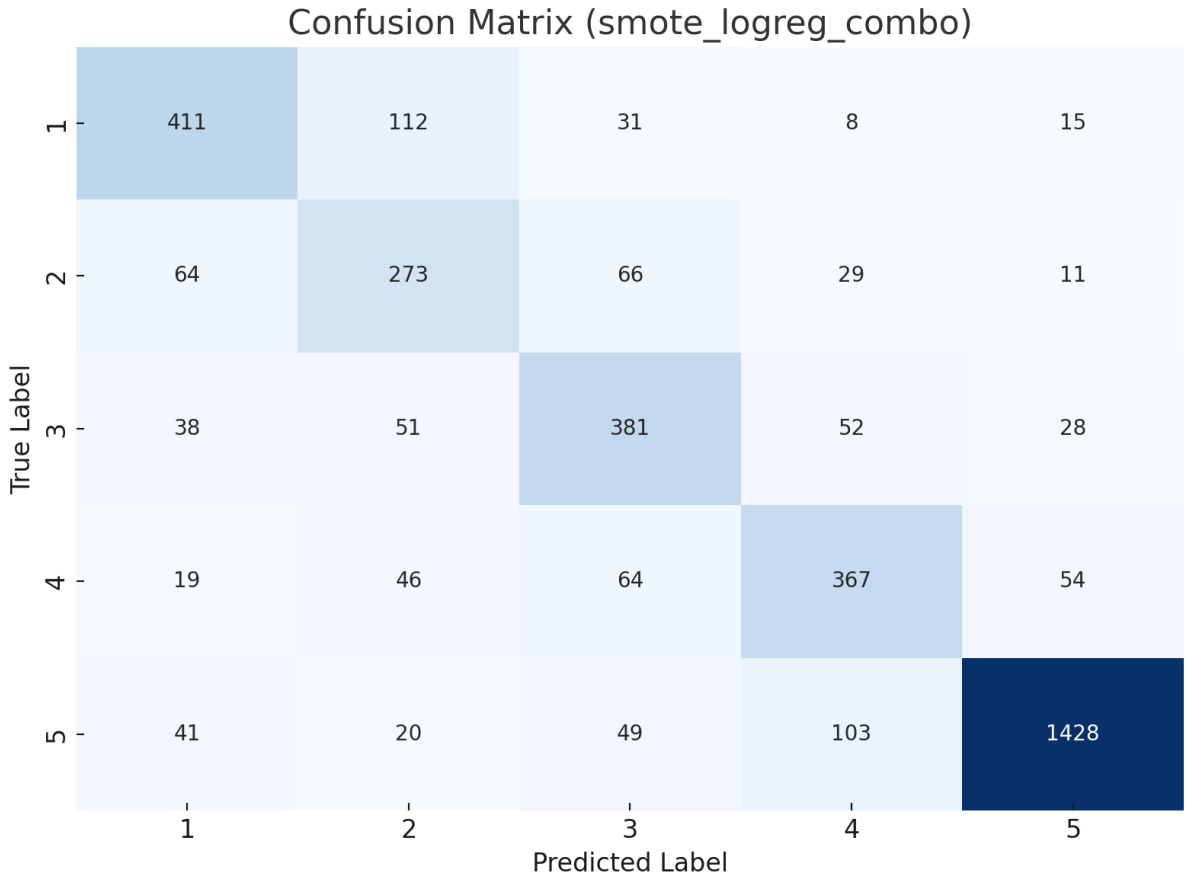
[3761/3761] Elapsed: 8s | ETA: 0s

✓ All responses received.

=== Classification report (smote_logreg_combo) ===

Клас	Precision	Recall	F1-Score	Support
1	0.717	0.712	0.715	577
2	0.544	0.616	0.578	443
3	0.645	0.693	0.668	550
4	0.657	0.667	0.662	550
5	0.930	0.870	0.899	1641
Accuracy			0.760	3761

Macro avg	0.698	0.712	0.704	3761
Weighted avg	0.770	0.760	0.764	3761



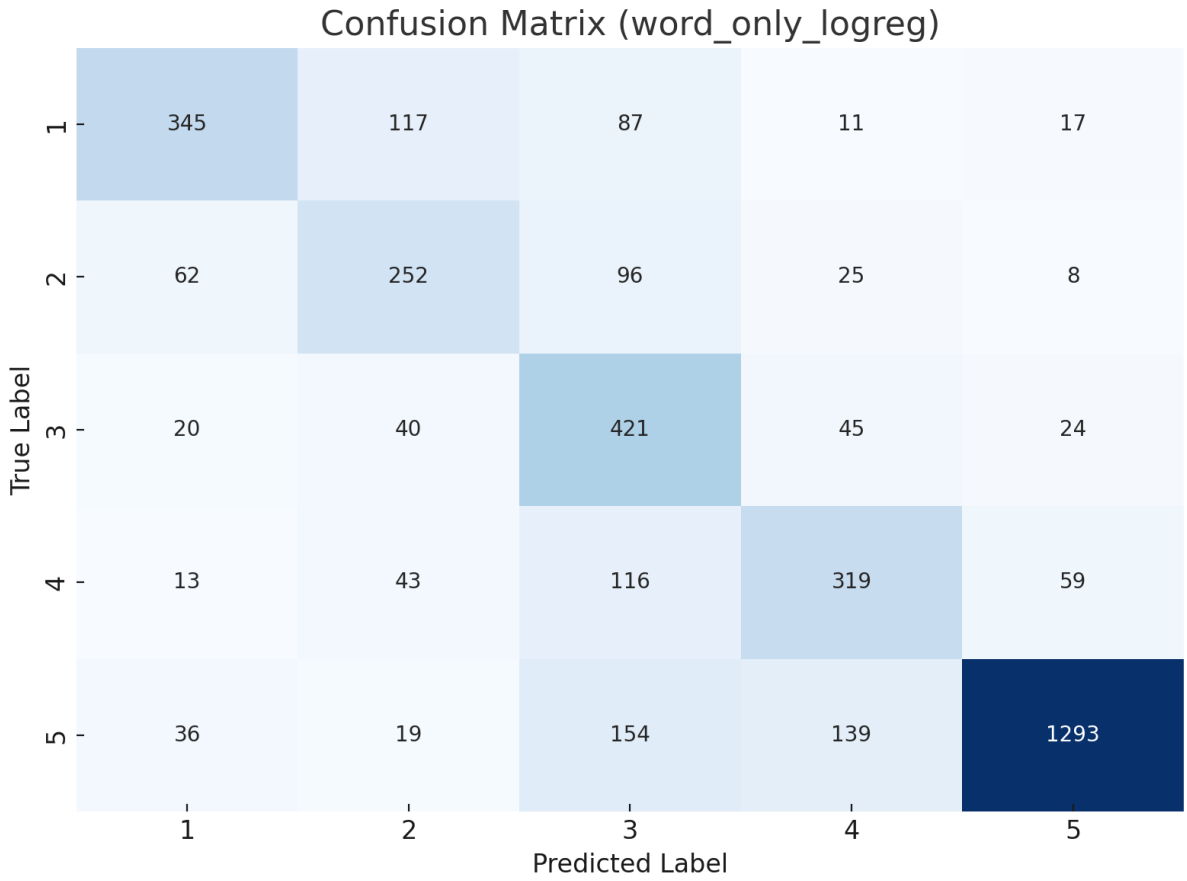
[3/9] Training word_only_logreg … done in 0.3s

[3761/3761] Elapsed: 2s | ETA: 0s

✓ All responses received.

=== Classification report (word_only_logreg) ===

Клас	Precision	Recall	F1-Score	Support
1	0.725	0.598	0.655	577
2	0.535	0.569	0.551	443
3	0.482	0.765	0.591	550
4	0.592	0.580	0.586	550
5	0.923	0.788	0.850	1641
Accuracy			0.699	3761
Macro avg	0.651	0.660	0.647	3761
Weighted avg	0.734	0.699	0.709	3761



[4/9] Training nb_combo … done in 1.2s

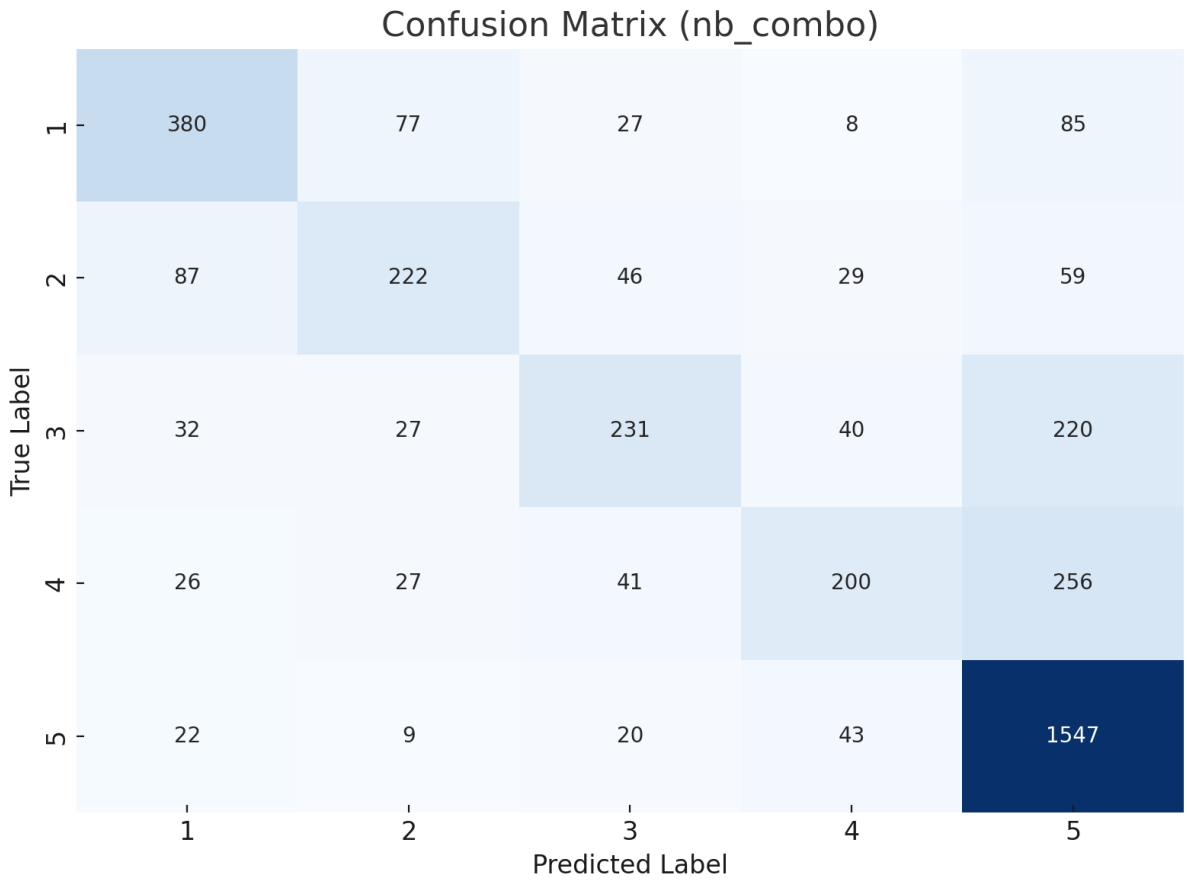
[3761/3761] Elapsed: 8s | ETA: 0s

✓ All responses received.

