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

Автоматичне визначення сарказму в українськомовних текстах

Кваліфікаційна робота бакалавра
студентки 4 курсу
освітньої програми
«Прикладна (комп'ютерна) лінгвістика та англійська мова»
спеціальності – 035.10 Філологія (прикладна лінгвістика)
галузі знань – 03 гуманітарні науки
Сніжанни Іванівни Ботвин
Науковий керівник:
Валентина РОБЕЙКО

«Допущено до захисту»
Протокол засідання
кафедри української мови та прикладної лінгвістики
протокол №15 від «06» 06 2024 року
завідувач кафедри _____(підпис)
к.філол.н., доц. Сергій РІЗНИК

КИЇВ – 2024

АНОТАЦІЯ

Кваліфікаційна робота спрямована на дослідження теми автоматичної класифікації українськомовних текстів на саркастичні або несаркастичні. **Актуальність теми** зумовлена відсутністю розробок для автоматичного виявлення сарказму саме в текстах, написаних українською мовою. Кінцева **мета** даного дослідження — запропонувати рішення для класифікації українськомовних текстів на такі, що містять сарказм, або ж ні. **Об'єктом** дослідження є українськомовні тексти. **Предмет дослідження** — лінгвістичні прояви сарказму в обраних текстах та способи їх автоматичної ідентифікації.

У першому розділі вказано такі відмінності сарказму від гумору, іронії, сатири: недоброзичливість, агресивність, наявність конкретної цілі, відсутність наміру викликати зміни в суспільстві. Розглянуто підходи для завдання автоматичної ідентифікації сарказму в текстах, а саме: правила, традиційне машинне навчання та глибоке. Визначено, що для такої бінарної класифікації тексту використовуються короткі та довгі тексти з можливим додаванням контексту різного типу. Також описано проблематику сарказму з боку автора та читача; вказано причини, чому сарказм важко визначити в тексті. До того ж проілюстровано основні ознаки сарказму на прикладі українськомовних текстів: гіперболу, пунктуаційні знаки, прагматичні ознаки (емотикони, емоджі, написання слова великими літерами), невідповідність, пародіювання російської вимови.

У другому розділі описано створення навчальної вибірки для завдання автоматичної ідентифікації саркастичного тексту. Також створено синтетичні саркастичні текстові дані та зроблено їх порівняння зі справжніми саркастичними даними. Проведено експерименти з моделями машинного навчання, включно з коригуванням гіперпараметрів та додаванням синтетичних даних. Запропоновано та опубліковано у вільному доступі систему, яка приймає текст від користувача та подає мітку для вказаного повідомлення — сарказм / не сарказм.

Ключові слова: сарказм, автоматичне виявлення сарказму, машинне навчання, класифікація текстів, синтетичні дані.

SUMMARY

The bachelor's thesis is aimed at exploring the topic of automatic classification of Ukrainian texts into sarcastic or non-sarcastic. **The relevance of the topic** is due to the lack of solutions for automatic detection of sarcastic texts written in Ukrainian. **The goal of this study** is to propose a system to the aforementioned text classification problem. **The object of the study** is Ukrainian texts. **The subject of the study** is linguistic features of sarcasm in selected texts and ways of their automatic identification.

The first section outlines the following differences between sarcasm and humor, irony, satire: unkindness, aggressiveness, the presence of a specific goal, and the lack of intention to cause changes in society. In addition, approaches to the task of automatic identification of sarcasm in texts are considered, namely: rules, traditional machine learning and deep learning. It is determined that for such a binary text classification, short and long texts are used with the possible addition of different types of context. We also describe the problems of sarcasm perception by the author and the reader; the reasons why sarcasm is difficult to identify in the text. In addition, the main features of sarcasm are illustrated using the examples of Ukrainian texts, namely: hyperbole, punctuation, pragmatic features (emoticons, emojis, capitalization), inconsistency, parody of Russian pronunciation.

The second section describes a dataset creation for the task of automatic classification of Ukrainian texts into sarcastic or non-sarcastic. We also create synthetic sarcastic data and compare it with real sarcastic data. Experiments with machine learning models were conducted, including adjusting hyperparameters and adding synthetic data. Finally, we propose and publish in an open source a system that accepts text from the user as input and provides a label: sarcasm or not.

Keywords: sarcasm, automatic sarcasm detection, machine learning, text classification, synthetic data.

ЗМІСТ

ВСТУП

Сарказм – таке висловлювання, буквальне значення якого відрізняється від того, яке мовець насправді хоче донести [72]. Його використання не тільки в живому спілкуванні, але й в онлайн-розмовах як способу передати власні думки та емоції завжди надзвичайно популярне. Проте висока складність виявлення сарказму в тексті може призводити до хибного тлумачення, непорозумінь та конфліктів.

Наприклад, сарказм часто є засобом вираження негативних настроїв через використання емоційно позитивно забарвленої лексики, тому розпізнати справжню тональність тексту стає непросто як для людей, так і для програм сентимент-аналізу. Це може призвести, наприклад, до збільшення кількості несправжніх “позитивних” відгуків на комерційних сайтах.

Загальною практикою в соціальних мережах є блокування неприйнятного контенту, до якого може випадково потрапити сарказм. Можна зробити припущення, що автоматичні системи виявлення сарказму рідко є частиною алгоритмів з перевірки текстів, наприклад, на прийнятність (тобто, текст не повинен містити дискримінаційних, образливих чи неправдивих висловлювань).

Розроблення комп’ютерних моделей для автоматичної класифікації текстів на сарказм, не сарказм розпочалось відносно недавно та впродовж останніх років набуває все більшої популярності. Автоматична ідентифікація сарказму визначається як проблема бінарної класифікації, що має на меті виявити чи є поданий текст саркастичним, чи ні [26]. Зважаючи на те, що сарказм — це не певна властивість однієї мови, то виникає загальна необхідність розробки таких систем для конкретних мов, зокрема для української. Поява все більшої кількості великих мовних моделей дозволяє застосовувати нові підходи до розв’язання різноманітних завдань, зокрема бінарної класифікації текстів.

Як нам відомо, досі таких систем для українськомовних текстів не розроблялось. Тому **актуальність теми** визначається відсутністю спеціальних

напрацювань щодо автоматичної класифікації саме українськомовних текстів на саркастичні або несаркастичні.

Практичне значення цього дослідження полягає в створенні набору даних, які містять тексти двох класів: сарказм, не сарказм. Ці дані були використані для тренування різних моделей машинного навчання. Сформована навчальна вибірка може бути джерелом аналізу загальновідомих та специфічних для українськомовних текстів ознак сарказму, а також використовуватися як набір даних для експериментів з алгоритмами машинного навчання.

Кінцева **мета** дослідження — створити модель (моделі) на основі машинного навчання для класифікації українськомовних текстів на дві категорії: сарказм або не сарказм.

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

1. Описати, чим сарказм відрізняється від інших різновидів комічного, а саме від гумору, іронії та сатири;
2. Провести огляд типів даних та підходів, що використовуються у галузі автоматичної ідентифікації сарказму;
3. Розглянути проблематику сарказму з боку автора та читача; причини, чому сарказм важко визначити в тексті;
4. Вказати основні ознаки сарказму та проілюструвати їх, якщо такий відповідник є, на українськомовних текстах;
5. Сформувати набір даних з саркастичними та несаркастичними текстами, порівняти згенеровані великими мовними моделями синтетичні саркастичні дані та справжні;
6. Здійснити тренування декількох алгоритмів машинного навчання, проаналізувати отримані результати, опублікувати навчену (навчені) модель (моделі) у вільному доступі.

Об’єктом дослідження є українськомовні тексти. **Предмет дослідження** — лінгвістичні прояви сарказму в обраних текстах та способи їх автоматичної

ідентифікації. **Матеріал дослідження** — 5378 повідомлення для кожного з класів (сарказм, не сарказм), опубліковані на платформі X та Telegram. **Методи**, що були використані в дослідженні: загальні методи (аналіз, синтез, моделювання), методи емпіричного дослідження (експеримент, порівняння), кваліметричний метод.

Методологічною основою дослідження є праці українських і зарубіжних вчених у галузі лінгвістики, автоматичного опрацювання природної мови, а також дослідження бінарної класифікації тексту на саркастичний або несаркастичний. Бакалаврська робота є продовженням наших попередніх курсових робіт “Автоматичне визначення сарказму в тексті” [1], “Автоматичне визначення сарказму на основі текстів з соціальної мережі Твіттер” [2]. Кваліфікаційна робота логічно розвиває й доповнює наші попередні напрацювання новими матеріалами для дослідження, а також отриманими результатами та висновками.

Структура й обсяг кваліфікаційної роботи. Робота складається зі вступу, трьох розділів: теоретичної частини (Розділ 1. «Огляд досліджень та основних проблем у галузі автоматичного визначення сарказму») та практичних результатів (Розділ 2. «Дані для алгоритмів машинного навчання»; Розділ 3. «Алгоритми машинного навчання»), містить висновки після кожного розділу та загальні висновки, список використаних джерел, додатки. Загальний обсяг бакалаврської роботи становить 77 сторінок. З них основного тексту 43 сторінок, список використаних джерел (93 найменування) – на 14 сторінках та додатки на 12 сторінках.

РОЗДІЛ 1. ОГЛЯД ДОСЛІДЖЕНЬ ТА ОСНОВНИХ ПРОБЛЕМ У ГАЛУЗІ АВТОМАТИЧНОГО ВИЗНАЧЕННЯ САРКАЗМУ

1.1. Розмежування сарказму, гумору, іронії та сатири

Окрім сарказму серед інших видів комічного існує також гумор, іронія, сатира. По черзі розглянемо кожен з них.

Термін “сарказм” походить від пізньолатинського слова sarcasmus (“насмішка”), який своєю чергою бере коріння від грецьких слів sarkasmos, тобто “знущання”, та sarkazein — “роздирати плоть як собака або кусати свої губи в люті” [44, с. 38]. Типово сарказм розглядається як гірке, глузливе висловлювання, де є невідповідність прямого й буквального значень.

Розрізнити сарказм та гумор не так складно, адже гумору властивий доброзичливий тон. Навіть у тому випадку, коли гумор направлений на недоліки людини, то їх демонстрація має життєрадісне забарвлення. Інакше кажучи, мовець сміється з іншими людьми з метою розвеселити слухачів. З сарказмом усе навпаки: мовець сміється над кимось чи чимось з метою розкритикувати, образити [85].

Сарказм часто прирівнюють до іронії або розглядають його як різновид іронії. Науковці розрізняють аж до восьми її форм [44]. Оскільки сарказм — суто лінгвістичне явище, то в намаганнях розмежувати його та іронію ми насамперед маємо на увазі той тип іронії, що проявляється при комунікації, тобто вербальну іронію.

Вербальна іронія — це іронія, що реалізована за допомогою слів. У будь-якому випадку ми маємо мовця (чи письменника), який має намір іронічно донести якесь повідомлення, та аудиторію, що зрозуміє або не зрозуміє значення повідомлення [24].

Одним із факторів, що розмежовує сарказм та вербальну іронію, є наявність конкретної цілі або жертви у випадку використання сарказму [90]. Причому жертвою може бути як окрема особа, так і група людей.

Наступним критерієм є жорстокість або агресивність. Сарказм нерідко

використовують як інструмент для вираження негативних емоцій, нещадної критики. В іронії наявна градація: її можна використовувати для лагідних та гострих зауважень. Однак сарказм сягає найвищої точки болючості, образливості [24].

Нарешті, сарказм є більш відвертим та очевидним, аніж іронія. Він не залишає місця для здогадок чи сумнівів [24].

Сарказм схожий на сатиру, але також відрізняється від неї. Сатира — такий різновид комічного, завданням якого є висміяти, зганьбити чи викрити вади, дурість, злочини людини, груп людей, держав або соціуму загалом з метою спровокувати зміни. Сатира насамперед охоплює сферу політики, її ціллю є пороки суспільства. Сарказм же часто направлений тільки на одну людину чи групу людей, що володіють окремими негативними якостями, але він не несе місію викликати зміни [6].

1.2. Автоматичне визначення сарказму в текстах

Для автоматичного виявлення сарказму в тексті використовуються комп'ютерні моделі двох типів: моделі на основі вмісту та на основі контексту. У моделях першого типу класифікація тексту проводиться з урахуванням лексичних ознак. Наприклад, у дослідженні [27] як лексичні ознаки були використані уніграми, кожна з яких представляє одне слово в тексті, та ознаки, що базуються на словниках з різними категоріями слів. Такі словники були об'єднані зі списком вигуків (наприклад, "ah", "oh", "yeah") та пунктуаційних знаків (наприклад, "!", "?"). Ще одне дослідження використовує лінгвістичні ознаки, що базуються на теорії невідповідності. До прикладу, невідповідністю може бути наявність поряд слів з протилежною тональністю як у реченні "I love being ignored", де love має позитивну тональність, а ignored — негативну [38].

Щодо контексту, то, наприклад, до уваги можна взяти те, у якій спільноті користувачів певний текст був опублікований. Так, речення "I really am proud of Obama" з великою ймовірністю є саркастичним, якщо воно було

опубліковане у форумі, який часто відвідують консерватори [33].

Для автоматичного визначення сарказму спочатку дослідники використовували методи, що базуються на певних правилах. Прикладом такого методу є вищезгадані лінгвістичні ознаки. Хоча такі методи легше пояснити, вони не є достатньо ефективними. Так, цей підхід покладається чітко на якість зібраних правил та їхню кількість, тому існує ризик неправильної класифікації, якщо якась з ознак була пропущена [88].

Крім того, застосовуються статистичні методи машинного навчання, коли алгоритми тренуються на проанотованих даних. Такий підхід більш надійний, ніж засновані на правилах підходи в тому, наприклад, що відбувається адаптація до нових даних і шаблонів. Проте одним із мінусів статистичних методів є те, що вони покладаються на певну кількість якісних даних.

Наразі популярними є різні типи нейронних мереж, що мають перевагу в автоматичному витягуванні ознак, які при ручному конструюванні не могли б бути виявлені [52]. У такому випадку модель завдяки, наприклад, використанню певного типу архітектури, може виявляти складніші випадки сарказму в тексті, адже вона матиме інформацію про контекст (до прикладу, попередні слова). Причому не обов’язково тренувати нейронну мережу з нуля. Останнім часом все більше використовуються попередньо натреновані великі мовні моделі, які вже володіють статистичними знаннями про мову. Обравши велику мовну модель та маючи певну кількість специфічних даних, можна донавчити модель для виконання конкретного завдання. Та все ж такий підхід вимагає наявності технічних ресурсів певної потужності. Насамкінець, тут стикаємося з явищем “чорного ящика”, адже не завжди можна пояснити, чому модель передбачила саме такий клас для тексту.

1.3. Типи даних

Найпоширенішим типом даних, які використовують для навчання систем автоматичної ідентифікації сарказму, є короткі тексти. Їх джерелом

слугують соціальні мережі, серед яких найбільш популярна платформа X (Twitter). Повідомлення, опубліковане в X, називається твітом. Для анотації твітів використовують два підходи. Перший — мануальна анотація, коли анотатори вручну відносять конкретний текст до однієї з двох категорій, а саме: сарказм та не сарказм. Другий підхід — використання Twitter developer APIs, за допомогою яких автоматично формують базу повідомлень за двома категоріями відповідно до хештегів, тобто ключових слів, або емодзі, що вказують на певну емоцію. Другий підхід є більш популярним, бо допомагає швидко створити масштабний набір даних та зберігає людський час. Наприклад, для маркування саркастичного та несаркастичного текстів, написаних англійською мовою, зазвичай використовують хештеги *#sarcasm*, *#sarcastic*, *#not*, *#irony* та *#angry*, *#happy*, *#blessed*, *#frustrated* відповідно [35]. Інколи дослідники припускали, що твіти, які не містять хештег *#sarcastic*, є несаркастичними [10].

Оскільки сарказм значно залежить від контексту та екстралінгвістичних знань, то щоб визначати його з більшою точністю також використовували або попередні пости користувача на сторінці, або ж увесь діалог, у якому з'явилось саркастичне повідомлення [66, 93].

Дещо менш поширеним типом даних є довгі тексти: відгуки на книги, фільми, придбані товари та дискусії на форумах загалом. Наприклад, популярним ресурсом даних на основі діалогів став серіал "Друзі", де кожну репліку героїв відносили до одного з двох класів [36].