=== Classification report (nb_combo) ===

precision recall f1-score support

Клас	Precision	Recall	F1-Score	Support
1	0.695	0.659	0.676	577
2	0.613	0.501	0.552	443
3	0.633	0.420	0.505	550
4	0.625	0.364	0.460	550
5	0.714	0.943	0.812	1641
Accuracy			0.686	3761
Macro avg	0.656	0.577	0.601	3761
Weighted avg	0.674	0.686	0.664	3761



[5/9] ▶️ Training svc_combo … done in 82.0s

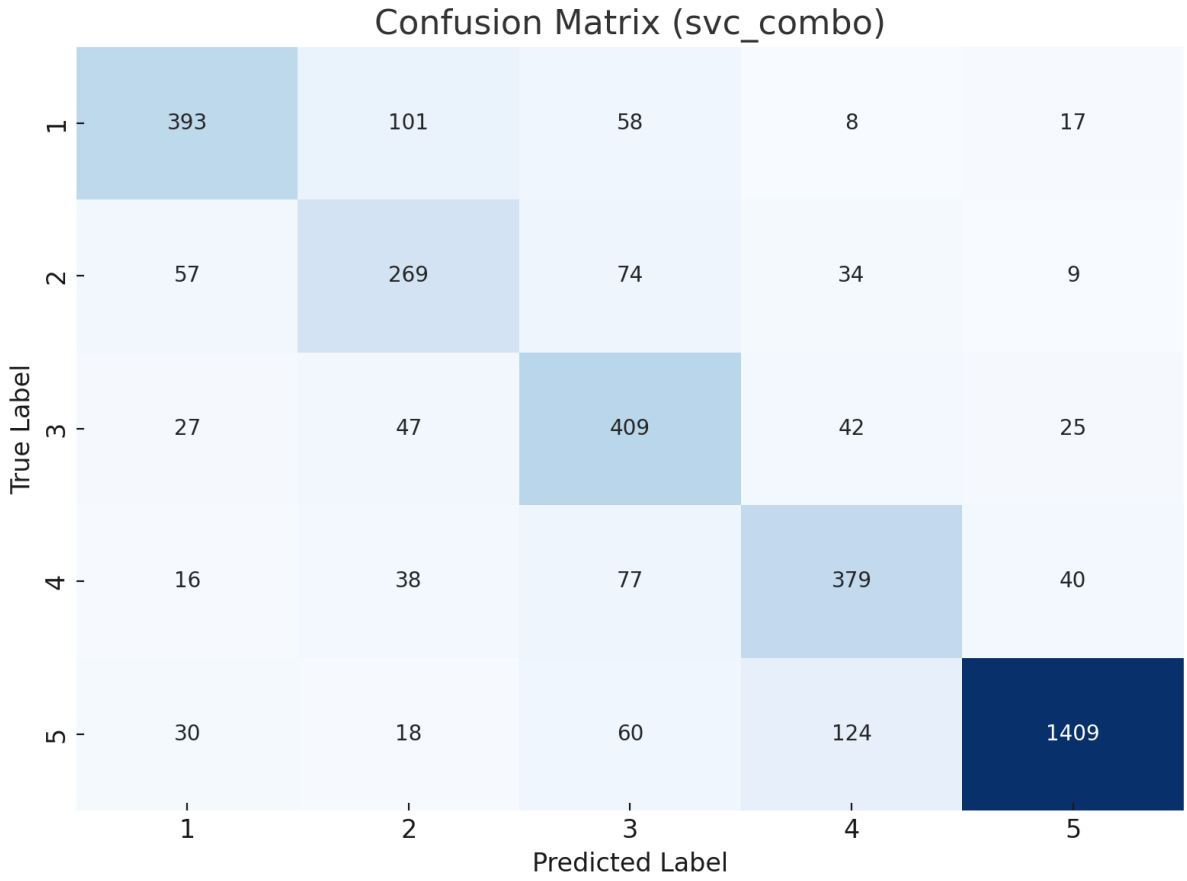
[3761/3761] Elapsed: 32s | ETA: 0s

✓ All responses received.

=== Classification report (svc_combo) ===

Клас	Precision	Recall	F1-Score	Support
1	0.751	0.681	0.715	577
2	0.569	0.607	0.587	443
3	0.603	0.744	0.666	550
4	0.646	0.689	0.667	550
5	0.939	0.859	0.897	1641
Accuracy			0.760	3761

Macro avg	0.702	0.716	0.706	3761
Weighted avg	0.775	0.760	0.765	3761



[6/9] ▶ Training knn_combo … done in 1.2s

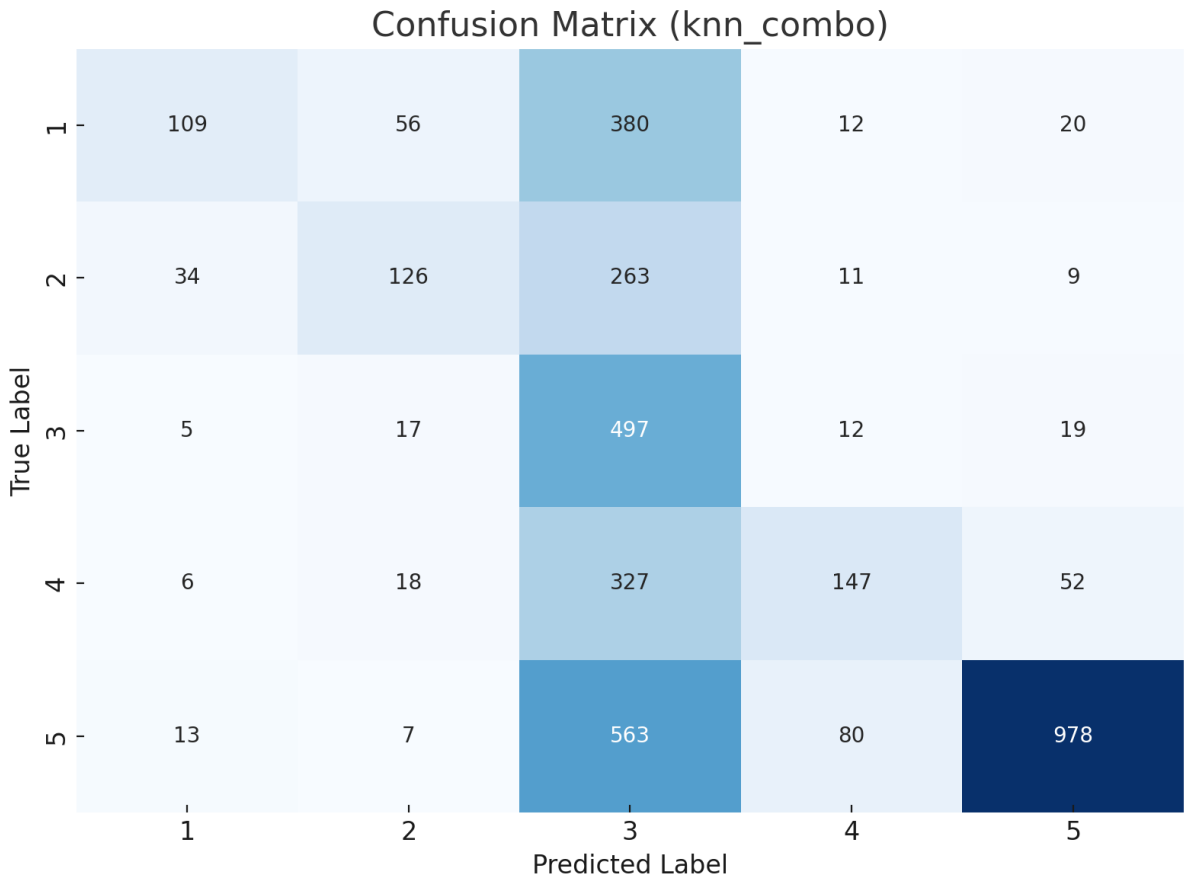
[3761/3761] Elapsed: 106s | ETA: 0s

✓ All responses received.

=== Classification report (knn_combo) ===

precision recall f1-score support

Клас	Precision	Recall	F1-Score	Support
1	0.653	0.189	0.293	577
2	0.562	0.284	0.378	443
3	0.245	0.904	0.385	550
4	0.561	0.267	0.362	550
5	0.907	0.596	0.719	1641
Accuracy			0.494	3761
Macro avg	0.586	0.448	0.428	3761
Weighted avg	0.680	0.494	0.513	3761



]

[7/9] ▶️ Training dt_combo … done in 25.7s

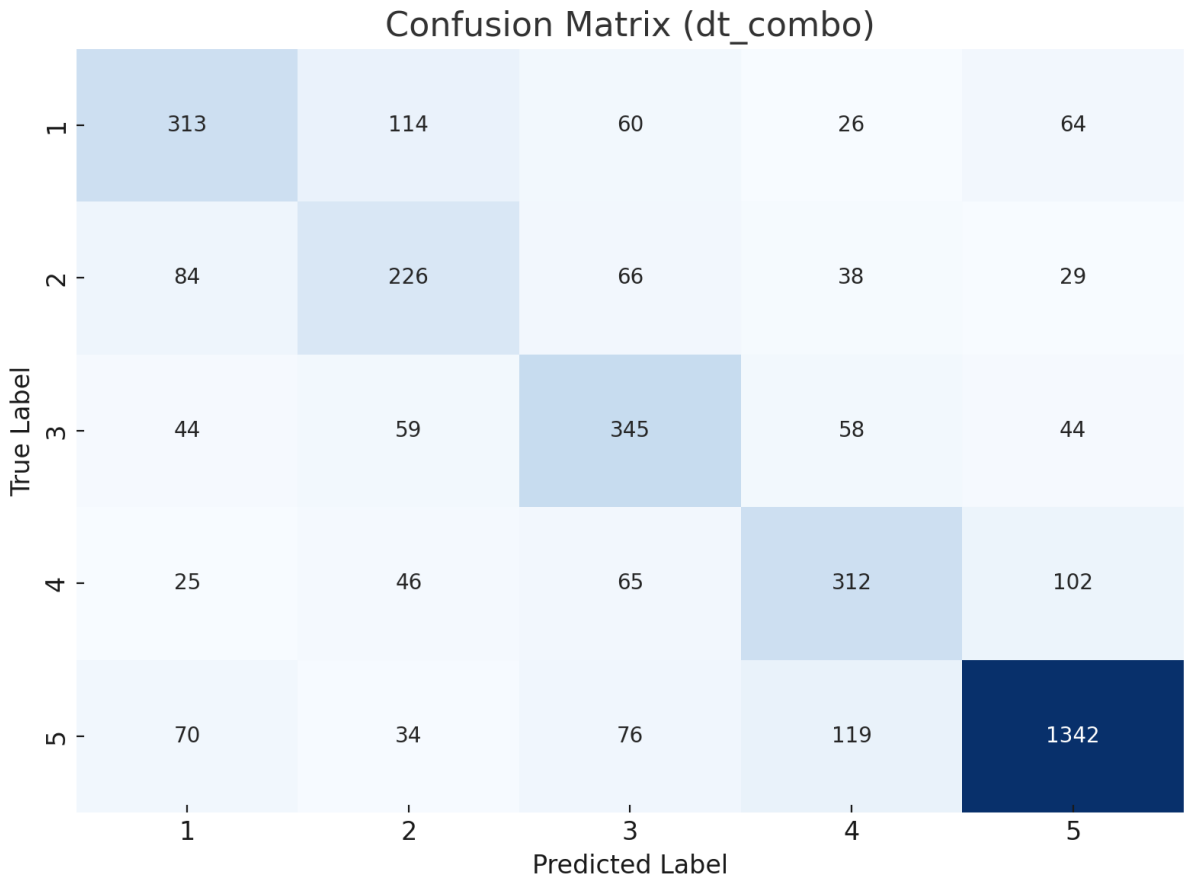
[3761/3761] Elapsed: 5s | ETA: 0s

✓ All responses received.

=== Classification report (dt_combo) ===

precision recall f1-score support

Клас	Precision	Recall	F1-Score	Support
1	0.584	0.542	0.562	577
2	0.472	0.510	0.490	443
3	0.564	0.627	0.594	550
4	0.564	0.567	0.566	550
5	0.849	0.818	0.833	1641
Accuracy			0.675	3761
Macro avg	0.607	0.613	0.609	3761
Weighted avg	0.680	0.675	0.677	3761



[8/9] ▶️ Training rf_combo … done in 51.6s

[3761/3761] Elapsed: 229s | ETA: 0s

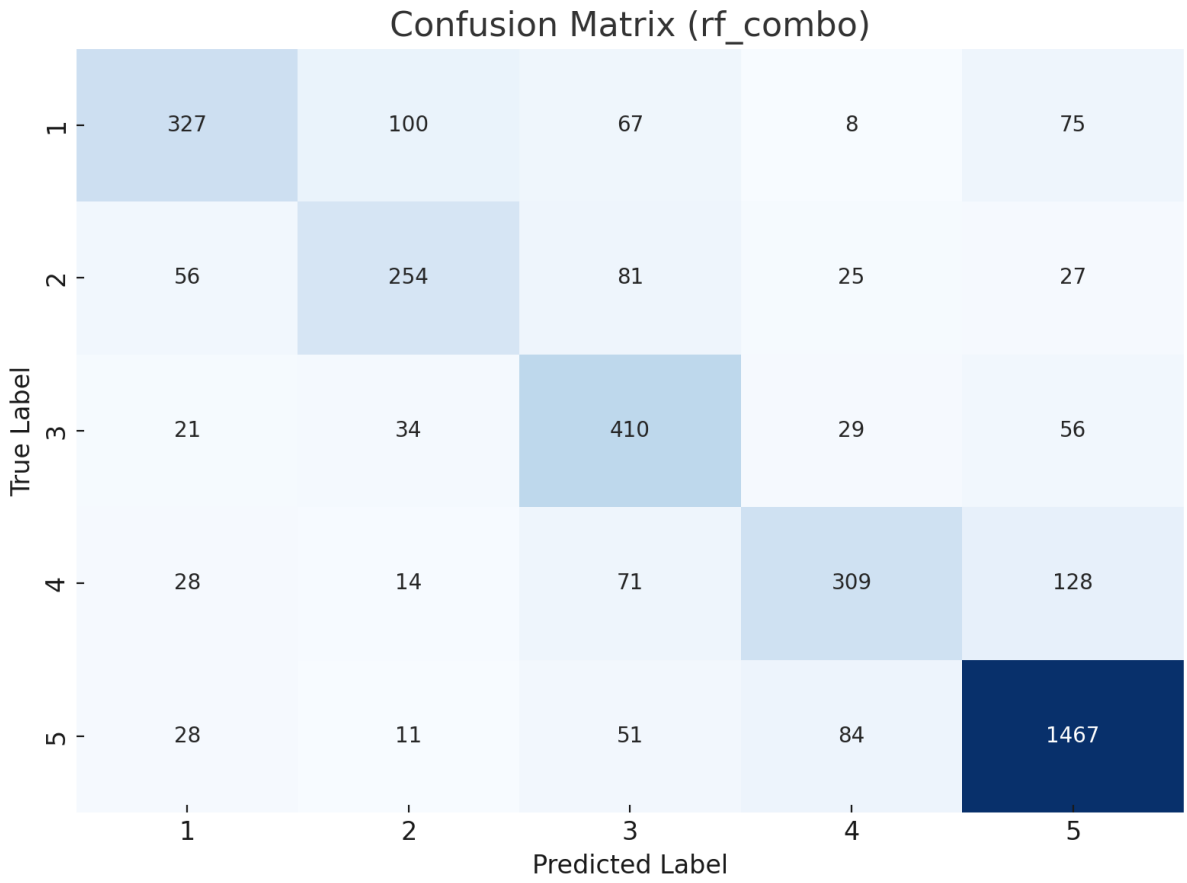
✓ All responses received.

=== Classification report (rf_combo) ===

precision recall f1-score support

Клас	Precision	Recall	F1-Score	Support
1	0.711	0.567	0.631	577
2	0.615	0.573	0.593	443
3	0.603	0.745	0.667	550

4	0.679	0.562	0.615	550
5	0.837	0.894	0.864	1641
Accuracy			0.736	3761
Macro avg	0.689	0.668	0.674	3761
Weighted avg	0.734	0.736	0.731	3761

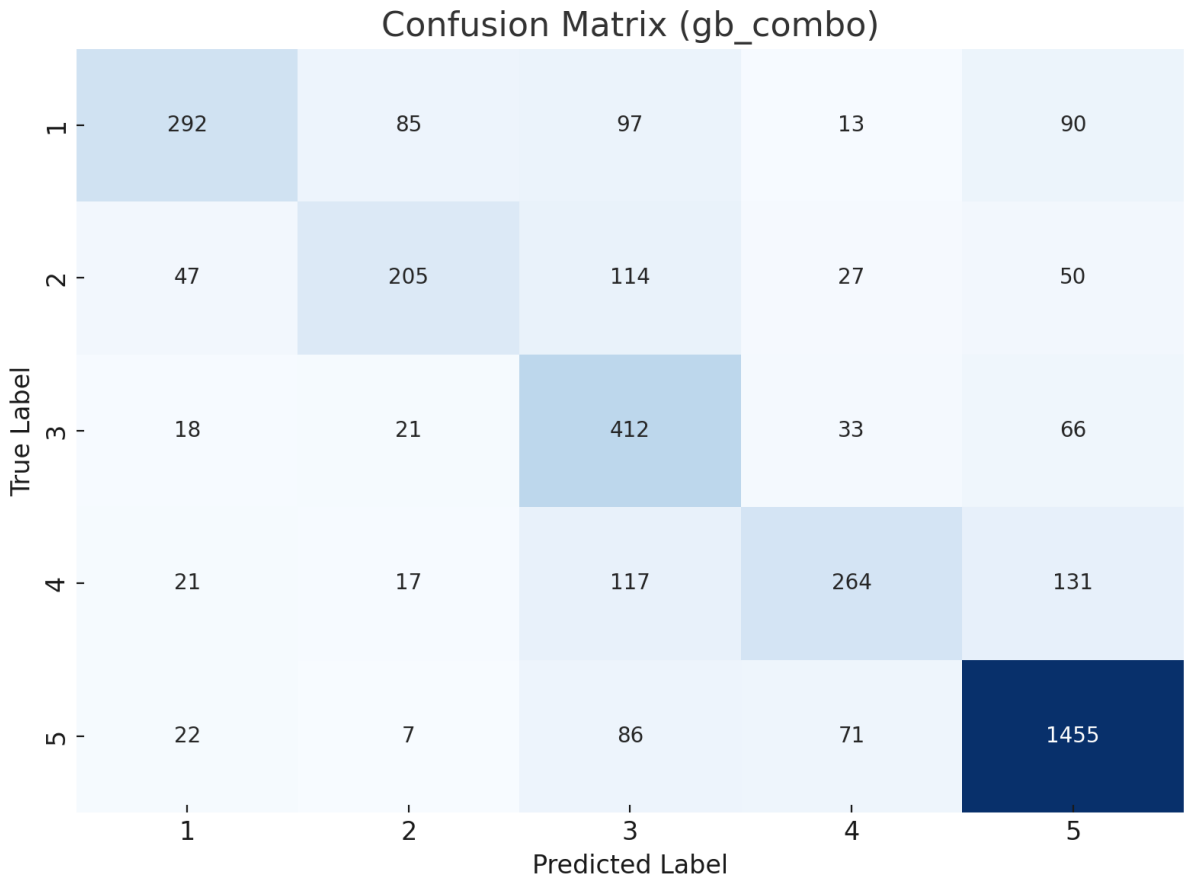


[9/9] ▶ Training gb_combo … done in 263.9s

[3761/3761] Elapsed: 7s | ETA: 0s

✓ All responses received.

Клас	Precision	Recall	F1-Score	Support
1	0.730	0.506	0.598	577
2	0.612	0.463	0.527	443
3	0.499	0.749	0.599	550
4	0.647	0.480	0.551	550
5	0.812	0.887	0.848	1641
Accuracy			0.699	3761
Macro avg	0.660	0.617	0.624	3761
Weighted avg	0.706	0.699	0.692	3761



Додаток 4 Неконтрольоване машинне навчання

=== LOGREG_COMBO ===

	precision	recall	f1-score	support
1	0.603	0.553	0.577	618
2	0.063	0.047	0.054	107
3	0.114	0.081	0.095	172
4	0.201	0.402	0.268	363
5	0.853	0.772	0.810	2501
accuracy			0.648	3761
macro avg	0.367	0.371	0.361	3761
weighted avg	0.693	0.648	0.666	3761

=== SMOTE_LOGREG_COMBO ===

	precision	recall	f1-score	support
1	0.561	0.639	0.598	618
2	0.071	0.121	0.089	107
3	0.100	0.186	0.130	172
4	0.199	0.229	0.213	363
5	0.863	0.738	0.796	2501
accuracy			0.630	3761
macro avg	0.359	0.383	0.365	3761
weighted avg	0.692	0.630	0.656	3761

=== WORD_ONLY_LOGREG ===

	precision	recall	f1-score	support
1	0.319	0.777	0.452	618
2	0.054	0.103	0.071	107
3	0.159	0.099	0.122	172
4	0.247	0.309	0.274	363
5	0.918	0.548	0.686	2501
accuracy			0.529	3761
macro avg	0.339	0.367	0.321	3761
weighted avg	0.695	0.529	0.565	3761

=== NB_COMBO ===

	precision	recall	f1-score	support
1	0.573	0.615	0.593	618
2	0.000	0.000	0.000	107
3	1.000	0.012	0.023	172
4	0.330	0.099	0.153	363
5	0.783	0.936	0.853	2501
accuracy			0.733	3761
macro avg	0.537	0.332	0.324	3761
weighted avg	0.693	0.733	0.680	3761

=== SVC_COMBO ===

	precision	recall	f1-score	support
1	0.541	0.647	0.590	618
2	0.099	0.093	0.096	107
3	0.147	0.174	0.160	172
4	0.244	0.289	0.264	363
5	0.860	0.786	0.822	2501
accuracy			0.668	3761
macro avg	0.378	0.398	0.386	3761
weighted avg	0.694	0.668	0.679	3761

=== KNN_COMBO ===

	precision	recall	f1-score	support
1	0.583	0.261	0.360	618
2	0.000	0.000	0.000	107
3	0.105	0.035	0.052	172
4	0.317	0.072	0.117	363
5	0.718	0.958	0.821	2501
accuracy			0.688	3761
macro avg	0.345	0.265	0.270	3761
weighted avg	0.609	0.688	0.619	3761

=== DT_COMBO ===

	precision	recall	f1-score	support
1	0.466	0.448	0.457	618
2	0.045	0.065	0.053	107
3	0.110	0.134	0.120	172
4	0.171	0.204	0.186	363
5	0.805	0.762	0.783	2501
accuracy			0.608	3761
macro avg	0.319	0.323	0.320	3761
weighted avg	0.635	0.608	0.621	3761

=== RF_COMBO ===

	precision	recall	f1-score	support
1	0.571	0.518	0.543	618
2	0.000	0.000	0.000	107
3	0.171	0.041	0.066	172
4	0.262	0.077	0.119	363
5	0.767	0.933	0.842	2501
accuracy			0.715	3761
macro avg	0.354	0.314	0.314	3761
weighted avg	0.637	0.715	0.663	3761

=== GB_COMBO ===

precision recall f1-score support

1 0.598 0.434 0.503 618
2 0.059 0.009 0.016 107
3 0.182 0.012 0.022 172
4 0.329 0.069 0.114 363
5 0.748 0.959 0.840 2501

accuracy 0.717 3761
macro avg 0.383 0.297 0.299 3761
weighted avg 0.637 0.717 0.654 3761