Серед звичайних навчальних вибірок для автоматичного визначення сарказму існують і такі, які залучають не тільки текст. До прикладу, люди часто використовують сарказм у розмові з іншими, тому було створено також набір даних на основі телефонних дзвінків, де кожне використання фрази "yeah right" було анотоване як саркастичне або ж ні [80]. Окрім того, існують мультимодальні набори даних, які залучають, наприклад, візуальний компонент. Так, текст "What a wonderful weather!" поєднується з зображенням,

де наближається буря [12].

Щодо обсягу навчальних вибірок, то дослідження [10] використовує збалансований набір даних, де кількість текстів для кожної з двох категорії становить 9767. У роботі [27] було створено навчальну вибірку з 900 текстів для трьох категорій: сарказм, емоційно позитивно та емоційно негативно забарвлені тексти. Корпус, створений на основі діалогів у серіалі "Друзі", містить 1888 саркастичних реплік [36]. Дослідження [42] презентує корпус з 533 мільйонами коментарів, де 1.3 мільйон — саркастичні тексти. Джерелом цієї навчальної вибірки слугувала платформа Reddit. Отже, обсяг текстових даних для вирішення завдання автоматичної ідентифікації сарказму не фіксований, а змінюється залежно від обраного дослідження.

1.4. Сарказм з боку мовця та читача. Золотий стандарт

Попередньо ми згадували про мануальну анотацію, коли анотатори вручну відносять конкретний текст до наперед визначених категорій, а саме: сарказм та не сарказм. Саме тут ми стикаємося з розмежуванням сарказму з боку автора висловлювання та зі сторони того, хто це висловлювання отримує. Для ефективного вирішення тих завдань, що стосуються аналізу настроїв, необхідно розрізняти сарказм з погляду як автора, так і читача, щоб уникнути будь-яких потенційно неправильних інтерпретацій [71].

У тих випадках, коли автор зі свого боку першопочатково наділяє текст саркастичністю, сарказм називається "навмисним" ("intended"). І навпаки, коли текст сприймається як саркастичний з погляду отримувача, тоді сарказм є "осмисленим" ("perceived") [57]. Соціальне, культурне походження мовця та аудиторії можуть впливати на те, як повідомлення буде сприйняте, й наскільки ефективною буде комунікація. Наприклад, рівень знайомства оратора з цільовою аудиторією може впливати на ймовірність використання ним сарказму [56].

Так, дослідження [22] демонструє, як регіональні відмінності можуть впливати на використання сарказму. У ньому студенти коледжів Нью-Йорку та

Теннессі брали участь у завданнях, які вимірювали їхнє використання сарказму: вони надавали визначення термінам іронія та сарказм, відповідали на питання про свій досвід та ймовірність використання сарказму. Виявилося, що учасники з Нью-Йорку схильні частіше формувати завершення повідомлень, що містять сарказм, аніж учасники зі штату Теннессі. Якщо враховувати й інші демографічні фактори, то, наприклад, учасники з Нью-Йорку та чоловіки повідомляли, що використовують сарказм частіше, ніж учасники з Теннессі та жінки відповідно.

Результат анотації текстів також може бути різним, якщо маємо культурні відмінності, що описано в дослідженні [37]. У цьому експерименті науковці використали два набори саркастичних текстових даних, що містили повідомлення з соціальної мережі “X” та дописи на дискусійних форумах, для виявлення розбіжностей в анотації між учасниками з США та Індії. За результатами дослідження було зроблено висновок, що учасники з Індії, на відміну від їхніх колег з США, більше погоджуються одне з одним. Водночас вони ж і частіше натрапляють на труднощі у випадку незнайомих ситуацій та іменованих сутностей. Щоб проілюструвати, у дослідженні згадується такий приклад: “It’s sunny outside and I am at work. Yay”. Це висловлювання інтерпретується як несаркастичне анотаторами з Індії, адже така погода є типовою для їхнього клімату. Заразом анотатори із США сприймають його як саркастичне. Зважаючи на подібні відмінності, все ж варто зазначити, що за результатами вищезгадані труднощі в анотації сарказму призводять до статистично незначного погіршення при його класифікації.

Інше дослідження було зосереджене на виявленні того, чи можуть графічні зображення емоцій, або емотикони, впливати на розуміння сарказму учасниками різного віку [32]. Результати цього експерименту підтверджують, що сприйняття текстів може бути різним, адже старші учасники були більш схильні до буквальної інтерпретації текстів, ніж їхні молодші колеги: вони

повідомляли як про частіше використання сарказму, так і частіше тлумачення неоднозначного тексту як саркастичного.

Враховуючи таке розмежування сарказму з боку автора та читача, вже були спроби створити текстові навчальні вибірки, де повідомлення були проанотовані самими авторами тексту. Як зазначають одні з перших дослідників такої розробки: “Попередні моделі, які досягали високої продуктивності у виявленні сарказму в наборах даних, що відображають осмислений сарказм (мітки експертів) або сарказм із хештегами (дистанційне спостереження), раптово не змогли виявити сарказм, як це мав на увазі автор” [58, с. 1287]. З іншого боку, модель, що використовувалась у дослідженні [62], була менш ефективною при виявленні сарказму з боку читача. Припускається, що детальніша інформація щодо контексту користувача (наприклад, певних деталей щодо сторінки в соціальних мережах) могла б допомогти спрогнозувати, як він чи вона можуть реагувати на повідомлення.

Якщо перед нами покладено завдання аналізу громадських думок, то треба враховувати намір автора повідомлень; якщо потрібно виявити мову ненависті, то необхідно брати до уваги можливі інтерпретації аудиторією, що має різне культурне, соціальне походження [58]. Зазначимо, що наше поточне дослідження зосереджене на вирішенні першого з цих двох завдань.

При обговоренні анотації навчальних вибірок варто згадати й про таке поняття в машинному навчанні, як золотий стандарт. Золотим стандартом заведено вважати такі дані, що були підготовлені або перевірені вручну, а тому є максимально об’єктивними [61]. Хоча текстові дані такого типу для вирішення завдання автоматичної ідентифікації сарказму широко використовуються (наприклад, популярний проанотований корпус діалогових текстів [59] або ж повідомлень із соціальної мережі “X” [84]), використання дистанційного спостереження для отримання міток також є поширеною практикою (до прикладу, навчальна вибірка отримана за допомогою тегів з платформи “X” [10] або дані отримані з Реддіт за тегом /s [42]).

Обидва вищезгадані підходи мають свої недоліки. Як ми вже зазначали, тлумачення тексту як саркастичного може залежати від, наприклад, соціокультурних особливостей. Тому виникає питання, кого обирати анотатором даних для отримання якомога об'єктивнішого результату. Крім того, неоднозначне повідомлення, де намір автора незрозумілий, буде призводити до низького рівня згоди між анотаторами. У випадку ж автоматичного отримання текстів за тегами виникає проблема шуму. Наприклад, сама мітка саркастичності може міститися в повідомленні іншому, аніж те повідомлення, що містить текст із сарказмом. До того ж маркування теж можуть мати неоднозначну семантику. Так, хештег #сарказм може бути як самим маркування саркастичного тексту, так і просто частиною текстового повідомлення, де знак решітки використовується, щоб більше користувачів побачило повідомлення.

1.5. Ознаки сарказму

Соціальні мережі нерідко впроваджують обмеження на довжину повідомлень. Наприклад, на платформі X першопочатково повідомлення могло містити до 160 символів стандартного кодування Юнікод, 20 з яких були збережені для команд та назв користувачів. З часом ліміт символів був збільшений до 280 [19]. Як бачимо, за цими обмеженнями, повідомлення є короткими. Щоб висловити свою думку, емоції наскільки чітко й зрозуміло, наскільки це можливо з дотриманням ліміту, користувачі використовують ознаки, які можуть допомогти зрозуміти про що саме йдеться у тексті. Наприклад, це можуть бути емоджі (емотикони), хештеги, певні слова чи частини мови. Звісно, за відсутності голосу, тону, міміки в текстовій комунікації подібні маркери є вкрай важливими для виявлення сарказму. Тому варто зробити огляд тих основних саркастичних ознак, зокрема лінгвістичних, які були використані вже в наявних дослідженнях. Такі ознаки проілюстровані на основі текстів з наших даних.

Гіпербола як ознака сарказму. Використання гіперболи збільшує ймовірність того, що текст містить сарказм [28]. Відповідно до Великої

Української Енциклопедії, гіпербола — це засіб увиразнення мови, основою якого є підкреслене перебільшення [4]. Крім того, використовуючи цей різновид образної мови мовець може висловлювати протилежність тому, що він або вона має на увазі [70]. Подібна ознака невідповідності часто згадується у присвячених сарказму дослідженнях [91, 41, 17]. Одним зі способів, яким можна виразити гіперболу, є слова з більшою інтенсивністю. Такі інтенсифікатори, або ж слова-підсилювачі, не є культурними артефактами лише певної мови, а навпаки: вони використовуються в усіх мовах і мають відповідні еквіваленти в них [8]. Слова-підсилювачі дають можливість вийти за межі нейтрального й натомість посилюють емоційну виразність мови, створюючи так звану емфазу [9]. Лексичні засоби, такі як слова та словосполучення, граматичні, лексико-граматичні засоби можуть використовуватися для передачі емфатичних конструкцій [7].

Таблиця 1.5.1

Засоби вираження	**Приклади з нашої навчальної вибірки**
Вигук	@caesarius14_ ***Ого вау***, неочікувано /сарк
Прислівник	@CrimeaUA1 Після цього віде мені стало ***остаточно*** зрозуміло хто на росії хозяїн...***звичайно*** що рузкі (сарказм) Підординцями були такими і є
Подовження голосних	"Люди, яким скучно жити, а ви рахували калорії?? це т***ааааа***к в***ееесеее***ло! сарказм*)"
Підсилювальна частка	"@andersostlund Та ну, ви що!!? Корупція - то ***лише*** в Україні!*сарказм*"
Прикметник (нагромадження)	Б.М.В. is a bitch! виходячи з мого внутрішнього переконання і ***вільної, неупередженої, обєктивної*** оцінки. сарказм

Емфатичність (гіпербола) як ознака сарказму

Наприклад, у повідомленні з наших даних “Обожнюю твою пунктуальність. *сарказм*” дієслово “обожнюю” очевидно є більш емоційно сильним: за інтенсивністю воно сильніше виражає позитивний сентимент, аніж дієслово “люблю” з такою ж позитивною оцінкою, але меншою інтенсивністю.

Для емфатичності також можуть бути використані такі частини мови, як: прислівники (особливо кількісно-означальні), прикметники, підсилювальні частки та вигуки. Подовження літер, а частіше саме голосних звуків, також може свідчити про намір автора наголосити на чомусь. Приклади використання деяких засобів емфатичності проілюстровано в табл. 1.5.1.

Пунктуація як ознака сарказму. Знаки пунктуації — не менш важливий елемент при визначенні сарказму в тексті. Найважливішими серед них є знак оклику “!”, знак питання “?”, лапки ““”, еліпсис “…”. [21, 47]. Еліпсис використовують для того, щоб позначити відсутність певної частини тексту або ж незакінчену думку [41]. Еліпсис зазвичай частіше трапляється в саркастичних текстах, аніж у буквальних; його більше асоціюють із критикою (буквальною та саркастичною), але рідше з буквальною похвалою [82]. Послідовність зі знаків оклику або знаків запитань також може бути маркером сарказму, але не завжди. У дослідженні [14] зазначено, що послідовність знаків оклику може виражати також вдячність, особливо якщо текст містить мало слів, але закінчується великою кількість знаків оклику. Звісно, якщо брати до уваги тільки цю ознаку, то вона може виявитись не зовсім ефективною для ідентифікації сарказму, але в комбінації з іншими ознаками може зробити свій вагомий внесок.

Таблиця 1.5.2.

Засоби вираження	**Приклади з нашої навчальної вибірки**
Послідовність знаків оклику	@samspiesonyou Ви шо, це ж вищий пілотаж**!!1!1**/сарк
Лапки	@m_for_mixacb і коли це станеться ? бо заїбали літачки та ракетки з території **″**вільної (сарказм), незалежної (знов сарказм)**″** РБ
Послідовність знаків питання	@bioravlyk Ви що, забороняєте жінкам народжувати**????** Сексистка! /Сарк

Пунктуація як ознака сарказму

У саркастичних текстах також можуть використовуватися лапки або, якщо бути точним, то так звані scare quotes. За визначенням Кембриджського онлайн

словника це “лапки (= символи « » або ' '), які іноді ставлять навколо слова чи фрази в письмовому реченні, щоб показати, що слово вжито особливим чином або таким чином, який може бути неправильним чи неправдивим” [68]. Щоб зазначити іншу тональність в тексті, саме scare quotes можуть бути використані [28]. Приклади ознак на основі пунктуації наведено в табл 1.5.2.

Прагматичні ознаки сарказму.. Прагматика вивчає як природна мова використовується в процесі комунікації з урахуванням наміру мовця та слухача; вона пов’язана з прихованим значенням тексту [15]. До прагматичних ознак інколи зараховують пунктуацію, про яку вже було сказано, а також емоджі, емотикони, написання слова великими літерами, згадки про користувачів [38, 63, 12].

Таблиця 1.5.3.

Засоби вираження	**Приклади з нашої навчальної вибірки**
Написання слова великими літерами	@dw_ukrainian @Light1065S @poroshenko Мабуть ***ВАС ТЕЖ КУПИЛИ***!!! хто заказав цей пост? не вірю що громадянин України міг таке опублікувати *сарказм
	як це так ??? жінка??? ***МАЄ ПРАВО***???! ***ВИРІШУВАТИ***???? що робити зі ***СВОЇМ ТІЛОМ***???? ну це вже вопше.... /сарказм
Емотикони та емоджі	@HetmanXIII Дуже достовірна і важлива інформація, дякую **;)** (Сарказм)
	@Vetal_Lukas @Bordfunker133 @sunny_whale_2 Бо ти осмілився на жіноче тіло дивитись. Як тобі не соромно (сарказм).😉

Прагматичні засоби як ознака сарказму

Написання слова великими літерами, як і велика кількість знаків оклику, використовується, щоб надати більшої інтенсивності певній емоції в тексті, тобто — для емфатичності. Щодо зображення емоцій, а саме емотиконів та емоджі, то вони хоча й пов’язані між собою, але водночас і мають певні

відмінності. Так, емотикон — “схематичне зображення людського обличчя, яке використовують у чатах та соцмережах для передавання емоцій” [5]. Емоджі — це зображення невеликого розміру, яке може відтворювати як людське обличчя, так й інші речі [23]. Емотикони роблять першопочатково неоднозначний текст більш саркастичним [32]. Серед емотиконів майже виключно у саркастичних повідомленнях використовувались емотикон з язиком “:p” та той, що підморгує “;)” [82]. Наприклад, другий з цих двох емотиконів позначає своєрідну прірву між контекстом у вигляді самого повідомлення та виразом обличчя, а, отже, наштовхує на думку про приховане значення сказаного [32]. Приклади вищеперерахованих прагматичних ознак наведені в табл. 1.5.3.

Невідповідність як ознака сарказму. Неодноразово в дослідженнях з автоматичного виявлення сарказму науковці зосереджувались на невідповідності [39, 67, 38, 47]. Традиційно таку невідповідність трактують як вираження мовцем думки, що є протилежною до того значення, що виражене буквальним висловлюванням [54]. Наприклад, сарказм може виникати, коли зіставляється слово з позитивною тональністю та дія або стан з негативною тональністю [67]. Дослідження [38] також враховує таку невідповідність між сентиментами, але класифікує її як явну. Крім цього, тут також враховується прихована невідповідність через фрази. Як приклад наводиться речення “I love this paper so much that I made a doggy bag out of it” з зіставленням слова love та фразеологізму made a doggy bag out of it. Приклади використання невідповідності в наших саркастичних текстах наведено в табл. 1.5.4.

За нашим спостереженням українськомовним саркастичним текстам характерна ще одна ознака: своєрідне перенесення звучання слів російською мовою на українські слова. Наприклад, повідомлення “@verctana @MGorokhovska "**пачіму ваш порох падписал минск и вот я должен єту вайну терпеть** "” або ж “Це були комуністи, якшо шо) от вам і **вєлікій савєцкій саюз** /сарк нег”. Можемо припустити, що такі повідомлення зазвичай

використовуються з метою висміяти певну групу людей, а саме тих людей, що є носіями проросійських сентиментів.

Таблиця 1.5.4.

невідповідність	Ммммм, знову **повітряна тривога**. Як я це **люблю**. /сарк /нег
	обожнюю митися в **крижаній воді** коли на вулиці +8° /САРК І ДУЖЕ НЕГ
	ще й **герпес** вискочив, падлюче. **дуже приємно** /сарк

Невідповідність як ознака сарказму в наших саркастичних текстах

1.6. Чому сарказм важко визначити в тексті?

Тема автоматичної ідентифікації сарказму набирає своєї популярності з публікації дослідження, що зосереджувалося на ідентифікації сарказму в усній мові [35]. Автори цієї роботи використали спектральні, контекстуальні та просодичні характеристики для вирішення завдання розпізнавання сарказму [80]. Щодо просодичних характеристик, то, наприклад, подовження голосних звуків, назалізація, збільшений час артикуляції та нечітка вимова можуть бути індикаторами сарказму в розмові [43]. Крім того, в особистому спілкуванні мовці можуть користуватись різними виразами обличчя (до прикладу, підняти одну брову, закотити очі) [16]. Однак, в онлайн розмовах ми переважно використовуємо саме текст як спосіб комунікації з іншою людиною. У текстовій формі ми втрачаємо, наприклад, тон голосу чи вирази обличчя, і відсутність такої додаткової інформації ускладнює автоматичне визначення сарказму.

Варто зазначити, що деякі аналогічні ознаки все ж таки в тексті можуть використовуватись. У попередньому підрозділі ми зазначали таку ознаку, як подовження голосних літер та емоджі / емотикони. До того ж разом із текстом може використовуватись візуальний супровід. Відомим прикладом серед користувачів Інтернету є зображення актора Ніколаса Кейджа, відоме за назвою “You Don’t Say?” (див. Додаток 1). Воно відображає лють, але комічну, і,

зокрема, використовується як саркастична відповідь на очевидне висловлювання [92].

Варто зазначити, що вибір джерел для створення навчальної вибірки також має свої тонкощі. На платформах X та Telegram можна спілкуватись не тільки за допомогою тексту, але й також використовувати зображення та посилання. Якщо текст містить сарказм, то для того, щоб його усвідомити, ймовірно треба буде перейти за прикріпленим посиланням або звернути увагу на прикріплене зображення, якщо такі є. Також згадані платформи дозволяють створювати так звані conversational threads, коли користувачі можуть відповідати на повідомлення одне одного. Якщо один з онлайн співрозмовників не зрозумів прихованого значення певного повідомлення й відповів на нього із серйозністю, то автор оригінального повідомлення може відповісти схожим текстом "#сарказм обов'язково ставити?". Отже, знову стикаємося із присутністю нерелевантних повідомлень.

Відсутність контексту та нерозуміння системами штучного інтелекту здорового глузду також ускладнює розуміння сарказму. Щодо контексту, то проілюструємо його важливість на прикладі тексту з нашої навчальної вибірки. Повідомлення "Офігенний день. П.С. сарказм" потенційно віднесемо до саркастичних текстів, адже маємо відповідне маркування від самого автора. Та за відсутності цього маркера навіть самій людині було б важко інтерпретувати це повідомлення. Якби мовець його висловлював в усному мовленні, ми могли б отримати підказки у вигляді, наприклад, виразу обличчя чи тону голосу. У текстовому ж вигляді таке повідомлення легко сприйняти буквально, особливо коли його автор використав мало або зовсім не використав аналогічних текстових саркастичних ознак. До прикладу, додаючи емотикон ":(" можна створити протиставлення позитивного та негативного сентименту між словом "офігенний" та відповідним емотиконом, що допомогло б читачам інтерпретувати це повідомлення як саркастичне з більшою ймовірністю.

Визначення здорового глузду за Кембриджським словником говорить, що це “базовий рівень практичних знань і суджень, які необхідні всім нам, щоб допомогти нам жити розумним і безпечним способом” [18]. Наприклад, повідомлення з наших даних “Співпадіння у стрічці, буває ж таке. І да, що таке Блумберг, хто там йому вірить? От портнов то ж авторитет! #сраказм https://t.co/fi2PM0IiW0” містить іменовані сутності, а саме Блумберг та Портнов (“#сраказм” — одне із ключових слів для маркування саркастичного тексту). Широкій аудиторії Блумберг може бути відомим як провідна компанія в інформаційній сфері, а Андрій Портнов як політичний діяч, якщо бути точніше — проросійський політичний діяч. Відповідно, не маючи цих знань і не зрозумівши таке протиставлення, можна сприйняти повідомлення буквально, а не саркастично.

Висновки до першого розділу

Сарказм як висловлювання з протиставленням буквального й справжнього значень можна відрізнити від інших видів комічного за його недоброзичливістю, агресивністю, наявністю конкретної цілі, відсутністю наміру викликати зміни в суспільстві.

Для автоматичного визначення сарказму застосовують підходи, засновані на правилах, традиційні методи машинного навчання та глибокі нейронні мережі. Для навчання моделей, що класифікують текст на саркастичний чи ні, використовують короткі та довгі тексти з різних джерел (соціальні мережі, книги, форуми, діалоги з фільмів тощо), що можуть поєднуватись з додатковою інформацією (наприклад, прикріплене зображення, посилання, попередні повідомлення).

Завдання автоматичної класифікації текстів вимагає багато якісних даних, проанотованих автоматично (за допомогою маркерів автора тексту) або мануально (за допомогою залучення експертів). Попри те, що згадані підходи до анотації даних мають свої недоліки, вони обидва часто використовуються для вирішення завдання ідентифікації сарказму.

У текстових даних немає характерних усній мові маркерів сарказму. А втім, саркастичні тексти все ж мають певні властивості, що можуть допомогти читачеві розтлумачити справжнє значення повідомлення. До таких ознак належать: гіпербола, пунктуаційні знаки, прагматичні ознаки (емотикони, емоджі, написання слова великими літерами), невідповідність, пародіювання російської вимови, де остання ознака характерна саме україномовним текстам.

Попри популярність теми автоматичного виявлення сарказму, це завдання має й складнощі: вибір джерела, способу збору та анотації текстових даних; часта відсутність контексту повідомлення та недосконаліть інтелектуальних систем щодо загальновідомих знань.

Для нашого дослідження актуальні короткі текстові повідомлення з соціальних мереж, месенджерів. Такі повідомлення можна автоматичного збирати за допомогою програмних бібліотек. Пошук може відбуватись за ключовими словами, що маркують текст як саркастичний. У такому випадку отримуємо анотацію текстів від самих авторів. Для класифікації повідомлень використаємо традиційне та глибоке машинне навчання, адже розробка правил вимагає великої кількості часу та є обмеженою в можливості адаптуватися до небачених раніше даних.

РОЗДІЛ 2. ДАНІ ДЛЯ АЛГОРИТМІВ МАШИННОГО НАВЧАННЯ

У цьому розділі описано формування навчальної вибірки, створення синтетичних саркастичних даних та їх порівняння зі справжніми саркастичними даними.

2.1. Формування навчальної вибірки

Попередньо було сформовано навчальну вибірку на основі соціальної мережі “X”. Для цього було використано бібліотеки snscrape [75] та twint [78], за допомогою яких з платформи “X” отримано повідомлення, що містили такі ключові слова: сарказм, сркзм, сарк, сраказм, sarcasm, sarc. Несаркастичними вважалися ті дані, які не містили відповідних ключових слів. У підсумку ми отримали 3795 текстів для кожного з класів (сарказм, не сарказм). Однак, через деякий час при спробі збору додаткових текстів ми помітили, що наші скрипти перестали працювати. Впевнившись, що проблема пов’язана не з самим скриптом, а з платформою, ми дійшли до висновку, що вказана ситуація виникла через нові обмеження платформи щодо збору даних [51]. Так виникло питання про пошук іншого джерела текстів.

Ми обрали Телеграм через легкість взаємодії з його API. Телеграм має канали, що є засобом поширення повідомлень на велику аудиторію [79]. Адміністратори можуть прикріплювати до каналів обговорення, де спілкуються користувачі. Ми обрали 23 українських Телеграм-канали й використали такі обговорення для формування додаткових текстів до наявної навчальної вибірки.

Цього разу було використано метод реактивного нагляду (reactive supervision) [71]. Він пов’язаний із використанням ключового повідомлення, яке вказує на сарказм у попередніх текстах. Табл. 2.1.1 ілюструє обмін повідомленнями в Телеграмі в обговореннях: користувач А написав саркастичне повідомлення, а користувач В, не розуміючи справжнє значення написаного тексту, відповідає справжнім повідомленням. Далі користувач А відповідає на повідомлення користувача В і вказує, що написане ним/нею попереднє

повідомлення було саркастичним. Так можна тримати анотування саркастичних повідомлень від самих авторів.

Таблиця 2.1.1

Користувач	Повідомлення з наших даних
A	та який профіт, там же 1-2 людини з інвалідністю та й годі 😁😁😁
B	По перше, мова йде не тільки про модернізацію транспорту для осіб з інвалідністю, а про загальну модернізацію. По друге, транспорт, як галузь економіки, сильно впливає на інші. Якщо коротко — розвиток транспорту стимулює розвиток інших галузей економіки.
A	це був сарказм, але дякую, що пояснив іншим абсолютно погоджуюся та підписуюся під кожним словом

Приклад гілки повідомлень з платформи Телеграм

Для пошуку відповідних повідомлень було використано чотири ключові слова: сарказм, sarcasm, сарк, sarc. Така гілка повідомлень могла містити від одного до трьох текстів. Після цього, щоб відсіяти непотрібні гілки текстів, тексти для кожного каналу були переглянуті вручну. У результаті ми отримали такі гілки повідомлень, з яких для створення саркастичної вибірки було залишено тільки перше повідомлення з кожного діалогу. Обсяг саркастичної вибірки з Телеграму становить 1583 тексти. Обсяг саркастичної вибірки з платформи “X” становить 3795 текстів. До попередніх несаркастичних текстових даних з мережі “X” були додані тексти кожного з 23 каналів з обговорень, що не містили ключових слів сарказм, sarcasm, сарк, sarc. У результаті ми отримали збалансований набір даних обсягом 10756 текстів по 5378 повідомлень на кожен з двох класів.

2.2. Створення синтетичних даних

У машинному навчанні чим більше даних ми маємо, тим краще. Зважаючи на популярність великих мовних моделей останнім часом та їхню здатність породжувати текст, ми вирішили не обмежуватись справжніми

повідомленнями й використали також синтетичні саркастичні дані для вирішення завдання класифікації текстів на сарказм та не сарказм.

Для генерації текстів ми вирішили використати дві великі мовні моделі від компанії OpenAI та Google DeepMind, а саме GPT-4 Turbo [31] та Gemini Experimental [81]. Для взаємодії з моделлю від OpenAI було використано OpenAI API [55]. Для вивчення цього API та експериментів з ним всі нові користувачі отримують безплатні токени загальною вартістю $5. Для взаємодії з Gemini Experimental було використано Vertex AI — платформу для розробки, створення та використання генеративного штучного інтелекту [87].

Великим мовним моделям можна надавати інструкції для виконання певних завдань. Такі інструкції називаються підказками (a prompt) [89]. Ми використали 3 види підказок: підказка з описом завданням (zero-shot prompting), підказка з описом завдання та одним прикладом справжнього саркастичного тексту (one-shot prompting), підказка з описом завдання та трьома прикладами саркастичного тексту (few-shot prompting) (див. Додаток 2). У кожній з інструкцій ми вказували великій мовній моделі згенерувати по 50 прикладів текстів. Було зроблено 20 ітерацій, у результаті чого ми мали б отримати разом по 1000 текстів на кожну модель.

Таблиця 2.2.1

Велика мовна модель	**Тип інструкції**	**Кількість текстів**	**Загальна кількість текстів**
Gemini Experimental	zero-shot	911	2939
	one-shot	1011	
	few-shot	1017	
GPT-4 Turbo	zero-shot	767	2554
	one-shot	928	
	few-shot	859	

Результат продукування синтетичних саркастичних текстів обраними великими мовними моделями

Бачимо, що обидві моделі не дотримувались інструкції щодо кількості згенерованих текстів: жодна з моделей не виконала умови 1000 прикладів тексту. Крім того, GPT-4 Turbo видає меншу кількість текстів, а Gemini Experimental — іноді видає більшу кількість, ніж вимагається.

Ми зіткнулись також із проблемою дотримання умови формату повідомлень, а саме формату JSON. Обидві моделі інколи порушували його структуру (наприклад, відсутність лапок та/або дужок там, де це потрібно). Такі хиби довелось виправляти вручну.

Окрім цього ми виявили, що отримані дані з Gemini Experimental містили 25 ідентичних повідомлень, водночас дані з GPT-4 Turbo не містили дублікатів.

Також хочемо описати інше спостереження. Інструкція з прикладами справжніх саркастичних текстів містила як зразки такі тексти на політичну тематику: “<користувач> <користувач> Вона не голосувала за пукіна….”, “Світу до росіян не існувало, запам'ятайте”. Припускаємо, що саме вони могли підштовхнути модель до продукування таких текстів, що потенційно можуть бути шкідливими. Наприклад, на початку деяких ітерацій модель Gemini Experimental створила тексти, які містять ознаки сарказму (наприклад, прислівники, частки, невідповідність). Завдяки ним або ж загальним знанням про світ ми можемо зрозуміти, що пряме значення тексту відрізняється від того, що насправді мається на увазі. Приклад таких текстів: “Російські вибори є зразком чесності та прозорості” або ж “Окупанти прийшли визволяти українців від нацистів. З нацистськими прапорами та символікою.”. Однак пізніше бачимо такі тексти, що часто виступали частиною пропаганди: “Українська мова - це просто діалект російської”, “Всі ці фотографії та відео з війни - фейки.”. Інколи ці тексти були написані від першої особи (“Я пишаюся тим, що я росіянин”, “Я підтримую спецоперацію.”) або ж декілька текстів поспіль формували ланцюжок (“Росія ніколи не здасться.” → “Ми - великий народ.” → “У нас особливий шлях.”). На нашу думку, такі тексти можуть свідчити про те, що великі мовні моделі можуть використовуватись не тільки з хорошими

намірами, але й для, наприклад, дезінформаційних кампаній, що дозволятиме продукувати велику кількість тексту зі шкідливим для безпеки країни змістом за короткий проміжок часу. При взаємодії з моделями Gemini можна вказати налаштування безпеки. Під час проведення експерименту наші налаштування безпеки повинні були блокувати той текст, який має високу ймовірність небезпеки. Можливо, змінивши налаштування, створення таких текстів вдасться уникнути.

Щодо моделі OpenAI, то наприкінці деяких ітерацій ми помітили генерацію тексту латинкою (“Tak, zvychayno, shcho ty poradyv meni zayty v neispravnyi lift. Adrenalin - ne robus cheeks,oni diskutno pislya etogo.”) або ж продовження тексту спеціальними символами (“Вау, і як я люблю, коли вєJackson8\x01#%*($.5\\'$cursors мн$().)]".00&!)().[]...”).

2.3. Порівняння справжніх та синтетичних даних

Щоб порівняти реальні та синтетичні дані було прийнято рішення сформувати вибірки однакового розміру за кількістю слів. Спочатку з усіх текстів було прибрано посилання, згадки користувачів, а також зведено повідомлення до нижнього регістру. Після цього було використано бібліотеку tokenize_uk [83], яка розділила тексти на окремі сегменти. На наступному кроці отримані сегменти були відфільтровані регулярними виразами так, щоб залишились тільки слова. Далі для кожного тексту було пораховано кількість слів. Насамкінець, повідомлення обиралися так, щоб сформувати однакові за обсягом слів вибірки. В результатів кожна вибірка містить по 30002 слова.

Щоб порівняти три вибірки було обраховано лексичні статистичні характеристики [3], а саме: обсяг тексту N, кількість окремих лексем V, кількість окремих словоформ $V_ф$, багатство словника, або ж індекс різноманітності, (V / N), середня повторюваність слова (N / V), індекс винятковості для тексту та для лем (V_1 / N та V_1/V відповідно), індекс концентрації та для тексту та для лем (V_{10} / N та V_{10} / V відповідно). Зазначимо, що V_1 це кількість слів із частотою 1 (hapax legomena), а V_{10} — кількість слів з

абсолютною частотою 10 та більше. Слова були зведені до початкової форми за допомогою бібліотеки stanza [77].

Таблиця 2.3.1

Синтетичні саркастичні дані		
	Gemini Experimental	*GPT-4 Turbo*
zero-shot	10001	10001
one-shot	10001	10001
few-shot	10000	10000
Справжні саркастичні дані		
Telegram	15001	
Twitter	15001	

Кількість слів за справжніми та синтетичними саркастичними вибірками

З табл. 2.3.2 можемо побачити, що слів та словоформ більше в справжніх саркастичних даних. Найменше їх у синтетичних даних моделі Gemini Experimental. Величина багатства словника, як і варіативності лексики, також найбільша, знову ж таки, для справжніх саркастичних даних. Індекс концентрації, так само як і середня повторюваність слова, найбільша для моделі Gemini Experimental.