Тренування на оцінці користувача, тестування за індексом

=== LOGREG_COMBO ===
fit in 124.3s
precision recall f1-score support

1 0.520 0.555 0.537 577
2 0.263 0.361 0.304 438
3 0.246 0.029 0.052 552
4 0.325 0.239 0.275 549
5 0.696 0.879 0.777 1644

accuracy 0.551 3760
macro avg 0.410 0.412 0.389 3760
weighted avg 0.498 0.551 0.505 3760

=== SMOTE_LOGREG_COMBO ===

fit in 5.7s

	precision	recall	f1-score	support
1	0.563	0.666	0.610	577
2	0.317	0.130	0.184	438
3	0.375	0.230	0.285	552
4	0.282	0.215	0.244	549
5	0.710	0.925	0.803	1644
accuracy			0.587	3760
macro avg	0.449	0.433	0.425	3760
weighted avg	0.530	0.587	0.544	3760

=== WORD_ONLY_LOGREG ===

fit in 14.7s

	precision	recall	f1-score	support
1	0.613	0.263	0.368	577
2	0.400	0.018	0.035	438
3	0.165	0.074	0.102	552
4	0.249	0.202	0.223	549
5	0.563	0.958	0.709	1644
accuracy			0.502	3760
macro avg	0.398	0.303	0.288	3760

weighted avg 0.447 0.502 0.418 3760

=== NB_COMBO ===

fit in 1.2s

	precision	recall	f1-score	support
1	0.564	0.638	0.599	577
2	0.000	0.000	0.000	438
3	0.000	0.000	0.000	552
4	0.210	0.046	0.075	549
5	0.545	0.989	0.702	1644
accuracy			0.537	3760
macro avg	0.264	0.334	0.275	3760
weighted avg	0.355	0.537	0.410	3760

=== SVC_COMBO ===

fit in 108.1s

	precision	recall	f1-score	support
1	0.540	0.679	0.602	577
2	0.304	0.064	0.106	438
3	0.196	0.080	0.113	552
4	0.289	0.230	0.256	549
5	0.653	0.906	0.759	1644

accuracy			0.553	3760
macro avg	0.396	0.392	0.367	3760
weighted avg	0.475	0.553	0.490	3760

=== KNN_COMBO ===

fit in 1.2s

	precision	recall	f1-score	support
1	0.452	0.258	0.329	577
2	0.333	0.007	0.013	438
3	0.190	0.022	0.039	552
4	0.233	0.038	0.066	549
5	0.490	0.974	0.652	1644
accuracy			0.475	3760
macro avg	0.340	0.260	0.220	3760
weighted avg	0.384	0.475	0.353	3760

=== DT_COMBO ===

fit in 26.8s

	precision	recall	f1-score	support
1	0.447	0.442	0.445	577
2	0.210	0.066	0.101	438
3	0.227	0.098	0.137	552
4	0.255	0.199	0.223	549

5 0.586 0.851 0.694 1644

accuracy 0.491 3760
macro avg 0.345 0.331 0.320 3760
weighted avg 0.420 0.491 0.436 3760

=== RF_COMBO ===
fit in 80.7s

	precision	recall	f1-score	support
1	0.551	0.520	0.535	577
2	0.182	0.005	0.009	438
3	0.250	0.018	0.034	552
4	0.212	0.046	0.075	549
5	0.528	0.978	0.686	1644
accuracy			0.517	3760
macro avg	0.345	0.313	0.268	3760
weighted avg	0.404	0.517	0.399	3760

=== GB_COMBO ===
fit in 247.3s

	precision	recall	f1-score	support
1	0.554	0.419	0.477	577
2	0.353	0.014	0.026	438

3	**0.000**	**0.000**	**0.000**	**552**
4	**0.250**	**0.040**	**0.069**	**549**
5	**0.507**	**0.989**	**0.670**	**1644**
accuracy			**0.504**	**3760**
macro avg	**0.333**	**0.292**	**0.249**	**3760**
weighted avg	**0.384**	**0.504**	**0.380**	**376**

Додаток 5 Великі мовні моделі zero-shot тестування

ChatGPT

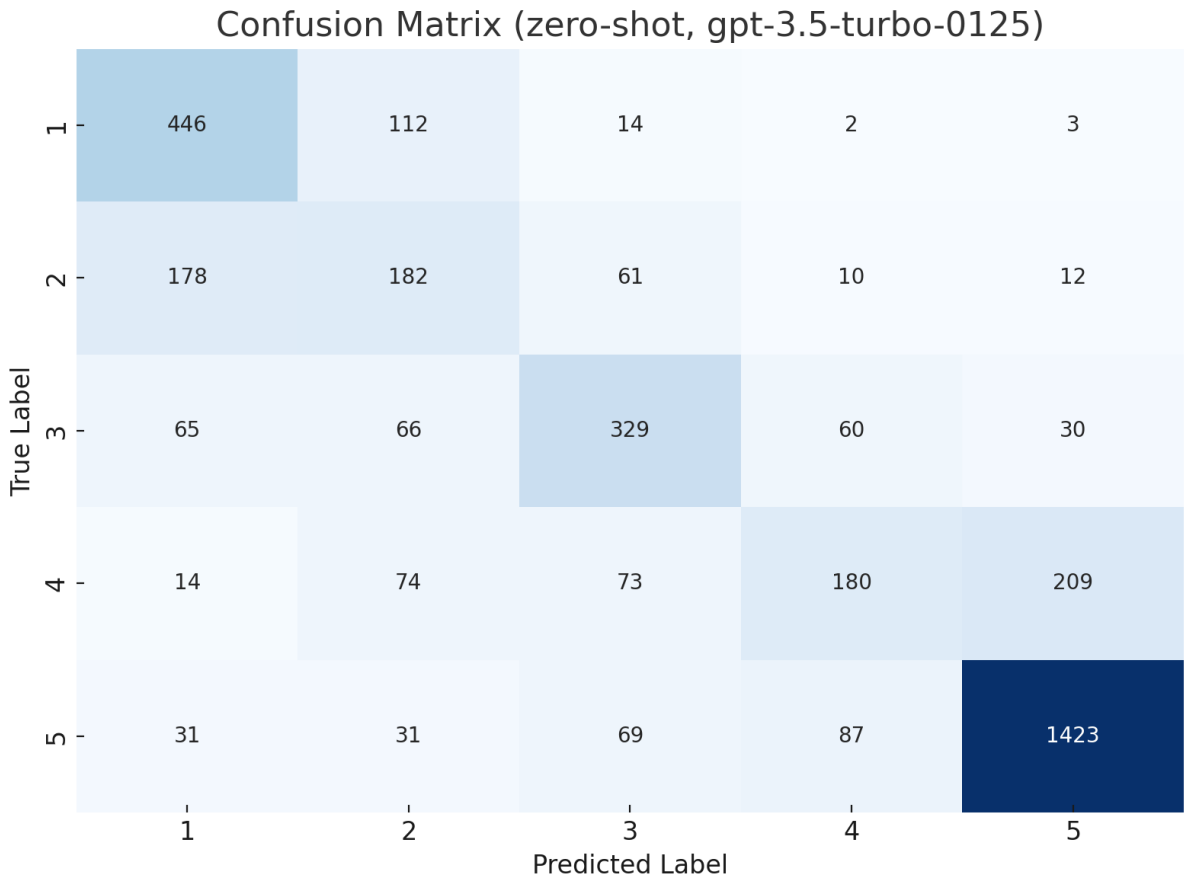
GPT3.5

3761/3761] Elapsed: 444s | ETA: 0s

✓ All responses received.

=== Classification report (zero-shot, gpt-3.5-turbo-0125, test_dataset.csv) ===

Клас	Precision	Recall	F1-Score	Support
1	0.608	0.773	0.680	577
2	0.391	0.411	0.401	443
3	0.603	0.598	0.600	550
4	0.531	0.327	0.405	550
5	0.849	0.867	0.858	1641
Accuracy			0.681	3761
Macro avg	0.596	0.595	0.589	3761
Weighted avg	0.675	0.681	0.673	3761



GPT4.1

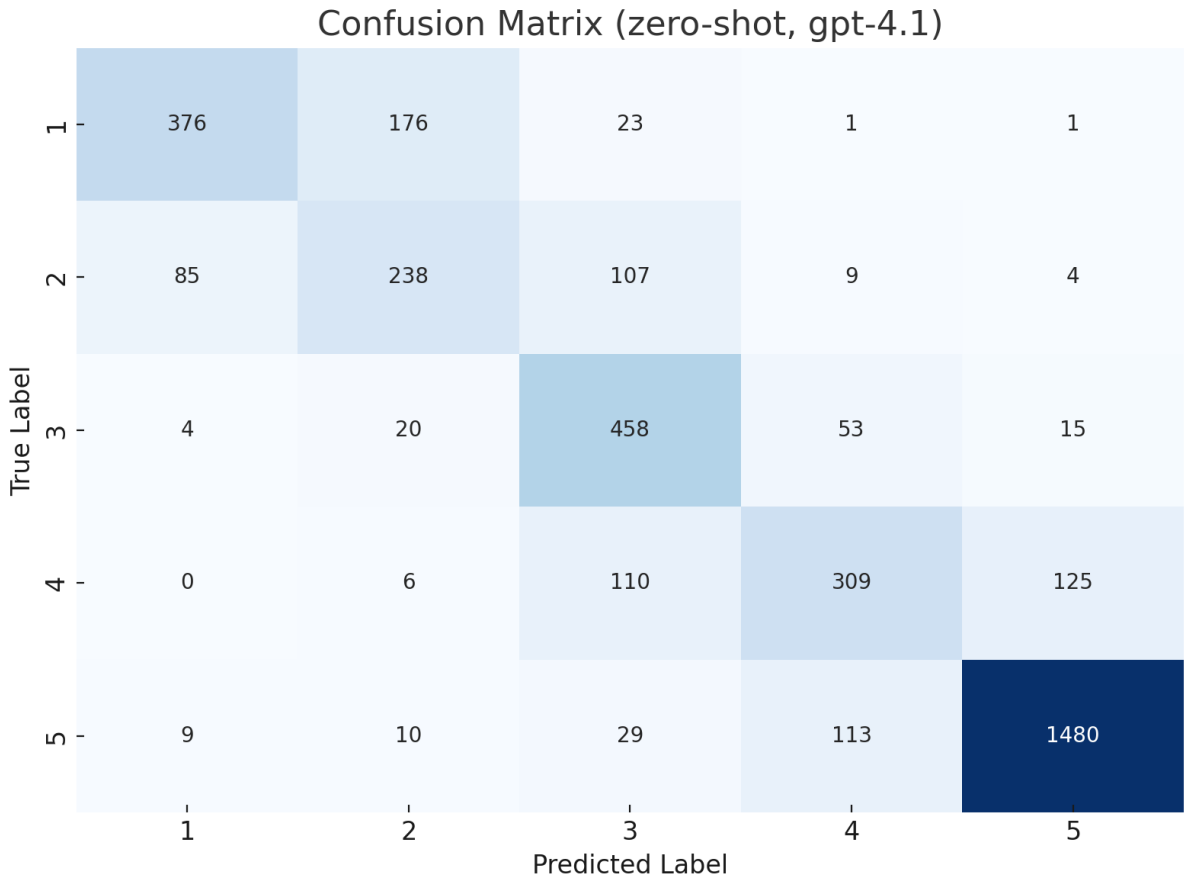
[3761/3761] Elapsed: 1846s | ETA: 0s

✓ All responses received.