Таблиця 2.3.2

	Gemini Experimental	**GPT-4 Turbo**	**Справжні саркастичні дані**
$V_ф$	5390.0	6652.0	10336.0
V	3447.0	4173.0	7439.0
V / N	0.115	0.139	0.248
N / V	8.704	7.19	4.033
V_1	3140.0	4274.0	7978.0
V_1 / N	0.105	0.142	0.266
V_1 / V	0.911	1.024	1.072
V_{10}	20337.0	18956.0	15043.0
V_{10} / N	0.678	0.632	0.501
V_{10} / V	5.9	4.543	2.022

Лексичні статистичні характеристики для трьох вибірок

Окрім статистичних характеристик, було визначено іменовані сутності для кожної вибірки. Для цього було використано бібліотеку spaCy [76]. Так, синтетичні дані, згенеровані моделлю OpenAI, містили найменшу кількість іменованих сутностей, а саме: Київ, Нью Йорк, Лондон, Макдональдс, Україна, ЄС та Укрзалізниця. Натомість синтетичні дані, згенеровані Gemini Experimental, містили понад 20 іменованих сутностей, а реальні саркастичні дані — більше сотні. Підрахувавши відносну абсолютну частоту для кожної іменованої сутності, було обрано 10 найчастотніших для реальних даних та даних з Gemini Experimental.

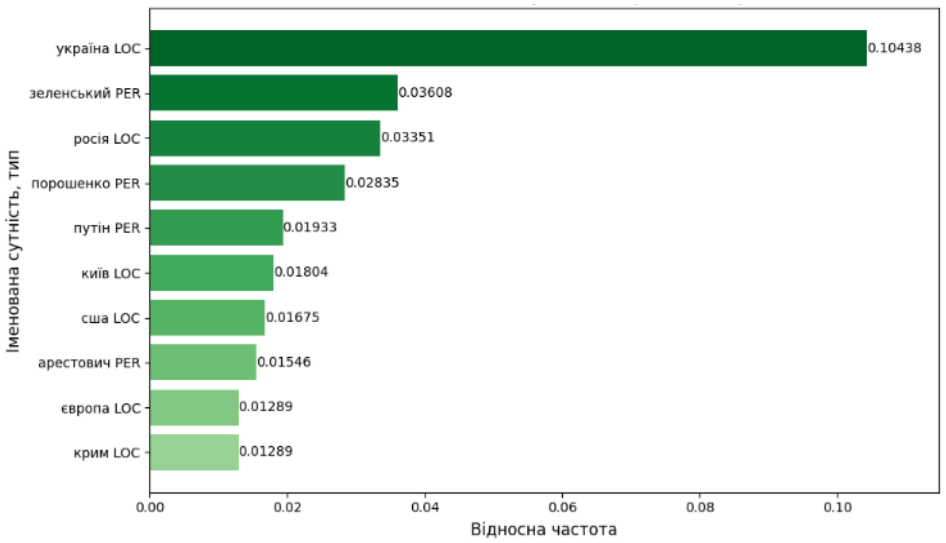


Рисунок 2.3.1. 10 найчастотніших іменованих сутностей в справжніх саркастичних даних

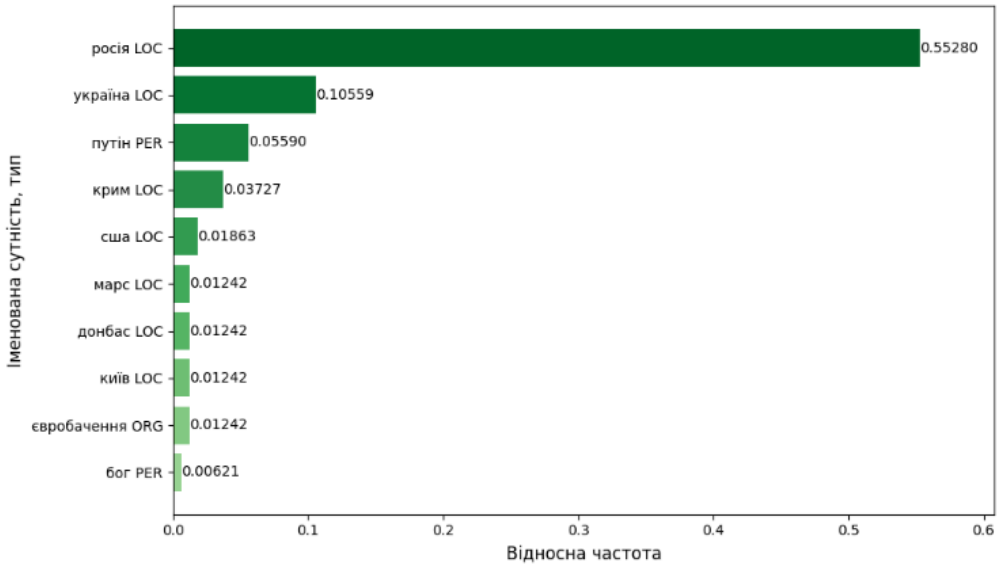


Рисунок 2.3.2. 10 найчастотніших іменованих сутностей в синтетичних саркастичних даних моделі Gemini Experimental

Бачимо, що обидві вибірки мають спільні іменовані сутності. Наприклад, Україна, росія, Зеленський, путін, США, Крим. Такі іменовані сутності свідчать про наявність політичної тематики.

За таким же принципом було обрано 10 найчастотніших біграм для трьох вибірок. Варто зазначити, що на рисунках зображено тільки ті біграми, які не містили стоп-слова як один із компонентів.

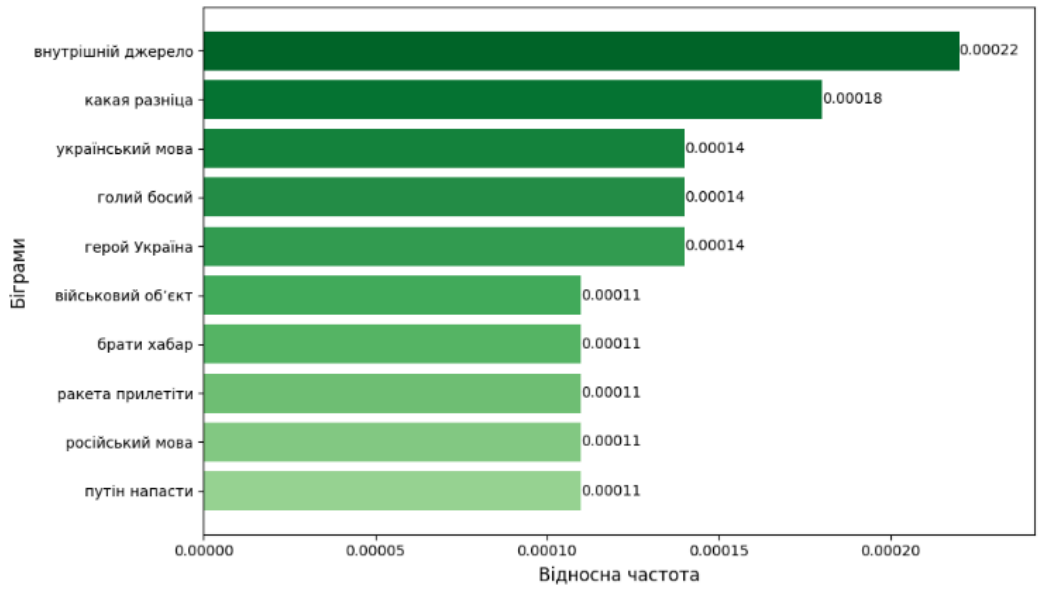


Рисунок 2.3.3. 10 найчастотніших біграм в справжніх саркастичних текстах

Деякі біграми зі справжніх саркастичних текстів також пов’язані політичною тематикою. Ба більше, такі біграми, як “військовий об’єкт”, “ракета прилетіти та “путін напасти” свідчать про безпосередній зв’язок з повномасштабним вторгненням рф в Україну.

Біграми з синтетичних даних не містять згадок про політику. Вони радше пов’язані з буденним життям. Можемо також помітити, що для обох вибірок є й спільна біграма, а саме “найкращий спосіб”. Крім того, обидві вибірки містять біграми з позитивним сентиментом (наприклад, “чудовий ідея”, “улюблений серіал”, “радість допомогти”, “радість піти”). Синтетичні дані з моделі Gemini Experimental мають також біграми з негативним сентиментом, як-от “телефон

розряджатися” та “комп’ютер зависати”, а біграми даних моделі OpenAI, пов’язані з часом, посідають перші місця.

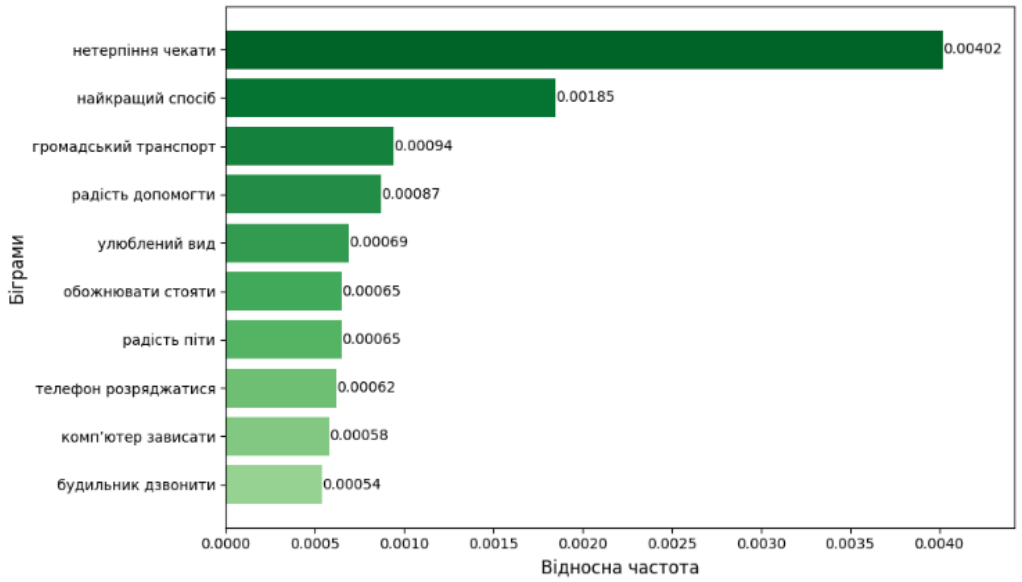


Рисунок 2.3.4. 10 найчастотніших біграм в саркастичних текстах моделі Gemini Experimental

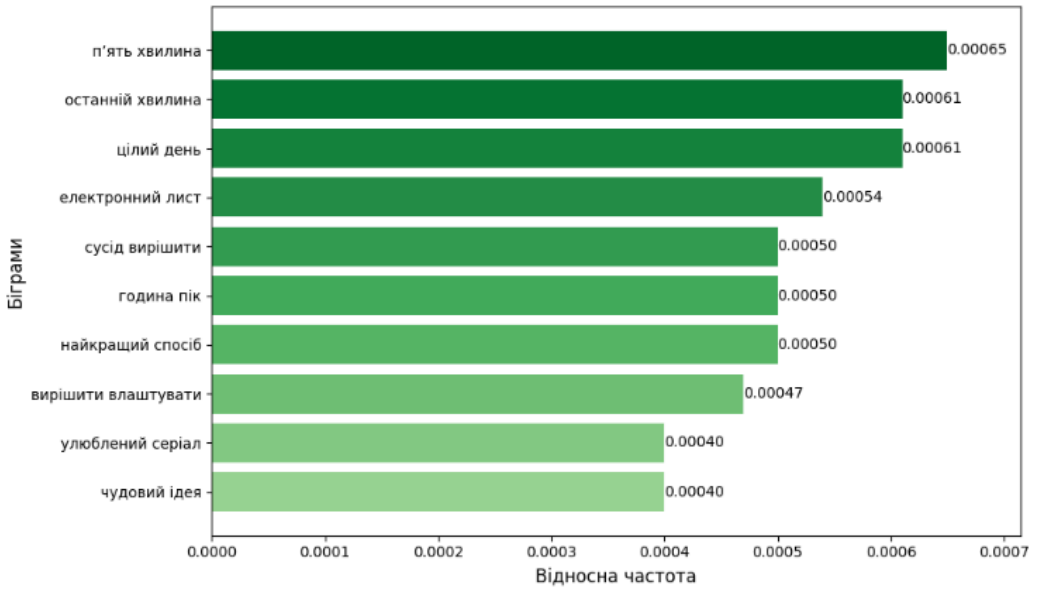


Рисунок 2.3.5. 10 найчастотніших біграм в саркастичних текстах моделі GPT-4 Turbo

Висновки до другого розділу

Навчальну вибірку, що містила 3795 саркастичних повідомлень з платформи “X”, було доповнено 1583 новими текстами, зібраними з платформи “Телеграм”.

Порівняння справжніх саркастичних даних та синтетичних саркастичних даних за статистичними лексичними характеристиками вказує на те, що найбагатша лексика в справжніх саркастичних даних. Якщо ж порівнювати використані дві великі мовні моделі, то за метриками кращі результати в моделі від OpenAI. Щодо визначення найчастотніших іменованих сутностей та біграм, то це дало нам інформацію про тематику даних (наприклад, наявність політичного складника).

Крім того, ми вирішили оцінити те, наскільки людина може відрізнити синтетичні саркастичні тексти від справжніх саркастичних текстів. Було створено форму, яка містила випадковим чином обрані 10 справжніх та 10 синтетичних саркастичних текстів з наших даних. Для кожного з них потрібно було обрати одну з міток: справжній, синтетичний. Подані тексти були перемішані.

Ми отримали відповіді від 13 осіб (див. Додаток 4). Середня кількість правильних відповідей — 15. Інакше кажучи, у 75 % випадків відповіді були правильні. Дев’ять з десяти синтетичних саркастичних текстів були щонайменше 4 рази обрані як справжні. Тільки один приклад синтетичного саркастичного тексту був всіма особами обраний як синтетичний: “Так, звісно, чуючи вас постійно скаржитися на роботу, моє життя стає яскравішим.”. Таке речення граматично побудоване неправильно. Крім того, у коментарях форми зазначили, що невдале поєднання дієслова *сказав* з іменником *історію* в синтетичному тексті “Дякую, що сказав ту ж історію вдесяте, вона не стала менш смішною.” вказує на його штучність. Також зазначалось, що синтетичні тексти можна відрізнити за пунктуацією та великою літерою на початку тексту. Можливо, ми отримали більшість правильних відповідей саме завдяки подібним ознакам штучних даних.

У підсумку можемо сказати, що синтетичні дані можуть бути використані для завдань класифікації текстів, зокрема класифікації текстів на саркастичні та

несаркастичні. Однак, варто пам’ятати, що такі дані варто додатково перевіряти, наприклад, на наявність пропагандистських елементів.

РОЗДІЛ 3. АЛГОРИТМИ МАШИННОГО НАВЧАННЯ

У цьому розділі описано використання машинного навчання для створення системи автоматичної класифікації тексту на два класи: сарказм, не сарказм. Серед алгоритмів машинного навчання ми обрали два традиційні алгоритми, а саме логістичні регресію та випадковий ліс. Крім того, ми застосували дві нейронні мережі, які використовують архітектуру трансформерів. Перша — попередньо натренована мовна модель RoBERTa, яку ми донавчали на сформованій навчальній вибірці, інша — велика мовна модель GPT-4 Turbo, що була використана як готове рішення. Насамкінець, було проаналізовано результати та опубліковано систему для автоматичної класифікації текстів на саркастичні та несаркастичні.

3.1. Тренування моделей для бінарної класифікації тексту

Передопрацювання даних для тренування традиційних алгоритмів машинного навчання розпочиналась з тексту без згадок користувачів та гіперпосилань. Наступним кроком було очищення повідомлень від спеціальних символів та знаків пунктуації, окрім знаку оклику, питання, еліпсису та лапок, оскільки вони можуть свідчити про наявність сарказму. Токенізація відбувалась з допомогою функціоналу nltk [53], а саме TweetTokenizer, оскільки цей токенізатор підходить саме для текстів з соціальних мереж (наприклад, він виділяє емодзі як окремий токен). Стоп-слова з тексту не забирались, оскільки ми помітили, що їхня відсутність дещо погіршує якість моделі. Насамкінець, слова було зведено до початкової форми функціоналом бібліотеки stanza [77].

Для тренування традиційних алгоритмів машинного навчання було використано бібліотеку scikit-learn [69]. Серед алгоритмів ми обрали логістичну регресію: її часто використовують як базове рішення, що задовольняє мінімальний рівень якості в розв’язанні певного завдання. Інший алгоритм — це алгоритм випадкового лісу. Ми розділили дані за принципом 80:20. Тобто, 80 % даних для тренування моделі та 20 % для перевірки її якості.

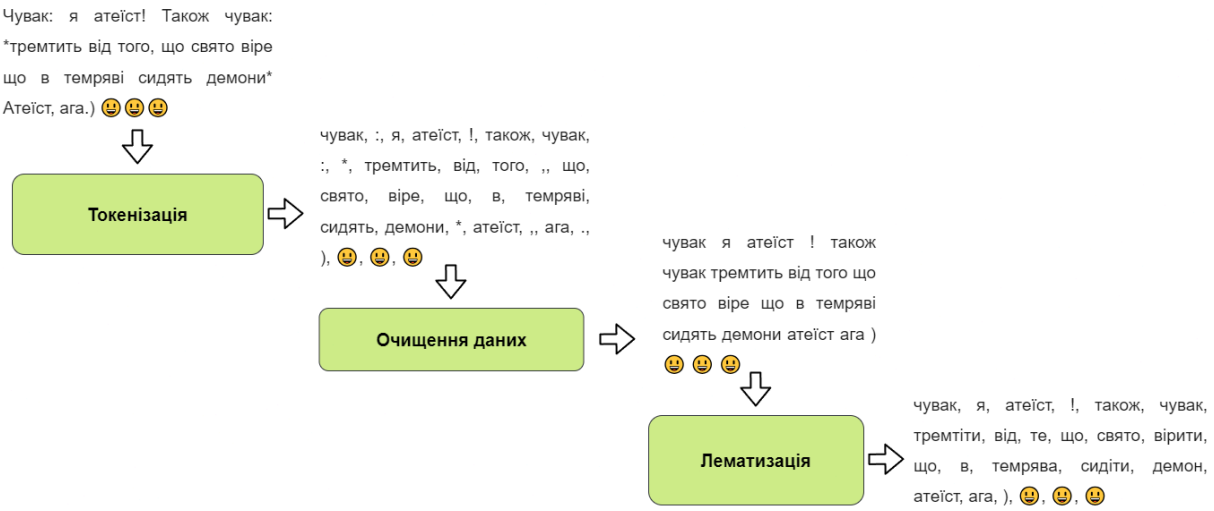


Рисунок 3.1.1. Етапи попереднього опрацювання даних для тренування традиційних алгоритмів машинного навчання

Для векторизації тексту, тобто його представлення як вектора чисел, було використано техніку TF-IDF (Term Frequency-Inverse Document Frequency). TF представляє абсолютну частоту слова, IDF — кількість усіх текстів, поділених на кількість текстів, що містять слово. Саме IDF надає словам, що зустрічаються часто, меншої важливості, а словам, що зустрічаються рідше, а тому можуть бути особливою характеристикою певного документа, більшої важливості.

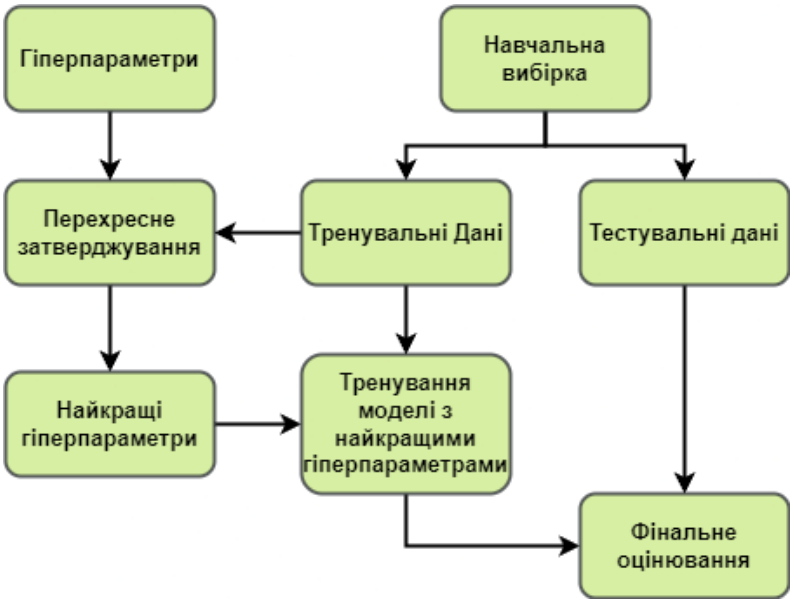


Рисунок 3.1.2. Етапи тренування традиційних алгоритмів машинного навчання

Тренування вказаних алгоритмів виконувалось у декілька етапів. Спершу ми використали алгоритми без жодних додатково вказаних гіперпараметрів (змінних, які не вивчаються моделлю, а задаються користувачем перед початком навчання) та перевірили їхню якість на тестових даних. Після цього для кожного з алгоритмів вказали певний набір значень для гіперпараметрів і використали техніку GridSearch [30] з перехресним оцінюванням для обрання найкращих значень для гіперпараметрів. Потім застосували алгоритми з найкращими гіперпараметрами для навчання на повній тренувальній вибірці.

Щоб перевірити, чи синтетичні дані допоможуть покращити якість алгоритмів для нашого завдання бінарної класифікації текстів, ми обрали синтетичні тексти від моделі GPT-4 Turbo, адже за лексичними статистичними характеристиками вони показали кращі показники за модель Gemini Experimental. Щоб зберегти баланс класів, саркастичні дані було доповнено справжніми несаркастичними текстами з обговорень вищезгаданих 23 Телеграм каналів.

Тепер зупинимося на використаних алгоритмах та їхніх параметрах.

Логістична регресія — модель, яка прогнозує ймовірність настання події. У нашому випадку, ймовірність того, чи є наш текст саркастичним. Було використано такі параметри цього алгоритму:

1. штраф (penalty) — параметр, що пов’язаний з регуляризацією. Це процес, під час якого зменшується вплив менш значущих змінних через надання їм коефіцієнтів близьких до нуля. Тобто застосовується штраф, коли модель використовує велику кількість змінних [60];
2. C — параметр, пов’язаний з силою регуляризації. Чим менший коефіцієнт, тим сильніша регуляризація й тим простіша модель з меншою кількістю параметрів та значними коефіцієнтами. За допомогою налаштування цього параметру можна віднайти баланс між складністю й простотою моделі [49];

3. розв’язувач (solver), що використовується для процесу оптимізації з метою зменшити похибку класифікації [74];
4. max_iter — максимальна кількість ітерацій необхідна, щоб розв’язувач віднайшов такі параметри, які дають найкращі результати на даному наборі даних [74].

Алгоритм випадкового лісу є прикладом ансамблевого навчання: використовується декілька класифікаторів в одній моделі задля підвищення її продуктивності. Замість того, щоб покладатися на результат одного дерева рішень, цей алгоритм об’єднує прогнози з кожного дерева й виводить остаточний результат на основі більшості голосів [50]. Для алгоритму випадкового лісу було обрано такі гіперпараметри [73]:

1. n_estimators, який визначає кількість дерев рішень, що будуть входити до випадкового лісу. Тобто, скільки дерев буде створено й використано для передбачення;
2. max_depth — гіперпараметр, який впливає на складність моделі, адже визначає максимальну кількість рівнів, або ж глибину, в кожному з дерев рішень;
3. min_samples_split, який визначає поріг розбиття вузла, зважаючи на кількість елементів даних, і задає необхідну для цього мінімальну кількість вибірок;
4. min_samples_leaf, що визначає мінімальну кількість елементів даних, дозволених у термінальних вузлах дерева.

При використанні традиційних алгоритмів машинного навчання в завданнях обробки природної мови виникає декілька труднощів. По-перше, такі моделі опрацьовують послідовності по одному токену за раз, тому інформація з розташованих далеко один від одного токенів могла втрачатися, що призводило до труднощів у розумінні взаємозв'язків між словами. По-друге, подібні алгоритми обмежені в розумінні контексту. Класичним прикладом ілюстрації цього є багатозначне англійське слово bank, яке може позначати, наприклад,

фінансову установу або берег річки. Щоб правильно вказати значення цього слова в конкретному тексті треба зважати на навколишні слова. Наприклад, у реченні “I saw a man near the bank” остаточно не можна визначити, значення вищевказаного слова. Однак коли ми дамо контекст “While I was swimming in the river, I saw a man near the bank”, то маємо більшу ймовірність, що йдеться саме про берег річки.

Розв’язання цієї проблеми було запропоновано у 2017 році [86], коли було представлено архітектуру нейронних мереж, а саме трансформери, які використовують механізм уваги. Він допомагає моделі зосередитись на важливих частинах вхідної послідовності під час передбачення, а не розглядати всі частини рівноцінно, що дозволяє краще вловлювати взаємозв’язки у вхідному тексті

Прикладом моделі представлення мови, яка базується на трансформерах, є BERT [20]. Вона фіксує значення слова, зважаючи на його сусідів зліва та справа. Ми ж використали таку трансформерну модель як RoBERTa — вдосконалення BERT [48]. RoBERTa базується на таких самих принципах, як і її попередник, тільки більш оптимізованих. Головна відмінність полягає в процесі попереднього тренування моделі та використаних даних. При навчанні RoBERTa було прибрано завдання передбачення наступного речення, збільшено кількість тренувальних даних, час тренування та довжину послідовностей. Крім того, застосовувалось динамічне маскування — при кожному проходженні моделлю тренувальних даних маскувались різні токени. Такі масковані токени модель має передбачити, зважаючи на контекст, що формується немаскованими токенами.

Ми обрали попередньо навчену українськомовну RoBERTa від компанії YouScan [65]. Завдяки попередньому навчанню, під час якого зокрема використовувались тексти з соціальних мереж, модель уже володіє певними статистичними знаннями про мову. Тепер ми можемо донавчити її на меншій

кількості специфічних даних для виконання певного завдання. Такий процес називають файнтюнінгом (fine-tuning) [46].

Для підготовки даних та тренування RoBERTa було використано платформу Hugging Face [33]. Підготовка даних для донавчання трансформерної моделі відрізняється від підготовки даних для традиційних моделей машинного навчання. Другий підхід вимагає більше уваги щодо очищення даних. Ми прибирали певні спеціальні символи, проводили лематизацію слів. Для RoBERTa ми ж попередньо прибрали тільки імена користувачів та гіперпосилання. Надалі опрацювання тексту відбувалось за допомогою функціоналу згаданої платформи Hugging Face.

Кожна попередньо навчена модель має власний токенізатор. Наша модель має 52 тисячі токенів, які можна використовувати для представлення тексту. Деякі з токенів називаються спеціальними токенами, щоб вказати, наприклад, чи є токен початком або кінцем речення: bos_token: <s>, eos_token: </s>, unk_token: <unk>, sep_token: </s>, pad_token: <pad>, cls_token: <s>, mask_token: <mask>.

Подаючи текст токенізатору, ми отримуємо два поля:

1) вхідні ідентифікатори, які відповідають числовим кодуванням, що ставлять у відповідність кожному токену ціле число;
2) маска уваги: вказує моделі, які токени слід ігнорувати при обчисленні механізму уваги.

Якщо під час використання традиційних алгоритмів машинного навчання під час токенізації ми зберігали слова повністю, то в цьому випадку токенізація відбувається на рівні частин слів. Тобто, слово розбивається на менші частини.

Рисунок 3.1.3. Результат роботи токенізатора для моделі RoBERTa

Дотреновування моделі RoBERTa відбувалось на тих же 80% відсотках від усієї навчальної вибірки. Решту 20% даних були розділені на дві частини: перша використовується для коригування моделлю своїх внутрішніх параметрів з метою зменшити перенавчання, інша — для оцінки її якості. Щодо гіперпараметрів, то були використані такі:

1) кількість пройдених за одну ітерацію тренувальних текстів (batch size); у нашому випадку — 16. Більше значення для цього параметру вказуватиме на більшу кількість даних, що модель опрацьовує за один раз. Це пришвидшує процес тренування, але вимагає більшої місткості пам’яті;
2) епоха (epoch) — скільки разів модель проходитиме увесь тренувальний набір даних. Ми використали 4 епохи;
3) швидкість навчання (learning rate) визначає розмір кроку під час процесу знаходження оптимальних значень, при яких досягається найкращий результат з мінімізації похибки. Чим більше значення, тим більші кроки, тим швидше навчається модель. Однак при занадто швидкому навчанні можна пропустити значення, за яких досягається мінімум похибки. У нашому випадку значення дорівнювало 1e-5;

4) зменшення ваги (weight decay) відповідає за надання переваги меншим значенням ваг нейронної мережі, щоб уникнути перенавчання. Наше значення — 0.02.

Останньою моделлю, яку ми використали для бінарної класифікації тексту, стала вже згадана велика мовна модель GPT-4 Turbo. Взаємодія відбувалась через API, як і при генерації синтетичних даних. Моделі була надана інструкція, за якою для вказаних текстів вона мала проставити мітку 1 для тексту, що потенційно містить сарказм, і 0 для тексту, що не містить сарказм. Використовувались тексти з тих самих 20 % від навчальної вибірки, що й для попередніх моделей.

3.2. Аналіз отриманих результатів

Якість роботи обраних чотирьох моделей оцінювалась за метриками точності, повноти, надійності та міри F1. В основі оцінки цих метрик закладено поняття матриці невідповідності. Матриця невідповідностей (confusion matrix) подає число хибно позитивних (false positives), хибно негативних (false negatives), істинно позитивних (true positives) та істинно негативних (true negatives) випадків. Наведемо детальніше, що означають ці метрики та як вони обчислюються:

1) Точність (precision) характеризує, яка частина об’єктів, що була ідентифікована класифікатором як позитивна, є насправді позитивною. Точність демонструє здатність алгоритму відрізняти даний клас від інших класів:

$$\frac{TP}{TP + FP}$$

2) Повнота (recall) характеризує, яку частину об’єктів позитивного класу з усіх об’єктів позитивного класу ідентифікував алгоритм. Повнота демонструє здатність моделі ідентифікувати даний клас взагалі:

$$\frac{TP}{TP + FN}$$

3) Надійність (accuracy) подає кількість правильно класифікованих текстів серед усіх текстів і обчислюється за формулою:

$$\frac{TP + TN}{TP + FP + TN + FN}$$

4) Міра F1 (F1-score) — середнє гармонічне точності та повноти:

$$\frac{2 * Precision * Recall}{Precision + Recall}$$

Найкращі параметри для логістичної регресії вказані в табл. 3.2.1. Бачимо, що за метриками використання найкращих гіперпараметрів не дало жодних покращень. За метрикою повноти відбулось навіть зниження значення на 0,01 відсотка. Використання синтетичних даних змогло покращити цю ж метрику на 0,03 %.

Таблиця 3.2.1

Алгоритм	**Параметри**	**Дані**	**Accuracy**	**Precision**	**Recall**	**F1-score**
Логістична регресія	-	справжні	0.68	0.67	0.74	0.70
	C=10, max_iter=2000, penalty='l2', solver='saga'	справжні	0.68	0.67	0.73	0.70
		справжні + синтетичні	0.68	0.66	0.76	0.70

Результати якості логістичної регресії

Найкращі параметри для алгоритму випадкового лісу вказано в табл. 3.2.2. Їх застосування для двох метрик покращило результати, а для двох — погрішило. Так, повнота збільшилась на 0,11 %, а надійність — на 0,03 %; точність зменшилась на 0,02 %, а метрика F1 — на 0,03 %. Синтетичні дані не дали жодного покращення.

Таблиця 3.2.2

Алгоритм	**Параметри**	**Дані**	**Accuracy**	**Precision**	**Recall**	**F1-score**
Випадковий ліс	-	справжні	0.66	0.78	0.46	0.68
	max_depth=None, min_samples_leaf=2, min_samples_spli	справжні	0.69	0.76	0.57	0.65

	t=2, n_estimators=200	справжні + синтетичні	0.69	0.75	0.57	0.65

Результати якості алгоритму випадкового лісу

Результати тренування моделі RoBERTa та анотування тестових даних моделлю GPT-4 Turbo вказані в табл. 3.2.3. Так, для RoBERTa метрика надійності збільшилась лише на 0,1 % при використанні синтетичних даних, а метрика точності — на 0,8 %. Повнота та метрика F1 зменшились на 0,8 % та 0,2 % відповідно. Щодо великої мовної моделі, то за трьома метриками, а саме надійності, точності та F1, вона має найнижчі значення серед усіх експериментів. Тільки метрика повноти є вищою за всі експерименти з алгоритмом випадкового лісу та за експериментом моделі RoBERTa з синтетичними даними.