=== Classification report (zero-shot, gpt-4.1) ===

Клас	Precision	Recall	F1-Score	Support
1	0.793	0.652	0.716	577
2	0.529	0.537	0.533	443
3	0.630	0.833	0.717	550
4	0.637	0.562	0.597	550
5	0.911	0.902	0.906	1641

Accuracy			0.761	3761
Macro avg	0.700	0.697	0.694	3761
Weighted avg	0.767	0.761	0.760	3761



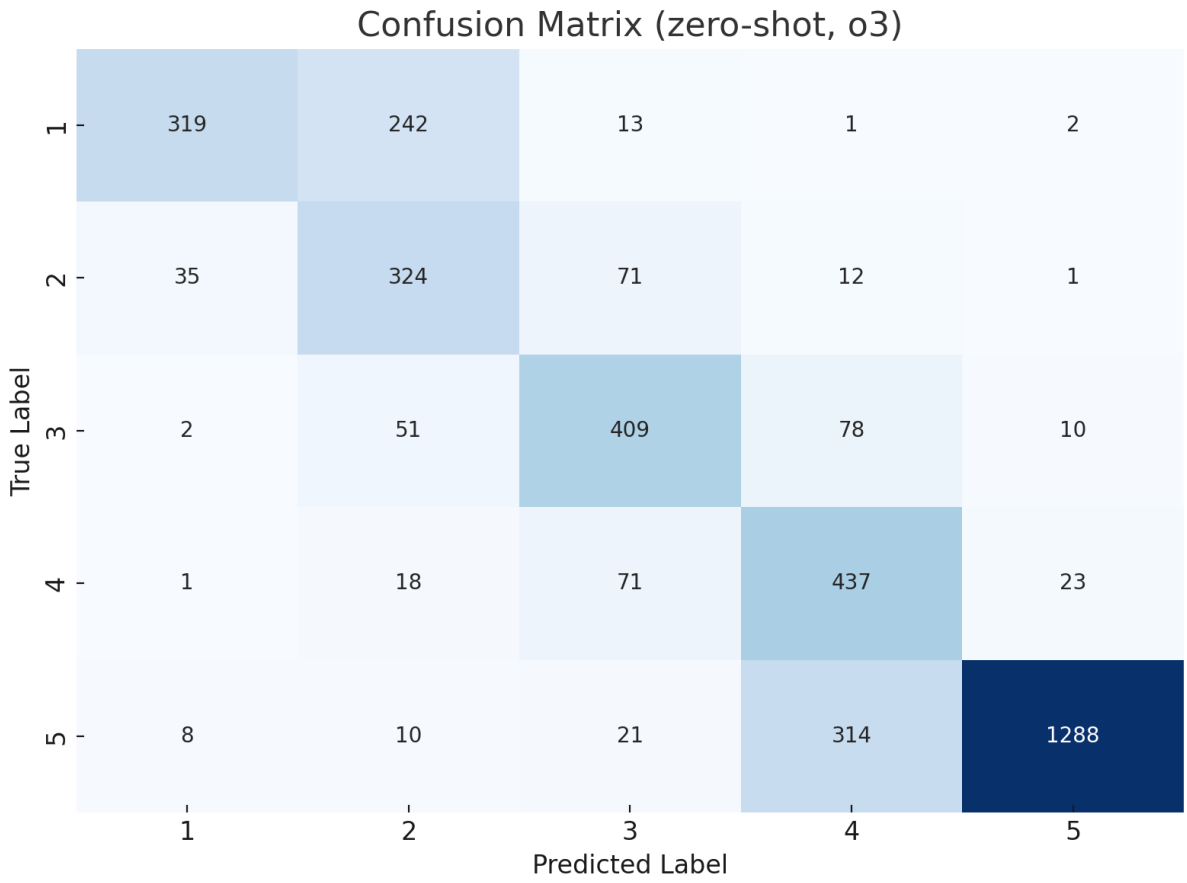
GPT o3

[3761/3761] Elapsed: 7018 | ETA: 0s

✓ All responses received.

=== Classification report (zero-shot, o3) ===

Клас	Precision	Recall	F1-Score	Support
1	0.874	0.553	0.677	577
2	0.502	0.731	0.596	443
3	0.699	0.744	0.721	550
4	0.519	0.795	0.628	550
5	0.973	0.785	0.869	1641
Accuracy			0.738	3761
Macro avg	0.713	0.721	0.698	3761
Weighted avg	0.796	0.738	0.750	3761



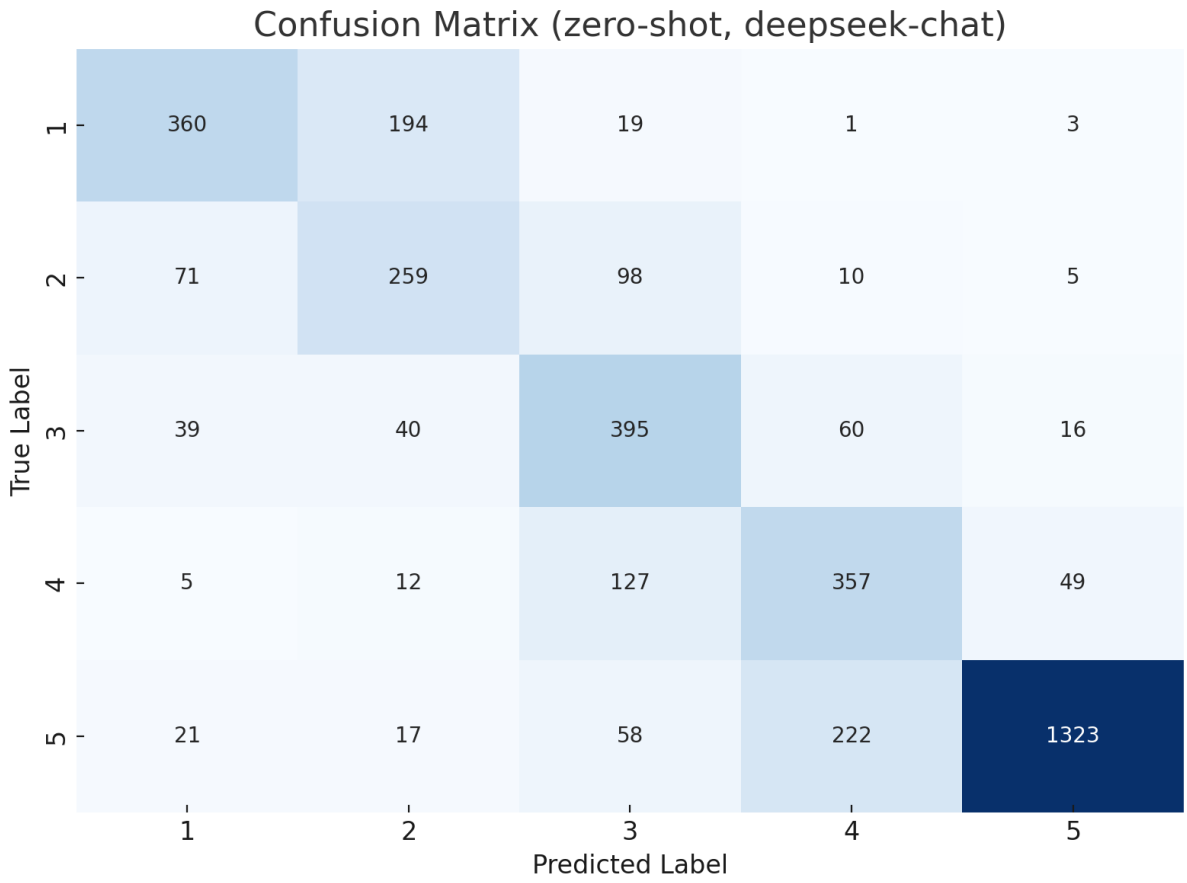
DeepSeek

deepseek-chat / deepseek-v3

[3761/3761] Elapsed: 574s | ETA: 0s

✓ All responses received.

Клас	Precision	Recall	F1-Score	Support
1	0.726	0.624	0.671	577
2	0.496	0.585	0.537	443
3	0.567	0.718	0.634	550
4	0.549	0.649	0.595	550
5	0.948	0.806	0.871	1641
Accuracy			0.716	3761
Macro avg	0.657	0.676	0.662	3761
Weighted avg	0.746	0.716	0.726	3761



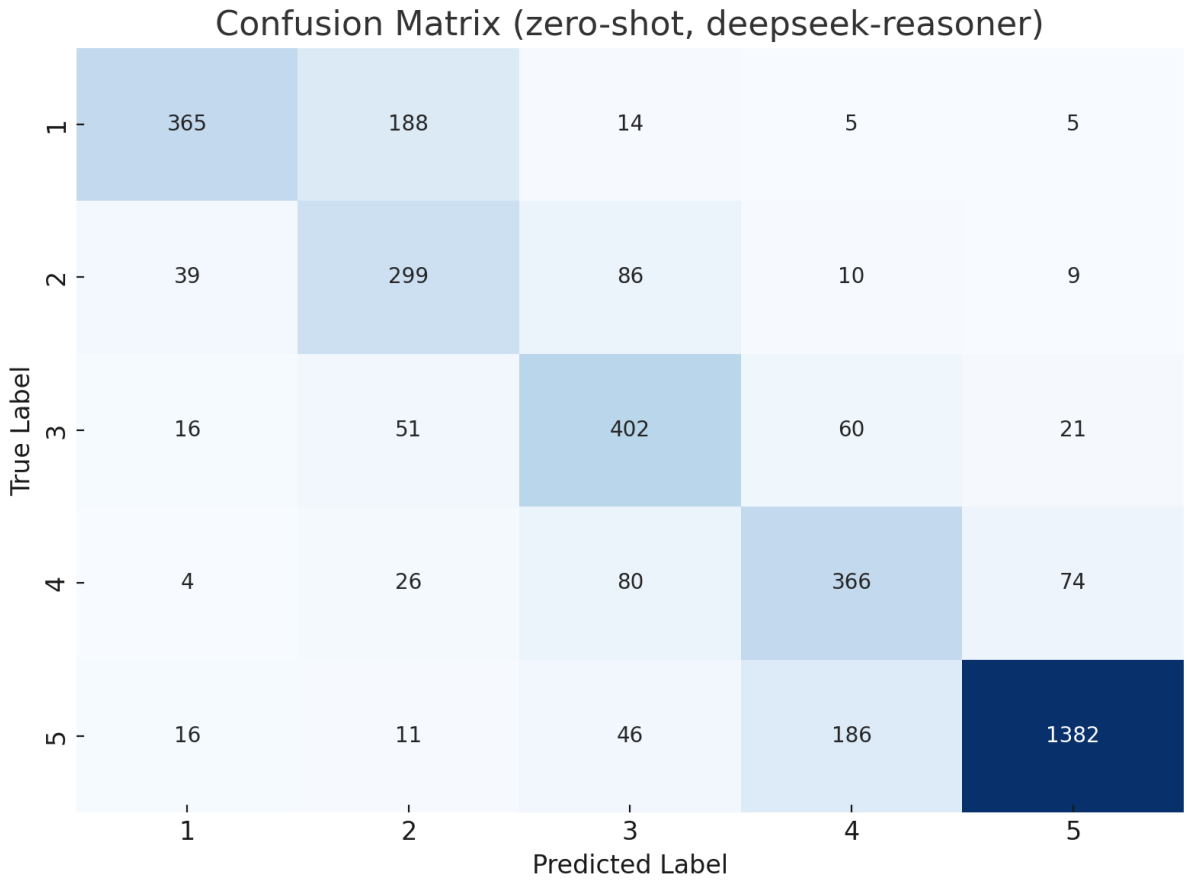
deepseek-r1/deepseek-reasoner

[3761/3761] Elapsed: 2684s | ETA: 0s

✓ All responses received.

Клас	Precision	Recall	F1-Score	Support
1	0.830	0.633	0.718	577
2	0.520	0.675	0.587	443
3	0.640	0.731	0.683	550
4	0.584	0.665	0.622	550
5	0.927	0.842	0.883	1641

Accuracy			0.748	3761
Macro avg	0.700	0.709	0.698	3761
Weighted avg	0.772	0.748	0.755	3761



Додаток 6 Великі мовні моделі fine-tuned тестування

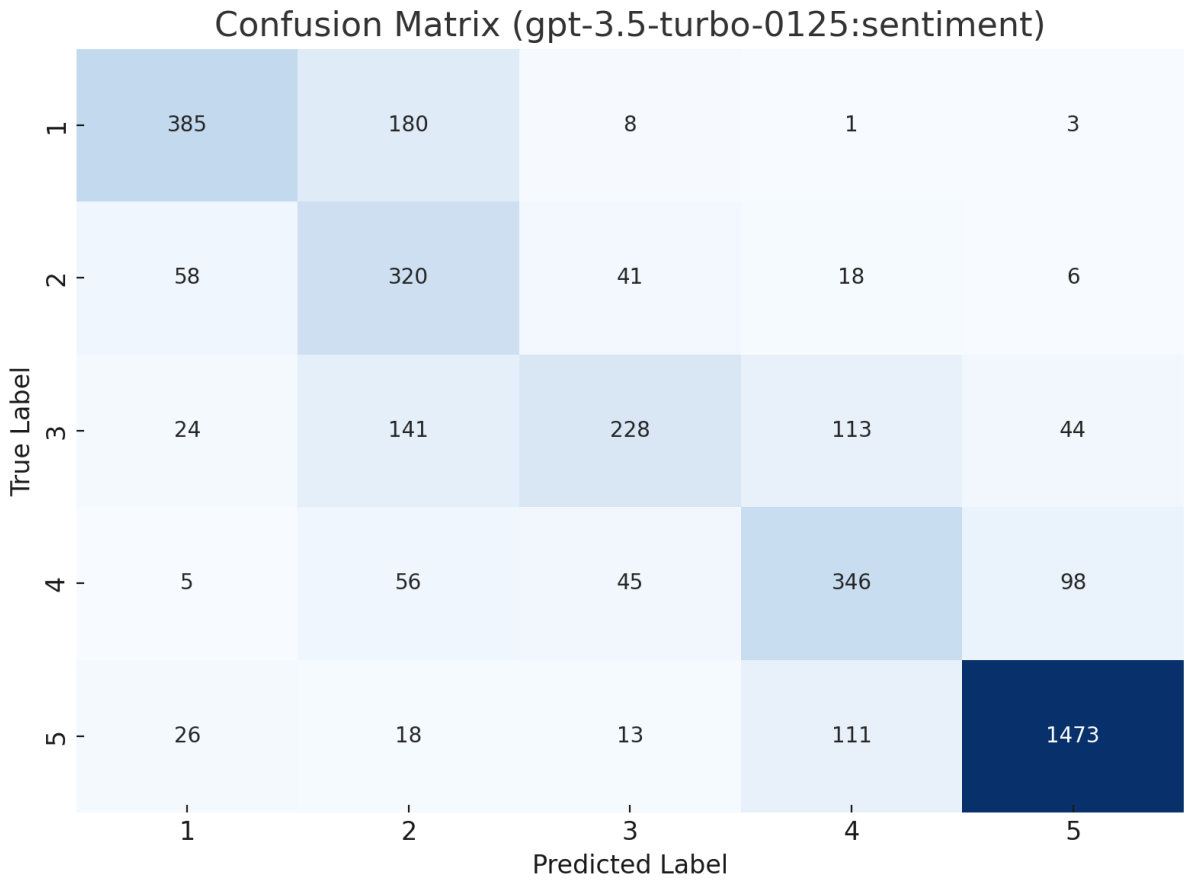
Донавчання на 100 відгуках

[3761/3761] Elapsed: 442s | ETA: 0s

✓ All responses received.

=== Classification report, gpt-3.5-turbo-0125:sentiment, test_dataset.csv) ===

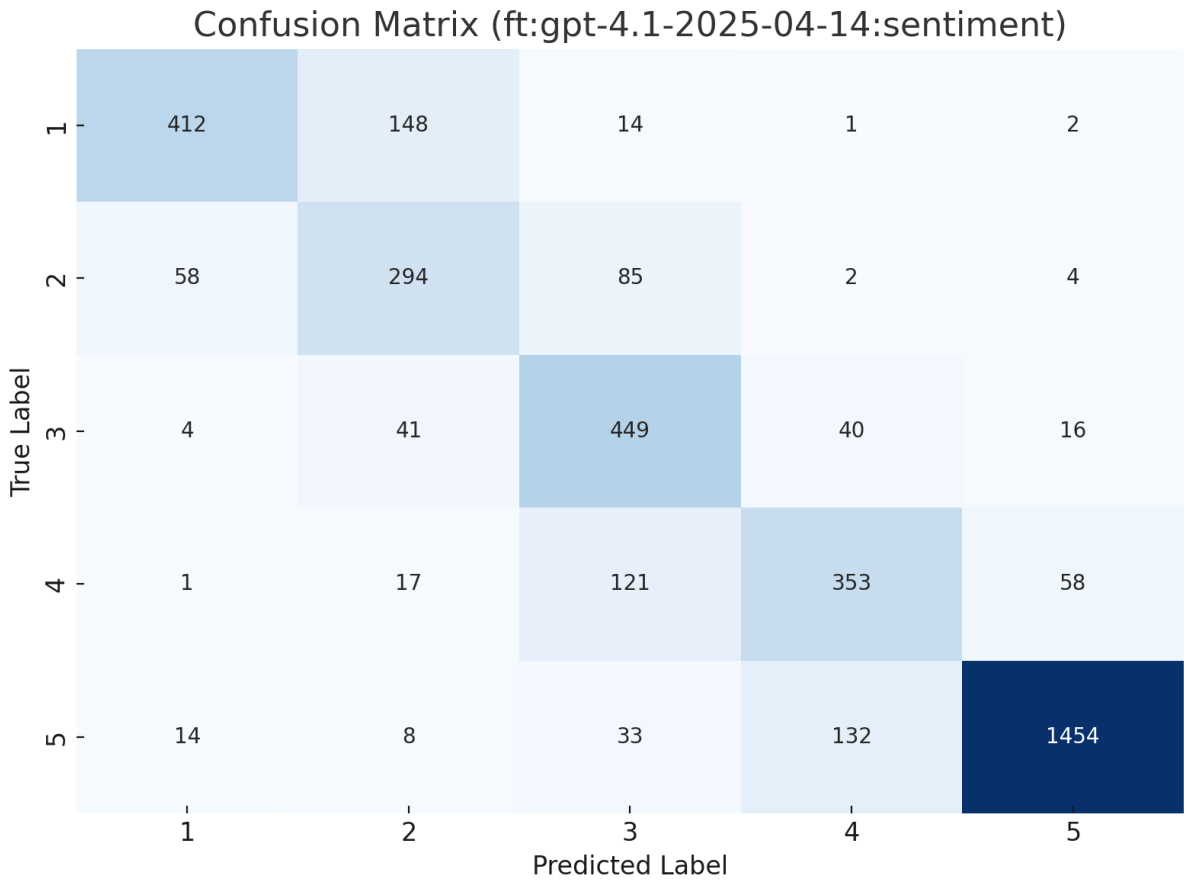
Клас	Precision	Recall	F1-Score	Support
1	0.773	0.667	0.716	577
2	0.448	0.722	0.553	443
3	0.681	0.415	0.515	550
4	0.587	0.629	0.608	550
5	0.907	0.898	0.902	1641
Accuracy			0.732	3761
Macro avg	0.679	0.666	0.659	3761
Weighted avg	0.753	0.732	0.733	3761



=== Classification report (gpt-4.1-2025-04-14:sentiment) ===

Клас	Precision	Recall	F1-Score	Support
1	0.843	0.714	0.773	577
2	0.579	0.664	0.618	443
3	0.640	0.816	0.717	550
4	0.669	0.642	0.655	550
5	0.948	0.886	0.916	1641
Accuracy			0.788	3761
Macro avg	0.735	0.744	0.736	3761
Weighted	0.802	0.788	0.792	3761

avg				



Донавчання на 1000 відгуків

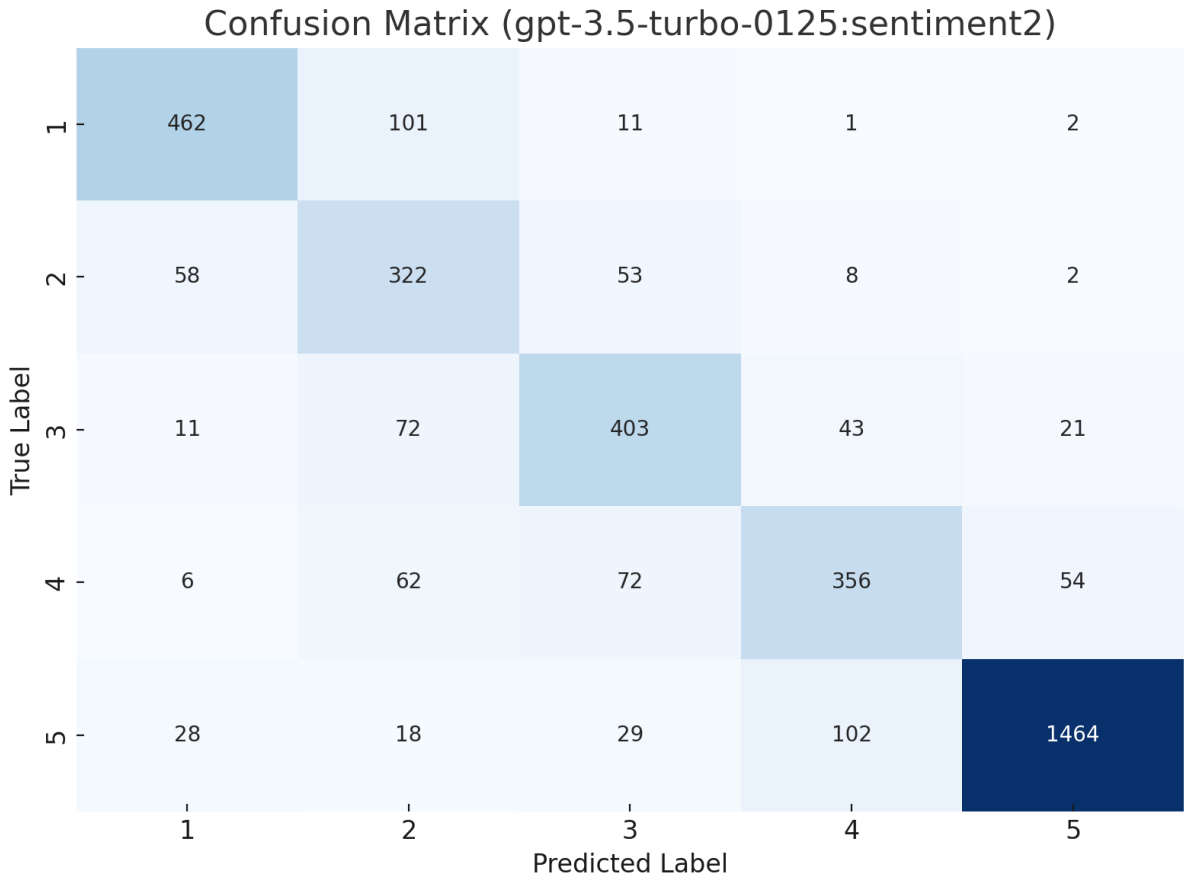
[3761/3761] Elapsed: 446s | ETA: 0s

✓ All responses received.