Таблиця 3.2.3

Алгоритм	**Дані**	**Accuracy**	**Precision**	**Recall**	**F1-score**
RoBERTa	справжні	0.75	0.78	0.70	0.74
RoBERTa	справжні + синтетичні	0.76	0.86	0.62	0.72
GPT-4 Turbo	-	0.58	0.56	0.68	0.61

Результати якості RoBERTa та GPT-4 Turbo

Також хочемо зауважити, що за графіками тренувальних та валідаційних втрат, а саме зростанням валідаційних втрат та зменшенням тренувальних втрат, можемо зробити висновок, що наприкінці 4 епохи модель RoBERTa була перенавчена (див. Додаток 3). Тобто, алгоритм вивчив детально тренувальні дані, але не зміг узагальнити отримані знання на валідаційних даних. Зважаючи на це, ми вказали параметр load_best_model_at_end, який наприкінці має завантажити найкращу модель з 4 епох.

3.3. Публікація моделей машинного навчання

Для того, щоб усі охочі могли використати навчені моделі, ми вирішили опублікувати їх. Було обрано три натреновані моделі: логістичну регресію без параметрів, алгоритм випадкового лісу з параметрами та модель RoBERTa. Усі моделі під час навчання використовували тільки справжні дані.

Для створення користувацького інтерфейсу ми використали бібліотеку gradio [29]. Зліва є текстове вікно, де користувач може написати текст, для якого потрібно вказати мітку: сарказм, не сарказм. Отриманий на вхід текст буде переданий до трьох вищезгаданих моделей машинного навчання. Після чого справа буде вказано мітку, яку передбачила кожна з моделей.

Gradio дозволяє поширювати посилання на створений інтерфейс користувача, але таке посилання працює обмежену кількість часу, а саме 72 години. Тому кінцеву програму було опубліковано на Hugging Face Spaces [34], що надає постійний доступ до застосунку.

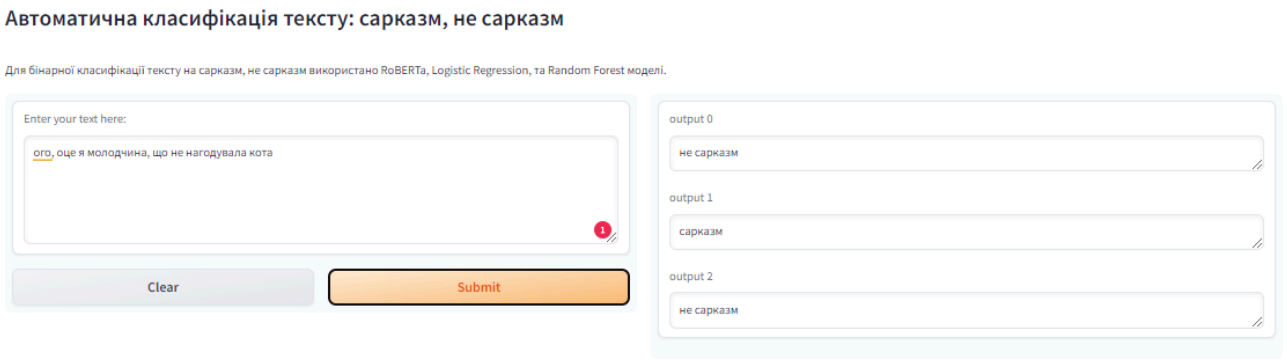


Рисунок 3.3.1. Інтерфейс користувача для класифікації тексту на саркастичний або несаркастичний

Опубліковані моделі були протестовані 5 користувачами (див. Додаток 5). Від них ми отримали 22 тексти, справжню мітку та передбачені моделями мітки, для яких підрахували вищезгадані метрики якості. Отримані результати наведено в табл. 3.3.1.

Моделі логістичної регресії та випадкового лісу мають значно кращі показники за модель RoBERTa. Логістична регресія, як і алгоритм випадкового лісу, отримали однакові результати. Модель RoBERTa передбачає мітки гірше

порівняно з двома іншими моделями. Низький показник метрики повноти вказує на те, що вона часто пропускає реальні саркастичні тексти.

Таблиця 3.3.1

Модель	Accuracy	Precision	Recall	F1-score
RoBERTa	0.59	0.78	0.50	0.61
Logistic Regression	0.86	0.92	0.86	0.89
Random Forest	0.86	0.92	0.86	0.89

Оцінка якості моделей на текстах від користувачів

Висновки до третього розділу

Підсумуємо результати, отримані від навчених нами моделей за найкращими результатами для кожної з метрик на тестових даних.

Метрика	Значення	Модель
Надійність	0.76	RoBERTa з додаванням синтетичних даних
Точність	0.86	RoBERTa з додаванням синтетичних даних
Повнота	0.76	Логістична регресія з параметрами та додаванням синтетичних даних
F1	0.74	RoBERTa зі справжніми даними

Найкращі показники для кожної з метрик серед моделей

Отже, 75 % найкращих значень серед метрик отримала RoBERTa. Також 75 % найкращих значень отримали моделі, що використовували синтетичні дані при тренуванні. Можемо припустити, що лідером серед моделей є RoBERTa саме завдяки своїй архітектурі з використанням трансформерів. Варто зазначити, хоча ми бачимо, що до фінальної таблиці увійшли переважно моделі з використанням синтетичних даних, але, порівнюючи результати в межах однієї моделі, використання синтетичних даних не завжди давало покращення.

Така ж ситуація з коригуванням гіперпараметрів: за деякими метриками для певних моделей вони давали як покращення, так і погіршення або взагалі нічого не змінювали. Причому якщо зміна й відбувалась, то це було менше ніж 1 %. Загалом на результат, який ми отримали, могли вплинути як якість та кількість синтетичних даних, так і обмежений набір значень для пошуку найкращих гіперпараметрів.

При тестуванні опублікованих моделей користувачами ми виявили, що найкращі результати показали моделі на основі традиційних алгоритмів машинного навчання, тоді як модель з архітектурою трансформерів мала гірші значення метрик. Це свідчить про те, що тестування на контрольній вибірці та реальними користувачами може давати різні результати.

ВИСНОВКИ

У цій роботі ми досліджували тему автоматичної класифікації українськомовних текстів на саркастичні або несаркастичні.

У першій частині було визначено, чим сарказм відрізняється від інших видів комічного – гумору, іронії, сатири. Також зроблено огляд підходів та наявних типів текстових даних для вирішення завдання автоматичної ідентифікації сарказму. Крім того, розглянуто такі проблематики: сарказм з боку автора та читача; причини, чому сарказм важко визначити в тексті. Також було описано основні ознаки сарказму на основі аналізу попередніх досліджень з відповідними прикладами ознак в українськомовних текстах.

Сарказм відрізняється від інших видів комічного за недоброзичливістю, агресивністю, наявністю конкретної цілі, відсутністю наміру викликати зміни в суспільстві. Для бінарної класифікації текстів на саркастичні та несаркастичні використовують правила, традиційне машинне навчання та глибоке. Такі моделі навчають на різних даних: довгих і коротких текстах з можливим додаванням контексту. При формуванні навчальних вибірок використовують автоматичну та мануальну розмітку текстів. Вибір джерела текстів, способу збору та анотації даних, відсутність контексту повідомлення, недосконаліть інтелектуальних систем щодо загальновідомих знань — це все те, що робить завдання автоматичного визначення сарказму в текстах ще складнішим. Зрештою, серед основних ознак сарказму було зазначено такі: гіпербола, пунктуаційні знаки, прагматичні ознаки (емотикони, емоджі, написання слова великими літерами), невідповідність, пародіювання російської вимови, де остання ознака характерна саме українськомовним текстам.

У другому розділі ми описали формування набору даних, що містить саркастичні та несаркастичні тексти; порівняли синтетичні саркастичні дані зі справжніми, провели тренування алгоритмів машинного навчання та опублікували три моделі у вільному доступі.

Обсяг фінального набору даних становить 10756 текстів, половина з яких — саркастичні тексти, а саме 3795 повідомлень з платформи “X” та 1583 зібраних нових повідомлень з платформи “Телеграм”. Щодо порівняння реальних та синтетичних даних, то за лексичними статистичними характеристиками ми з’ясували, що багатшу лексику мають все ж таки справжні дані, причому кращі показники серед штучно створених даних має модель від OpenAI. Серед навчених моделей машинного навчання більшість найкращих значень метрик на тестувальній вибірці отримала модель RoBERTa. При тестуванні моделі реальними користувачами кращими ж виявились традиційні алгоритми машинного навчання.

Коригування гіперпараметрів моделей, як і додавання синтетичних даних, змогло покращити значення метрик, але незначною мірою й не для всіх моделей. Ми опублікували три різні моделі машинного навчання у вільному доступі з можливістю написати в текстове поле певне повідомлення та отримати передбачену мітку від відповідних трьох моделей.

Ми бачимо декілька напрямків продовження нашого дослідження в майбутньому:

1. Доповнення навчальної вибірки, що включає не тільки додавання нових текстів з уже використаних платформ, а й пошук інших джерел, зокрема перспективним напрямком є синтетичні дані, адже все частіше з’являються нові мовні моделі або вдосконалюються наявні;
2. Пошук оптимального рішення продукування синтетичних даних великими мовними моделями. Використання різних видів інструкцій та коригування їх змісту, надання інших прикладів реальних даних, зміна параметрів моделей (наприклад, температури) — це все те, що могло б потенційно покращити генерування синтетичних даних, зокрема саркастичних;

3. Експерименти з попередньо тренованими мовними моделями. Тут можна використати не тільки іншу архітектуру (наприклад, згадану BERT), але й застосувати передавальне навчання (transfer learning): перевірити, чи зможе вже навчена модель для автоматичного визначення сарказму в англомовних текстах використати наявні знання для цього ж завдання, але на основі україномовних повідомлень.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Ботвин С. І. Автоматичне визначення сарказму в тексті / Сніжанна Іванівна Ботвин. – Київ: Київський національний університет імені Тараса Шевченка, 2022. – 24 с.
2. Ботвин С. І. Автоматичне визначення сарказму на основі текстів з соціальної мережі Твіттер / Сніжанна Іванівна Ботвин. – Київ: Київський національний університет імені Тараса Шевченка, 2023. – 44 с.
3. Бук С. Статистичні характеристики лексики основних функціональних стилів української мови: спроба порівняння / С. Бук // Лексикографічний бюлетень: Зб. наук. пр. — К.: Ін-т української мови НАН України, 2006. — Вип. 13. — С. 166-170. [Електронний ресурс] — Режим доступу: http://dspace.nbuv.gov.ua/handle/123456789/72846.
4. Гіпербола. ВУЕ. [Електронний ресурс] — Режим доступу: https://vue.gov.ua/%D0%93%D1%96%D0%BF%D0%B5%D1%80%D0%B1%D0%BE%D0%BB%D0%B0.
5. Емотикон // Словотвір. [Електронний ресурс] — Режим доступу: https://slovotvir.org.ua/words/emotykon.
6. Ковалевська Я. Л. САТИРИКО-САРКАСТИЧНИЙ ДИСКУРС У МЕДІАПРОСТОРІ США / Я. Л. Ковалевська // Науковий вісник Міжнародного гуманітарного університету. — 2015. [Електронний ресурс] — Режим доступу: http://www.vestnik-philology.mgu.od.ua/archive/v18/part_2/18.pdf.
7. Пилипенко І. О. Способи передачі емфази при перекладі з англійської мови на українську / Інна Олександрівна Пилипенко // International scientific and practical conference "The latest problems of modern science and practice" / Інна Олександрівна Пилипенко. — USA: International Science Group, 2021. — С. 405—406.

[Електронний ресурс] — Режим доступу: https://books.google.com.ua/books.

8. Цепенюк Т. ВІДТВОРЕННЯ ЛЕКСИЧНИХ ІНТЕНСИФІКАТОРІВ В УКРАЇНСЬКИХ ПЕРЕКЛАДАХ РОМАНІВ Д. СТІЛ / Тетяна Цепенюк // SCIENTIFIC LETTERS OF ACADEMIC SOCIETY OF MICHAL BALUDANSKY / Тетяна Цепенюк., 2017. — С. 135—137. [Електронний ресурс] — Режим доступу: http://dspace.tnpu.edu.ua/bitstream/123456789/24070/3/Tcepenuk.pdf.
9. Ягільнікі І. О. АНГЛО-УКРАЇНСЬКИЙ ПЕРЕКЛАД СЛІВ-ПІДСИЛЮВАЧІВ / І. О. Ягільнікі, Н. В. Денисенко // МОВА. СВІДОМІСТЬ. КОНЦЕПТ / І. О. Ягільнікі, Н. В. Денисенко., 2019. — (Випуск 9). — С. 170—174. [Електронний ресурс] — Режим доступу: https://pedagogy.lnu.edu.ua/wp-content/uploads/2015/03/Sbornyk-9.pdf#page=170.
10. Bamman D., Smith N. A. Contextualized sarcasm detection on twitter. In Proceedings of the 9th International AAAI Conference on Web and Social Media. 2015. [Електронний ресурс] — Режим доступу: https://doi.org/10.1609/icwsm.v9i1.14655.
11. Barbieri F., Saggion H., Ronzano F. Modelling Sarcasm in Twitter, a Novel Approach. Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, Baltimore, Maryland. Stroudsburg, PA, USA, 2014. [Електронний ресурс] — Режим доступу: https://doi.org/10.3115/v1/w14-2609.
12. Bharti S. K. et al. Sarcastic sentiment detection in tweets streamed in real time: a big data approach . Digital Communications and Networks. 2016. Vol. 2, no. 3. P. 108—121. [Електронний ресурс] — Режим доступу: https://doi.org/10.1016/j.dcan.2016.06.002.

13. Bin Liang, Chenwei Lou, Xiang Li, Min Yang, Lin Gui, Yulan He, Wenjie Pei, and Ruifeng Xu. 2022. Multi-Modal Sarcasm Detection via Cross-Modal Graph Convolutional Network. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1767—1777, Dublin, Ireland. Association for Computational Linguistics. [Електронний ресурс] — Режим доступу: https://aclanthology.org/2022.acl-long.124/.
14. Bouazizi M. and Otsuki Ohtsuki T., "A Pattern-Based Approach for Sarcasm Detection on Twitter," in IEEE Access, vol. 4, pp. 5477-5488, 2016. [Електронний ресурс] — Режим доступу: doi: 10.1109/ACCESS.2016.2594194.
15. Bushra N. A Critical Pragmatic Study of Sarcasms in American and British Interviews [Електронний ресурс] / N. Bushra, S. Bushra // 7th International Conference on Multidisciplinary Sciences (7thicomus). — 2021. [Електронний ресурс] — Режим доступу: https://www.researchgate.net/publication/363925404_A_Critical_Pragmatic_Study_of_Sarcasms_in_American_and_British_Interviews.
16. Cari R. Has Social Media Changed How We Use Sarcasm? / Romm Cari // The Cut. — 2016. [Електронний ресурс] — Режим доступу: https://www.thecut.com/2016/05/has-social-media-changed-how-we-use-sarcasm.html.
17. Chakrabarty T. FLUTE: Figurative Language Understanding through Textual Explanations / T. Chakrabarty, D. Ghosh, A. Saakyan // Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing / T. Chakrabarty, D. Ghosh, A. Saakyan., 2022. — С. 7139—7159. [Електронний ресурс] — Режим доступу: https://aclanthology.org/2022.emnlp-main.481/.