=== Classification report (gpt-3.5-turbo-0125:sentiment2, test_dataset.csv) ===

Клас	Precision	Recall	F1-Score	Support
1	0.818	0.801	0.809	577

2	0.560	0.727	0.633	443
3	0.710	0.733	0.721	550
4	0.698	0.647	0.672	550
5	0.949	0.892	0.920	1641
Accuracy			0.800	3761
Macro avg	0.747	0.760	0.751	3761
Weighted avg	0.811	0.800	0.804	3761

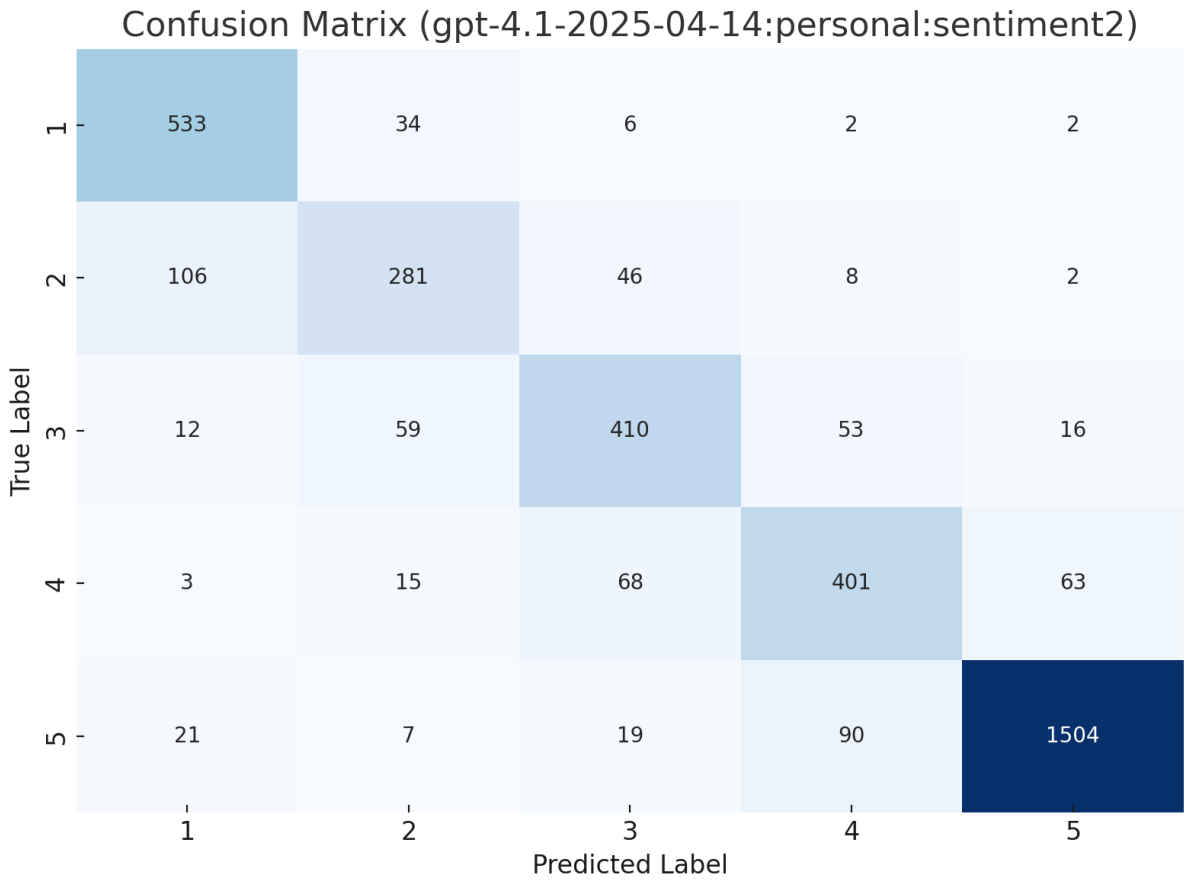


[3761/3761] Elapsed: 204s | ETA: 0s

✓ All responses received.

=== Classification report (gpt-4.1-2025-04-14:personal:sentiment2) ===

Клас	Precision	Recall	F1-Score	Support
1	0.790	0.924	0.851	577
2	0.710	0.634	0.670	443
3	0.747	0.745	0.746	550
4	0.724	0.729	0.726	550
5	0.948	0.917	0.932	1641
Accuracy			0.832	3761
Macro avg	0.784	0.790	0.785	3761
Weighted avg	0.833	0.832	0.831	3761



Додаток 7 Вплив мови запиту на результати класифікації великих мовних моделей.

Zero-shot

gpt-3.5

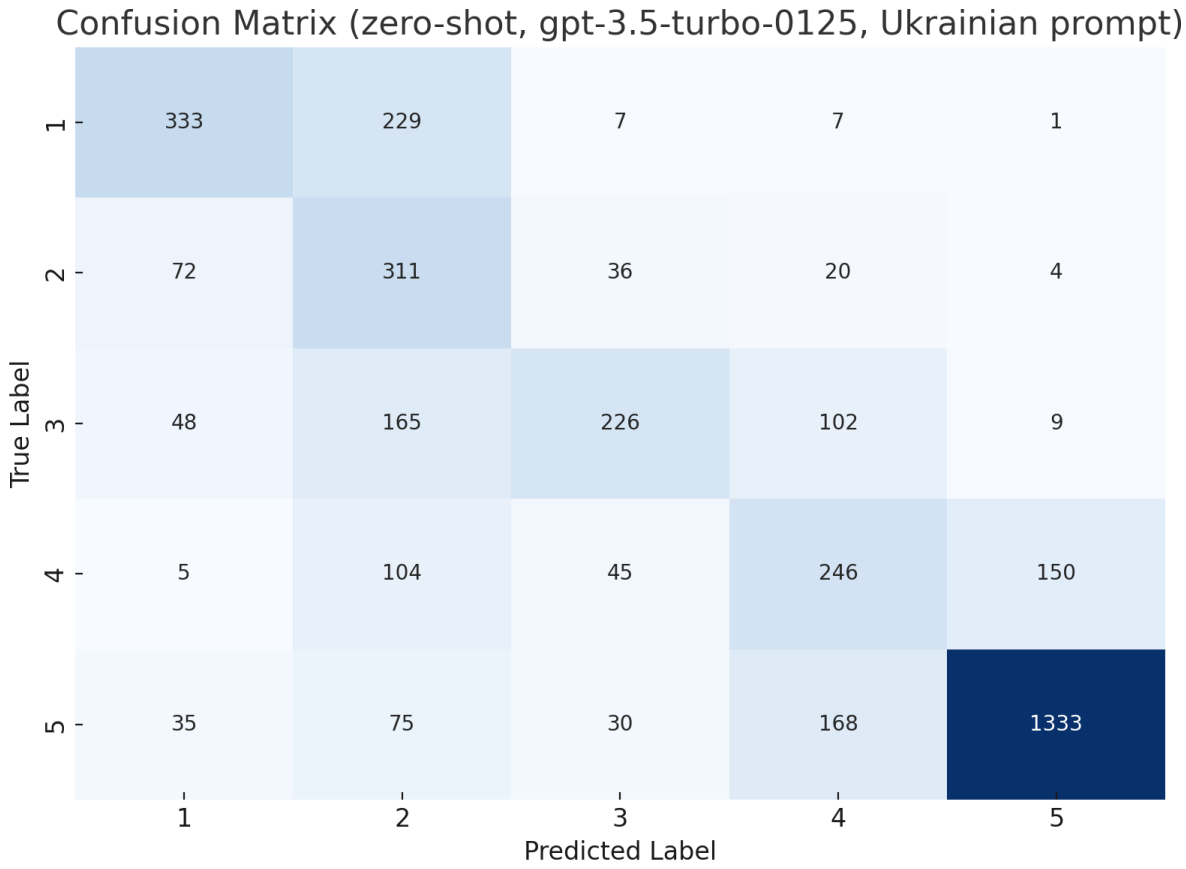
[3761/3761] Elapsed: 756s | ETA: 0s

✓ All responses received.

=== Classification report (zero-shot, gpt-3.5-turbo-0125, test_dataset.csv, ukrainian prompt) ===

Клас	Precision	Recall	F1-Score	Support
1	0.675	0.577	0.622	577

2	0.352	0.702	0.469	443
3	0.657	0.411	0.506	550
4	0.453	0.447	0.450	550
5	0.890	0.812	0.850	1641
Accuracy			0.651	3761
Macro avg	0.606	0.590	0.579	3761
Weighted avg	0.696	0.651	0.661	3761



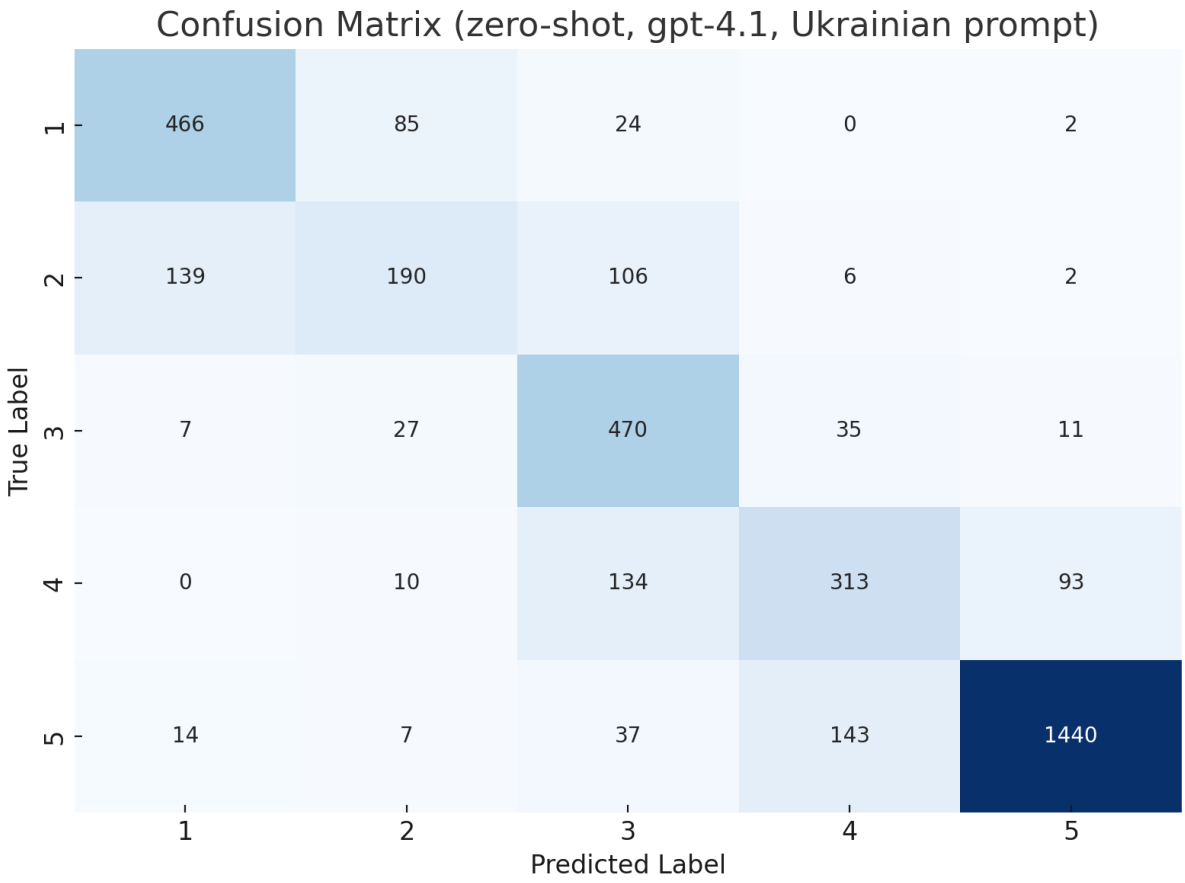
gpt-4.1

[3761/3761] Elapsed: 1853s | ETA: 0s

✓ All responses received.