18. Common sense // Cambridge Dictionary. [Електронний ресурс] — Режим доступу: https://dictionary.cambridge.org/dictionary/english/common-sense.
19. Counting characters when composing Tweets . [Електронний ресурс] — Режим доступу: https://developer.twitter.com/en/docs/counting-characters.
20. Devlin, Jacob et al. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.” North American Chapter of the Association for Computational Linguistics (2019). [Електронний ресурс] — Режим доступу: https://arxiv.org/abs/1810.04805.
21. Dmitry Davidov, Oren Tsur, and Ari Rappoport. 2010. Semi-Supervised Recognition of Sarcasm in Twitter and Amazon. In Proceedings of the Fourteenth Conference on Computational Natural Language Learning, pages 107—116, Uppsala, Sweden. Association for Computational Linguistics. [Електронний ресурс] — Режим доступу: https://aclanthology.org/W10-2914/.
22. Dress M. L. et al. Regional Variation in the Use of Sarcasm. Journal of Language and Social Psychology. 2008. Vol. 27, no. 1. P. 71—85. [Електронний ресурс] — Режим доступу: https://journals.sagepub.com/doi/pdf/10.1177/0261927X07309512.
23. Emoticon vs. emoji // Dictionary.com [Електронний ресурс] — Режим доступу: https://www.dictionary.com/compare-words/emoticon-vs-emoji.
24. Garmendia J. Irony. University of Cambridge ESOL Examinations, 2020. 178 p. [Електронний ресурс] — Режим доступу: https://www.scribd.com/document/522275061/Irony-Cambridge-University-Press-2018-Key-Topics-in-Semantics-and-Pragmatics.
25. Geron A. Hands-On Machine Learning with Scikit-Learn and TensorFlow / Aurelien Geron. — 2019. [Електронний ресурс] —

Режим доступу: http://powerunit-ju.com/wp-content/uploads/2021/04/Aurelien-Geron-Hands-On-Machine-Learning-with-Scikit-Learn-Keras-and-Tensorflow_-Concepts-Tools-and-Techniques-to-Build-Intelligent-Systems-OReilly-Media-2019.pdf.

26. Ghosh D., Vajpayee A., Muresan S. A Report on the 2020 Sarcasm Detection Shared Task. Proceedings of the Second Workshop on Figurative Language Processing, Online. Stroudsburg, PA, USA, 2020. P. 1-11. [Електронний ресурс] – Режим доступу: https://doi.org/10.18653/v1/2020.figlang-1.1.
27. González-Ibáñez R., Muresan S., Wacholder N. Identifying Sarcasm in Twitter: A Closer Look. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Portland, Oregon, USA, 2011. P. 581—586. [Електронний ресурс] — Режим доступу: https://aclanthology.org/P11-2102/.
28. Govindan V., Balakrishnan V. A machine learning approach in analysing the effect of hyperboles using negative sentiment tweets for sarcasm detection. Journal of King Saud University - Computer and Information Sciences. 2022. [Електронний ресурс] — Режим доступу: https://doi.org/10.1016/j.jksuci.2022.01.008.
29. Gradio [Електронний ресурс] – Режим доступу: https://www.gradio.app/.
30. Grid Search | Dremio. Dremio Unified Analytics Platform for a Self-Service Lakehouse. [Електронний ресурс] — Режим доступу: https://www.dremio.com/wiki/grid-search/.
31. GPT-4 Turbo in the OpenAI API [Електронний ресурс] – Режим доступу:

https://help.openai.com/en/articles/8555510-gpt-4-turbo-in-the-openai-api.

32. Howman H. E., Filik R. The role of emoticons in sarcasm comprehension in younger and older adults: Evidence from an eye-tracking experiment. Quarterly Journal of Experimental Psychology. 2020. Vol. 73, no. 11. P. 1729—1744. [Електронний ресурс] — Режим доступу: https://doi.org/10.1177/1747021820922804.
33. Hugging Face — The AI community building the future. Hugging Face. [Електронний ресурс] — Режим доступу: https://huggingface.co/.
34. Hugging Face Spaces [Електронний ресурс] – Режим доступу: https://huggingface.co/spaces.
35. Joshi A., Bhattacharyya P., Carman M. J. Automatic Sarcasm Detection. ACM Computing Surveys. 2017. Vol. 50, no. 5. P. 1—22. [Електронний ресурс] — Режим доступу: https://doi.org/10.1145/3124420.
36. A. Joshi et al. Harnessing sequence labeling for sarcasm detection in dialogue from TV series "Friends.". Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning. 2016. P. 146—155. [Електронний ресурс] — Режим доступу: https://aclanthology.org/K16-1015.pdf.
37. Joshi A. et al. How Do Cultural Differences Impact the Quality of Sarcasm Annotation?: A Case Study of Indian Annotators and American Text. Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities, Berlin, Germany. Stroudsburg, PA, USA, 2016. [Електронний ресурс] — Режим доступу: https://doi.org/10.18653/v1/w16-2111.
38. Joshi A., Sharma V., Bhattacharyya P. Harnessing Context Incongruity for Sarcasm Detection. Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint

Conference on Natural Language Processing (Volume 2: Short Papers), Beijing, China. Stroudsburg, PA, USA, 2015. [Електронний ресурс] — Режим доступу: https://doi.org/10.3115/v1/p15-2124.

39. Joshi A., Tripathi V., Patel K. et al. Are Word Embedding-based Features Useful for Sarcasm Detection? 2016. [Електронний ресурс] — Режим доступу: https://arxiv.org/abs/1610.00883.
40. Jurafsky D. Naive Bayes and Sentiment Classification / D. Jurafsky, J. Martin // Speech and Language Processing / D. Jurafsky, J. Martin., 2023. — (Third edition). [Електронний ресурс] — Режим доступу: https://web.stanford.edu/~jurafsky/slp3/4.pdf.
41. Kečkeš I. Detecting Sarcasm in Communication on Twitter / Ines Kečkeš // Sveučilište u Zagrebu Filozofski fakultet Odsjek za anglistiku. — 2022. [Електронний ресурс] — Режим доступу: https://repozitorij.ffzg.unizg.hr/islandora/object/ffzg:6479.
42. Khodak M., Saunshi N., and Vodrahalli K. 2018. A Large Self-Annotated Corpus for Sarcasm. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan. European Language Resources Association (ELRA). P. 641 — 646. [Електронний ресурс] — Режим доступу: https://aclanthology.org/L18-1102/.
43. Kim J. Developing conceptual understanding of sarcasm in L2 English through explicit instruction / J. Kim, J. Lantolf // Language Teaching Resource Vol 22(2) [Електронний ресурс] — Режим доступу: https://journals.sagepub.com/doi/pdf/10.1177/1362168816675521.
44. Kreuz R. Irony and Sarcasm. MIT Press, 2020. 232 p. [Електронний ресурс] — Режим доступу: https://pdfcoffee.com/irony-and-sarcasm-by-roger-j-kreuz-pdf-free.html.
45. Kumar L. K., Somani A., Bhattacharyya P. Approaches for Computational Sarcasm Detection: A Survey. Powai Mumbai,

Maharashtra, India, 2017. P. 1— 14. [Електронний ресурс] — Режим доступу: https://www.cfilt.iitb.ac.in/resources/surveys/Greroaches%20for%20Computati onal%20Sarcasm%20Detection:%20A%20Survey.pdf.

46. Kumari K. RoBERTa: A Modified BERT Model for NLP. Comet. [Електронний ресурс] — Режим доступу: https://www.comet.com/site/blog/roberta-a-modified-bert-model-for-nlp.
47. Lin Z. Modeling Intra and Inter-modality Incongruity for Multi-Modal Sarcasm Detection / Z. Lin, H. Pan, P. Fu. — 2020. [Електронний ресурс] — Режим доступу: https://aclanthology.org/2020.findings-emnlp.124.pdf.
48. Liu, Yinhan et al. “RoBERTa: A Robustly Optimized BERT Pretraining Approach.” ArXiv abs/1907.11692 (2019): n. pag. [Електронний ресурс] — Режим доступу: https://arxiv.org/abs/1907.11692.
49. Logistic Regression and regularization: Avoiding overfitting and improving generalization. Medium. [Електронний ресурс] — Режим доступу: https://medium.com/@rithpansanga/logistic-regression-and-regularization-avoiding-overfitting-and-improving-generalization-e9afdcddd09d.
50. Machine Learning Random Forest Algorithm - Javatpoint. [Електронний ресурс] — Режим доступу: https://www.javatpoint.com/machine-learning-random-forest-algorithm.
51. Mehta I. X updates its terms to ban crawling and scraping | TechCrunch. [Електронний ресурс] — Режим доступу: https://techcrunch.com/2023/09/08/x-updates-its-terms-to-ban-crawling-and-scraping/.
52. Moores B., Mago V. A Survey on Automated Sarcasm Detection on Twitter, 2022. [Електронний ресурс] — Режим доступу: https://arxiv.org/pdf/2202.02516.pdf.

53. NLTK [Електронний ресурс] – Режим доступу: https://www.nltk.org/index.html.
54. Noorhayu S. Sarcastic Expressions and the Influence of Social Distanceand Relative Power in The TV Series Friends / S. Noorhayu, A. Munandar // Lexicon Volume 1, Number 1. — 2020. [Електронний ресурс] — Режим доступу: https://journal.ugm.ac.id/lexicon/article/view/64585/30805.
55. OpenAI API [Електронний ресурс] – Режим доступу: https://openai.com/index/openai-api/.
56. Oprea S. V., Magdy W. The Effect of Sociocultural Variables on Sarcasm Communication Online. Proceedings of the ACM on Human-Computer Interaction. 2020. Vol. 4, CSCW1. P. 1—22. [Електронний ресурс] — Режим доступу: https://doi.org/10.1145/3392834.
57. Oprea S., Magdy W. Exploring Author Context for Detecting Intended vs Perceived Sarcasm. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy. Stroudsburg, PA, USA, 2019. [Електронний ресурс] — Режим доступу: https://doi.org/10.18653/v1/p19-1275.
58. Oprea S., Magdy W. iSarcasm: A Dataset of Intended Sarcasm. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online. Stroudsburg, PA, USA, 2020. [Електронний ресурс] — Режим доступу: https://doi.org/10.18653/v1/2020.acl-main.118.
59. Oraby S. et al. Creating and Characterizing a Diverse Corpus of Sarcasm in Dialogue. Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue, Los Angeles. Stroudsburg, PA, USA, 2016. [Електронний ресурс] — Режим доступу: https://doi.org/10.18653/v1/w16-3604.

60. Penalized Logistic Regression Essentials in R: Ridge, Lasso and Elastic Net - Articles - STHDA. STHDA - Accueil. [Електронний ресурс] — Режим доступу: http://www.sthda.com/english/articles/36-classification-methods-essentials/149-penalized-logistic-regression-essentials-in-r-ridge-lasso-and-elastic-net/.
61. Petkova G. The Gold Standard — The Key to Information Extraction and Data Quality Control. ontotext. [Електронний ресурс] — Режим доступу: https://www.ontotext.com/blog/gold-standard-key-to-information-extration-data-quality-control/.
62. Plepi J., Flek L. Perceived and Intended Sarcasm Detection with Graph Attention Networks. Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021), Online. Stroudsburg, PA, USA, 2021. [Електронний ресурс] — Режим доступу: https://doi.org/10.18653/v1/2021.wnut-1.12.
63. Pranali P. Literature survey of sarcasm detection / P. Pranali, C. Chaitali // International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET). — 2017. [Електронний ресурс] — Режим доступу: https://www.semanticscholar.org/paper/Literature-survey-of-sarcasm-detection-Chaudhari-Chandankhede/187ae7fe20591ad8f4174d6c24968ad79673510e.
64. Python. [Електронний ресурс] — Режим доступу: https://www.python.org/.
65. Radchenko V. We Trained the Ukrainian Language Model. YouScan. [Електронний ресурс] — Режим доступу: https://youscan.io/blog/ukrainian-language-model/.

66. Rajadesingan A., Zafarani R., Liu H. Sarcasm detection on twitter: A behavioral modeling approach. In Proceedings of the 8th ACM International Conference on Web Search and Data Mining. 2015. P. 97—106. [Електронний ресурс] — Режим доступу: https://ashwinrajadesingan.com/files/SarcasmDetection.pdf.
67. Riloff E. Sarcasm as Contrast between a Positive Sentiment and Negative Situation / E. Riloff, P. Surve, A. Qadir. — 2013. [Електронний ресурс] — Режим доступу: https://aclanthology.org/D13-1066.pdf.
68. Scare quotes. Cambridge Dictionary | English Dictionary, Translations & Thesaurus. [Електронний ресурс] — Режим доступу: https://dictionary.cambridge.org/dictionary/english/scare-quotes 48y
69. Scikit-learn [Електронний ресурс] — Режим доступу: https://scikit-learn.org/stable/index.html.
70. Shakhnoza R. HYPERBOLE AS A DEVICE OF SPEECH / Rustam Shakhnoza // Academic Research in Educational Science Volume 3 Issue 12. — 2022. [Електронний ресурс] — Режим доступу: https://ares.uz/storage/app/uploads/public/63a/2f6/ff2/63a2f6ff2c79d018856023.pdf.
71. Shmueli B., Ku L.-W., Ray S. Reactive Supervision: A New Method for Collecting Sarcasm Data. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Online. Stroudsburg, PA, USA, 2020. [Електронний ресурс] — Режим доступу: https://doi.org/10.18653/v1/2020.emnlp-main.201.
72. Skalicky S., Crossley S. Linguistic Features of Sarcasm and Metaphor Production Quality. Proceedings of the Workshop on Figurative Language Processing, New Orleans, Louisiana. Stroudsburg, PA, USA, 2018. [Електронний ресурс] – Режим доступу: https://doi.org/10.18653/v1/w18-0902.

73. Sklearn.ensemble.RandomForestClassifier. scikit-learn. [Електронний ресурс] — Режим доступу: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html.
74. Sklearn.linear_model.LogisticRegression. scikit-learn. [Електронний ресурс] — Режим доступу: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html.
75. Snscrape [Електронний ресурс] — Режим доступу: https://github.com/JustAnotherArchivist/snscrape.
76. spaCy [Електронний ресурс] – Режим доступу: https://spacy.io/.
77. Stanza — A Python NLP Package for Many Human Languages [Електронний ресурс] — Режим доступу: https://stanfordnlp.github.io/stanza/.
78. TWINT - Twitter Intelligence Tool [Електронний ресурс] — Режим доступу: https://github.com/twintproject/twint.
79. Telegram Channels. Telegram. [Електронний ресурс] — Режим доступу: https://telegram.org/tour/channels.
80. Tepperman J. Yeah right: Sarcasm recognition for spoken dialogue systems [Електронний ресурс] / J. Tepperman, D. Traum, S. Narayanan. — 2006. [Електронний ресурс] — Режим доступу: https://www.researchgate.net/publication/221491095_Yeah_right_Sarcasm_recognition_for_spoken_dialogue_systems.
81. Test experimental Gemini models – 2024. [Електронний ресурс] — Режим доступу: https://cloud.google.com/vertex-ai/generative-ai/docs/multimodal/gemini-experimental.
82. Thompson D., Filik R. Sarcasm in Written Communication: Emoticons are Efficient Markers of Intention. Journal of Computer-Mediated

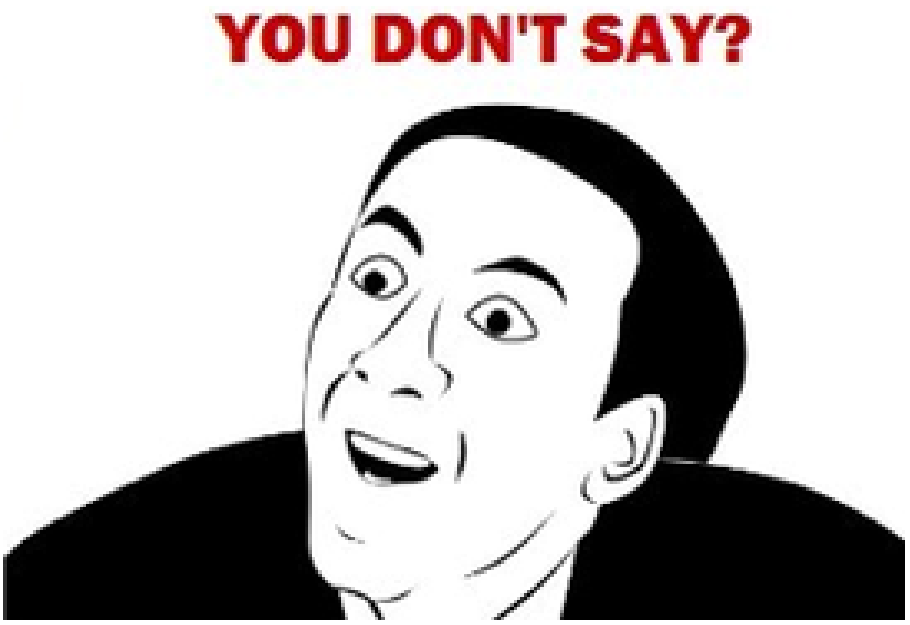
Communication. 2016. Vol. 21, no. 2. P. 105—120. [Електронний ресурс] — Режим доступу: https://doi.org/10.1111/jcc4.12156.