=== Classification report (zero-shot, gpt-4.1-2025-04-14, ukrainian prompt) ===

Клас	Precision	Recall	F1-Score	Support
1	0.744	0.808	0.775	577
2	0.596	0.429	0.499	443
3	0.610	0.855	0.712	550
4	0.630	0.569	0.598	550
5	0.930	0.878	0.903	1641
Accuracy			0.765	3761
Macro avg	0.702	0.708	0.697	3761
Weighted avg	0.771	0.765	0.763	3761



Fine-tuned

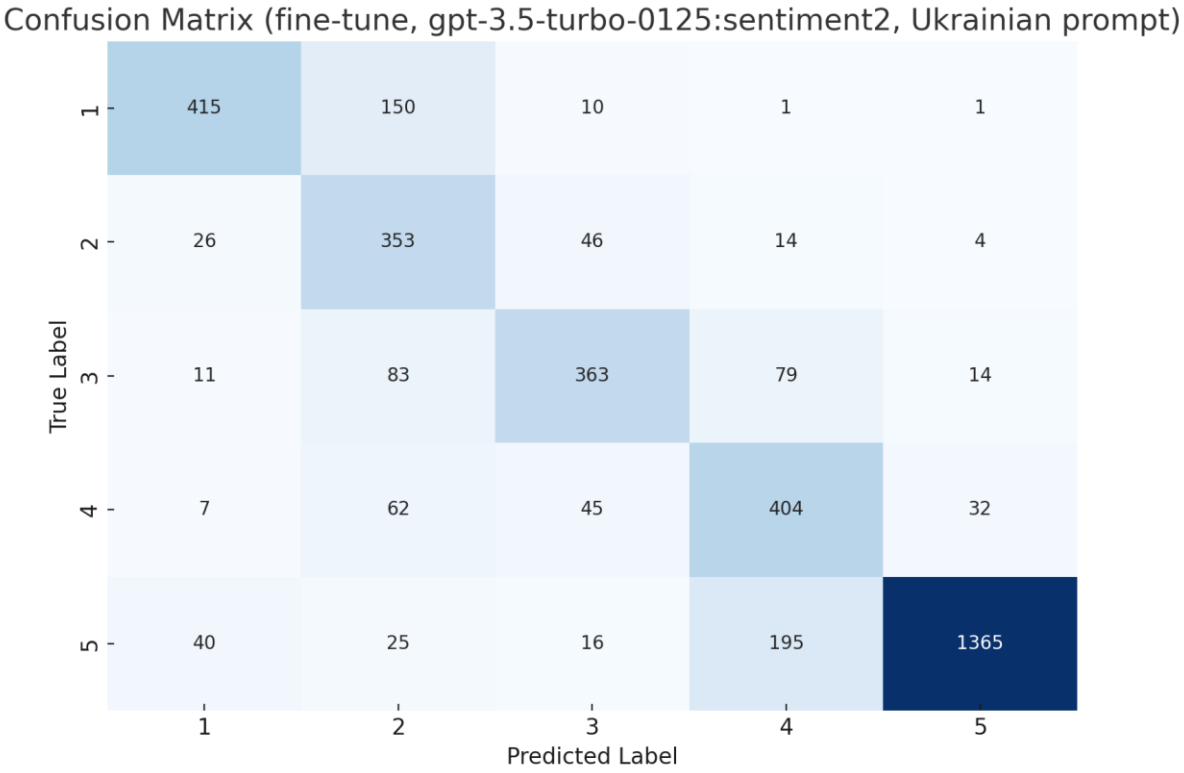
[3761/3761] Elapsed: 442s | ETA: 0s

✓ All responses received.

=== Classification report (fine-tune, gpt-3.5-turbo-0125:sentiment2, test_dataset.csv, ukrainian prompt ===

Клас	Precision	Recall	F1-Score	Support
1	0.832	0.719	0.771	577
2	0.525	0.797	0.633	443
3	0.756	0.660	0.705	550
4	0.583	0.735	0.650	550
5	0.964	0.832	0.893	1641

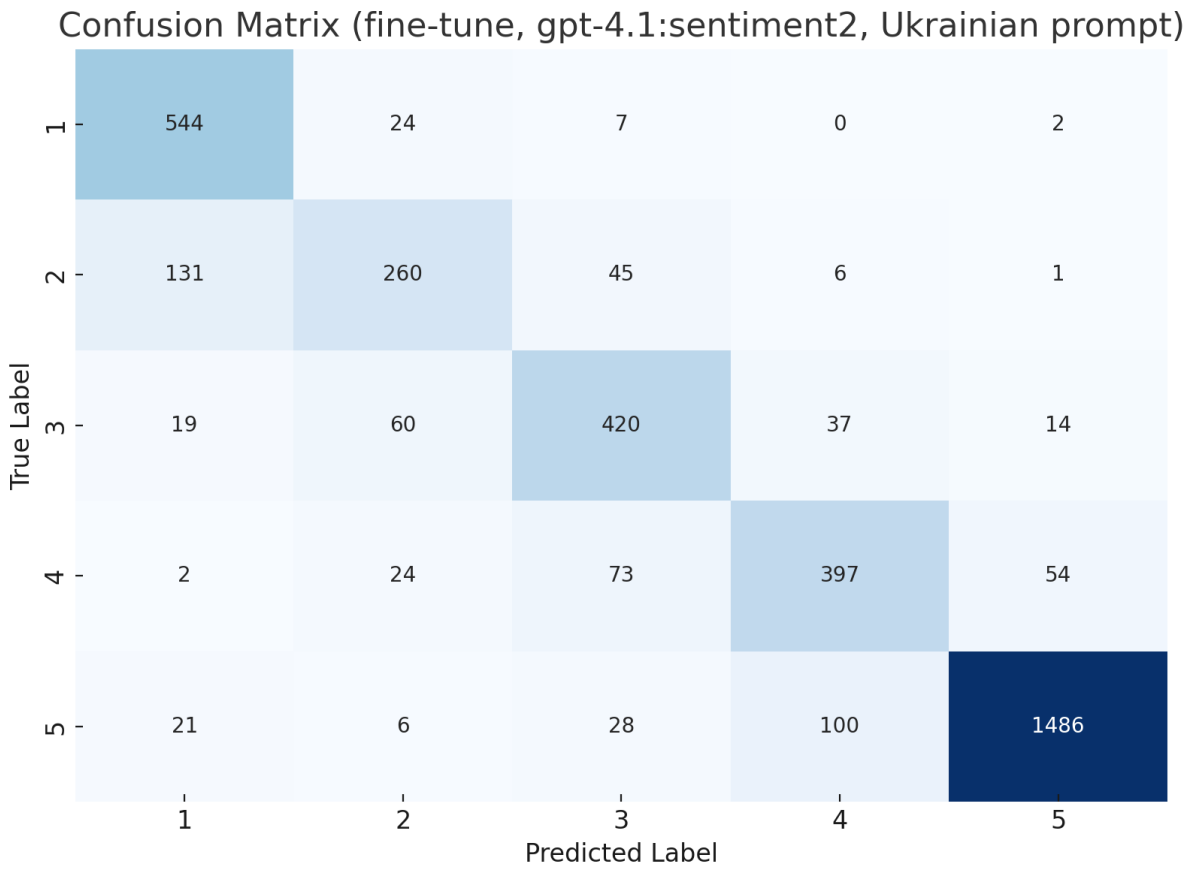
Accuracy			0.771	3761
Macro avg	0.732	0.748	0.730	3761
Weighted avg	0.806	0.771	0.781	3761



=== Classification report (fine-tune, gpt-4.1-2025-04-14:sentiment2, ukrainian prompt) ===

Клас	Precision	Recall	F1-Score	Support
1	0.759	0.943	0.841	577
2	0.695	0.587	0.636	443
3	0.733	0.764	0.748	550
4	0.735	0.722	0.728	550
5	0.954	0.906	0.929	1641

Accuracy			0.826	3761
Macro avg	0.775	0.784	0.777	3761
Weighted avg	0.829	0.826	0.825	3761

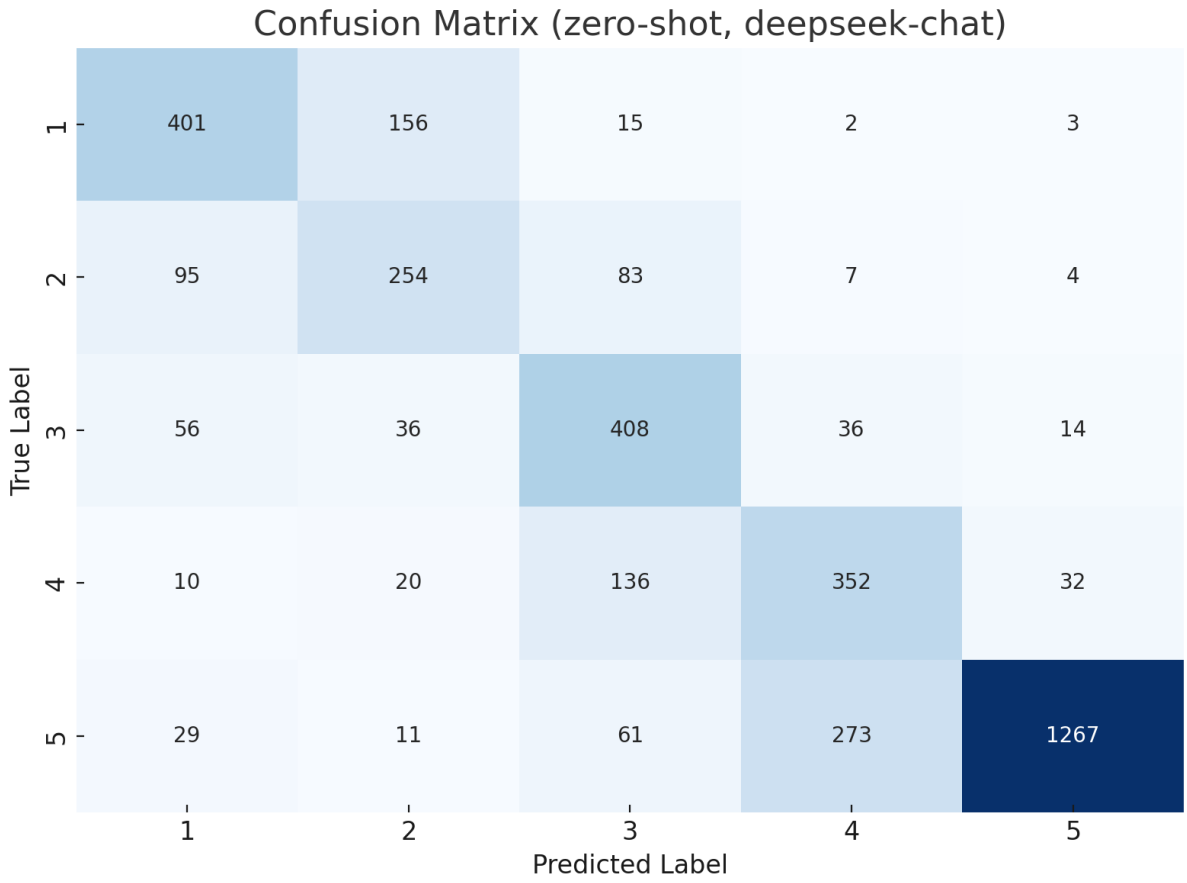


deepseek-chat

=== Classification report (zero-shot, deepseek-chat) ===

Клас	Precision	Recall	F1-Score	Support
1	0.679	0.695	0.687	577
2	0.532	0.573	0.552	443

3	0.580	0.742	0.651	550
4	0.525	0.640	0.577	550
5	0.960	0.772	0.856	1641
Accuracy			0.713	3761
Macro avg	0.655	0.684	0.665	3761
Weighted avg	0.747	0.713	0.723	3761



deepseek-reasoner

=== Classification report (zero-shot, deepseek-reasoner, ukrainian prompt) ===

Клас	Precision	Recall	F1-Score	Support
1	0.830	0.633	0.718	577
2	0.520	0.675	0.587	443
3	0.640	0.731	0.683	550
4	0.584	0.665	0.622	550
5	0.927	0.842	0.883	1641
Accuracy			0.748	3761
Macro avg	0.700	0.709	0.698	3761
Weighted avg	0.772	0.748	0.755	3761