83. Tokenize UK [Електронний ресурс] – Режим доступу: https://tokenize-uk.readthedocs.io/en/latest/readme.html.
84. Van Hee C., Lefever E., Hoste V. SemEval-2018 Task 3: Irony Detection in English Tweets. Proceedings of The 12th International Workshop on Semantic Evaluation, New Orleans, Louisiana. Stroudsburg, PA, USA, 2018. [Електронний ресурс] — Режим доступу: https://doi.org/10.18653/v1/s18-1005.
85. Vandaele J. “Each time we laugh” translated humour in screen comedy / Jeroen Vandaele. – 1999. [Електронний ресурс] — Режим доступу: https://www.semanticscholar.org/paper/%E2%80%9CEach-time-we-laugh%E2%80%9D-translated-humour-in-screen-Vandaele/a8829c30bf216f4741594a3ef628f178bdfc7ae8.
86. Vaswani, Ashish et al. “Attention is All you Need.” Neural Information Processing Systems (2017). [Електронний ресурс] — Режим доступу: https://arxiv.org/abs/1706.03762.
87. Vertex AI. [Електронний ресурс] – Режим доступу: https://cloud.google.com/vertex-ai?hl=uk.
88. Wallace B. C., Choe D. K., Charniak E. Sparse, Contextually Informed Models for Irony Detection: Exploiting User Communities, Entities and Sentiment. Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. 2015. P. 1035—1044. [Електронний ресурс] — Режим доступу: https://aclanthology.org/P15-1100.pdf.
89. What is a Large Language Model? | A Comprehensive LLMs Guide. Elastic — The Search AI Company | Elastic. [Електронний ресурс] — Режим доступу: https://www.elastic.co/what-is/large-language-models.

90. Wilson D. Irony comprehension: A developmental perspective. Journal of Pragmatics. 2013. Vol. 59. P. 40—56. [Електронний ресурс] — Режим доступу: https://doi.org/10.1016/j.pragma.2012.09.016.
91. Woi J. P. L., Juita N. Types of Sarcasm in the Comment Column of Male Netizen on the Youtube Account of Sukmawati Soekarno Putri News Video. Proceedings of the 5th International Conference on Language, Literature, and Education (ICLLE-5 2022). Paris, 2022. P. 52—58. [Електронний ресурс] — Режим доступу: https://doi.org/10.2991/978-2-494069-85-5_7.
92. You don't say? // Know Your Meme. [Електронний ресурс] — Режим доступу: https://knowyourmeme.com/memes/you-dont-say--3.
93. Your Sentiment Precedes You: Using an author's historical tweets to predict sarcasm / A. Khatri et al. Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis. Lisboa, Portugal, 2015. P. 25—30. [Електронний ресурс] — Режим доступу: https://aclanthology.org/W15-2905.pdf.

ДОДАТКИ

Додаток 1. Саркастичний вираз обличчя Ніколаса Кейджа



Додаток 2. Підказки для великих мовних моделей

Модель	**Тип підказки**	**Підказка**
Gemini Experimental	zero-shot	Ти багато спілкуєшся онлайн. Тобі відомо, що сарказм – таке висловлювання, буквальне значення якого відрізняється від того, яке мовець насправді має на увазі. Саркастичні тексти можуть містити такі ознаки: гіпербола, пунктуаційні знаки, прагматичні ознаки (емотикони, емоджі, великі літери), невідповідність, пародіювання російської вимови. Згенеруй 50 саркастичних текстів. Тексти повинні бути унікальні та покривати різні теми. Не використовуй слова сарказм, сарк, sarcasm, sarc у тексті. Поверни JSON у такому вигляді: {"тексти": list[саркастичні тексти]}
Gemini Experimental	one-shot	Ти багато спілкуєшся онлайн. Тобі відомо, що сарказм – таке висловлювання, буквальне значення якого відрізняється від того, яке мовець насправді має на увазі. Саркастичні тексти можуть містити такі ознаки: гіпербола, пунктуаційні знаки, прагматичні ознаки (емотикони, емоджі, великі літери), невідповідність, пародіювання російської вимови. Згенеруй 50 саркастичних текстів. Тексти повинні бути унікальні та покривати різні теми. Не використовуй слова сарказм, сарк, sarcasm, sarc у тексті. Поверни JSON у такому вигляді: {"тексти": list[саркастичні тексти]}. Приклад: {"тексти": list["<користувач> що тут сказати, Савченко об'єднує країну!! Так тримати!1"]}
Gemini Experimental	few-shot	Ти багато спілкуєшся онлайн. Тобі відомо, що сарказм – таке висловлювання, буквальне значення якого відрізняється від того, яке мовець насправді має на увазі. Саркастичні тексти можуть містити такі ознаки: гіпербола, пунктуаційні знаки, прагматичні ознаки (емотикони, емоджі, великі літери), невідповідність, пародіювання російської вимови. Згенеруй 50 саркастичних текстів. Тексти повинні бути унікальні та

		покривати різні теми. Не використовуй слова сарказм, сарк, sarcasm, sarc у тексті. Поверни JSON у такому вигляді: {"тексти": list[саркастичні тексти]}. Приклад: {"тексти": list["<користувач> <користувач> Вона не голосувала за пукіна….", "Світу до росіян не існувало, запам'ятайте\n", "Я не зрозумію мабуть ніколи, чому комусь лінь шити взуття по стандарту.Чому треба обов‘язково нашити купу шибздиків, і називати їх «МАЛОМЄРКІ»\nНАШО? Це прояв пекла на землі? Шиють взуття агенти диявола?""]}
GPT-4 Turbo	zero-shot	Ти багато спілкуєшся онлайн. Тобі відомо, що сарказм – таке висловлювання, буквальне значення якого відрізняється від того, яке мовець насправді має на увазі. Саркастичні тексти можуть містити такі ознаки: гіпербола, пунктуаційні знаки, прагматичні ознаки (емотикони, емоджі, великі літери), невідповідність, пародіювання російської вимови.Згенеруй 50 саркастичних текстів. Тексти повинні бути унікальні та покривати різні теми. Не використовуй слова сарказм, сарк, sarcasm, sarc у тексті. Поверни правильний RFC8259 JSON без жодних відхилень у такому форматі: {"тексти": list[саркастичні тексти]}
GPT-4 Turbo	one-shot	Ти багато спілкуєшся онлайн. Тобі відомо, що сарказм – таке висловлювання, буквальне значення якого відрізняється від того, яке мовець насправді має на увазі. Саркастичні тексти можуть містити такі ознаки: гіпербола, пунктуаційні знаки, прагматичні ознаки (емотикони, емоджі, великі літери), невідповідність, пародіювання російської вимови.Згенеруй 50 саркастичних текстів. Тексти повинні бути унікальні та покривати різні теми. Не використовуй слова сарказм, сарк, sarcasm, sarc у тексті. Поверни правильний RFC8259 JSON без жодних відхилень у такому форматі: {"тексти": list[саркастичні тексти]}. Приклад: {"тексти": list["<користувач> що тут сказати, Савченко об'єднує країну!! Так тримати!1"]}

GPT-4 Turbo	few-shot	Ти багато спілкуєшся онлайн. Тобі відомо, що сарказм – таке висловлювання, буквальне значення якого відрізняється від того, яке мовець насправді має на увазі. Саркастичні тексти можуть містити такі ознаки: гіпербола, пунктуаційні знаки, прагматичні ознаки (емотикони, емоджі, великі літери), невідповідність, пародіювання російської вимови.Згенеруй 50 саркастичних текстів. Тексти повинні бути унікальні та покривати різні теми. Не використовуй слова сарказм, сарк, sarcasm, sarc у тексті. Поверни правильний RFC8259 JSON без жодних відхилень у такому форматі: {"тексти": list[саркастичні тексти]}. Приклад: {"тексти": list["<користувач> <користувач> Вона не голосувала за пукіна….", "Світу до росіян не існувало, запам'ятайте\n", "Я не зрозумію мабуть ніколи, чому комусь лінь шити взуття по стандарту.Чому треба обов‘язково нашити купу шибздиків, і називати їх «МАЛОМЄРКІ»\nНАШО? Це прояв пекла на землі? Шиють взуття агенти диявола?""]}

Додаток 3. Графіки тренувальних та валідаційних втрат для моделі RoBERTa

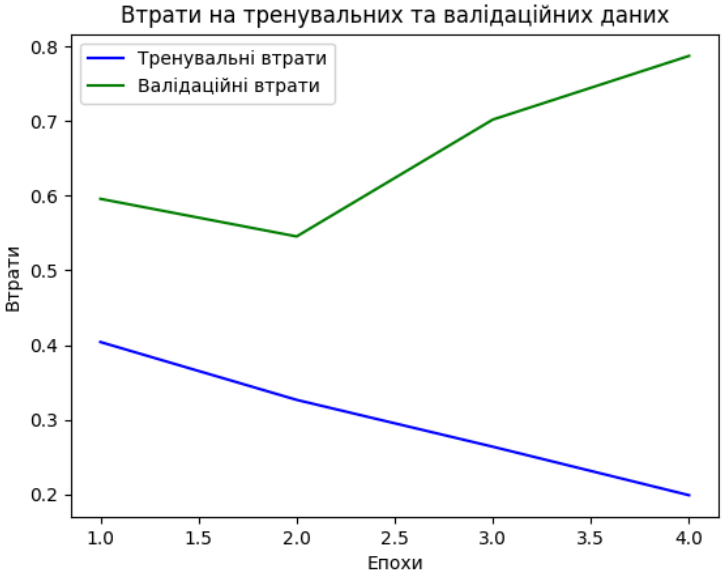


Рисунок 1. Тренувальні та валідаційні втрати для моделі RoBERTa зі справжніми даними

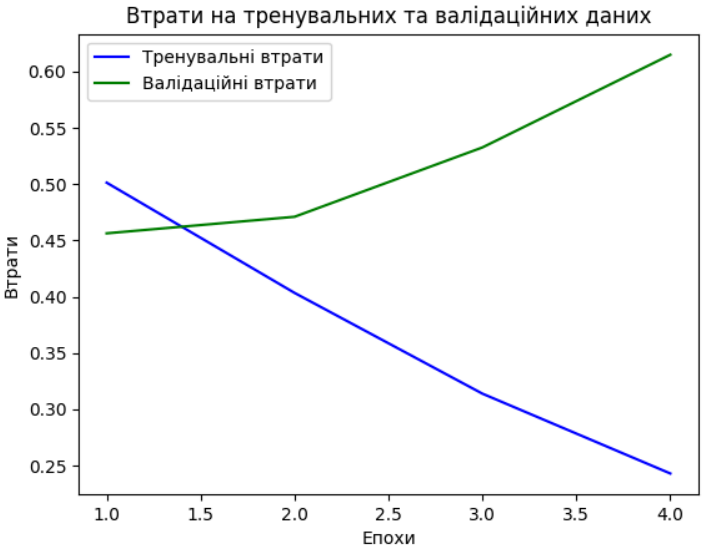


Рисунок 2. Тренувальні та валідаційні втрати для моделі RoBERTa з синтетичними даними

Додаток 4. Результати опитування про синтетичні та справжні саркастичні дані:

https://docs.google.com/spreadsheets/d/10K0B9P0I1O-httEY_w0BBt8ggavAiNQDCtrArzCzNVo/edit?usp=sharing

Додаток 5. Результати опитування про якість моделей від користувачів

Тексти	Справжня мітка	RoBERTa	Logistic Regression	Random Forest
о так, література дуже важливий предмет для програмістів	1	1	1	1
Звичайно, ти абсолютно правий — хто взагалі потребує здорового глузду, коли можна покладатися на щастя і випадковість?	1	0	1	1
Звісно, ти - експерт у кожній справі.	1	0	1	0
побудувати школу біля порохової вежі - просто геніальна ідея!	1	1	1	1
ну бо порох - наш гетьман, довіряю йому і більше нікому	1	0	1	1
ну звісно, от відірвусь від своїх справ і допоможу тобі	1	0	1	1
Ну звісно, найкращий спосіб вирішити всі проблеми — це просто нічого не робити. Хтось же інший точно все владнає.	1	1	1	1
Дуже цікаво почути твою думку.	1	0	0	1
люблю читати фармацевтичні книги :)	0	0	0	0
ага, а українці викопали Чорне море	1	1	1	1
це справді цікава тема	0	0	0	0
Ну, якщо ти думаєш, що понеділок — це найгірший день тижня, просто згадай, що є ще й вівторок!	0	1	0	1
Так, це дійсно прекрасна ідея - піти на прогулянку зараз, коли падає дощ.	1	0	1	1
ооо, так, давайте поставте ще дедлайнів	1	0	1	1

спати 2 години за ніч - вбивсто ментального здоров'я	0	1	0	0
з цими нескінченними семінарами можна застрелитись	1	1	0	0
Або я диплом, або він мене.	0	0	0	0
Це було дуже важко, але ми зробили це разом.	0	0	0	0
Львів - місто дощу	1	1	1	1
котики люблять їсти домашній сир	0	0	1	0
Росія - наймогутніша держава на планеті Земля	1	1	1	1
Твоя праця справді вражає	0	0	0	0

Додаток 6. Покликання на дані та моделі машинного навчання

1. Справжня навчальна вибірка з саркастичними та несаркастичними текстами: https://github.com/botvyns/bachelor_project/blob/main/1data_merge/dataset.csv
2. Синтетичні саркастичні дані від Gemini Experimental та GPT-4 Turbo: https://github.com/botvyns/bachelor_project/blob/main/2sampling_synth_real/synthetic_data_combined.csv
3. Опублікована модель на основі алгоритму логістичної регресії: https://huggingface.co/spaces/Snizhanna/sarcasm_detection/blob/main/lr_classifier_default.pkl
4. Опублікована модель на основі алгоритму випадкового лісу: https://huggingface.co/spaces/Snizhanna/sarcasm_detection/blob/main/rf_classifier_param.pkl
5. Опублікована модель RoBERTa: https://huggingface.co/Snizhanna/ukr-roberta-base-finetuned-sarc
6. Система для класифікації тексту на саркастичний, несаркастичний: https://huggingface.co/spaces/Snizhanna/sarcasm_detection

Додаток 7. Структура GitHub репозиторію:

https://github.com/botvyns/bachelor_project

1. Тека **0scraping**: дані з Телеграм каналу та скрипт для збирання даних;
2. Тека **1data_merge:** записник з кодом для об'єднання даних з платформ "Telegram", "X".

   Вхідні дані:

   1. `twitter_sarcastic_not_sarcastic.csv` (попередній набір даних з "X" з текстом двох класів: сарказм, не сарказм);
   2. `telegram_sarcastic.csv` (тільки саркастичні повідомлення з conversational threads);
   3. `telegram_not_sarcastic_23k.csv` (не саркастичні повідомлення з "Telegram").

   Вихідні дані:

   1. `dataset.csv` (об'єднані дані з двох платформ);
   2. `dataset_cleaned.csv` (видалено дублікати, забезпечено баланс класів).
3. Тека **2sampling_synth_real**: об'єднання синтетичних даних з Gemini, GPT-4 Turbo та створення рівних за словами вибірок для порівняння реальних/синтетичних даних.

   Вхідні дані:

   1. `dataset_cleaned.csv`;
   2. TXT-файли зі згенерованими синтетичними саркастичними даними від Gemini Experimental, GPT-4 Turbo без підказки, з однією підказкою, з кількома підказками.

   Вихідні дані:

   1. `synthetic_data_combined.csv`;
   2. `openai_sample.csv` (вибірка з GPT-4 Turbo);
   3. `gemini_sample.csv` (вибірка з Gemini Emerimental);
   4. `real_sarc_sample.csv` (вибірка зі справжніх саркастичних даних).

4. Тека **3comparing_synth_real**: обчислення лексичних статистичних характеристик, найчастотніших іменованих сутностей, біграм, триграм.

   Вхідні дані:

   1. `real_sarc_sample.csv`, `gemini_sample.csv`, `openai_sample.csv` (рівні за кількістю слів вибірки для порівняння синтетичних та саркастичних даних).

   Вихідні дані:

   1. `gemini_sample_lemmatized_ents_ngrams.csv`,`openai_sample_lemmatized_ents_ngrams.csv`,
      `real_sarc_sample_lemmatized_ents_ngrams.csv` (рівні за словами вибірки з інформацією про леми, іменовані сутності, біграми, триграми).
5. Тека **4prep_data_for_models**: попередня обробка даних для алгоритмів машинного навчання.

   Вхідні дані:

   1. `dataset_cleaned.csv`;
   2. `telegram_not_sarcastic_sample_left.csv` (частина `telegram_not_sarcastic_23k.csv`, яка не була використана для додавання несаркастичних даних до `dataset.csv`);
   3. `synthetic_data_combined.csv`.

   Вихідні дані:

   1. `dataset_ready_for_models.csv` (додатково очищені дані, готові для алгоритмів машинного навчання);
   2. `synth_openai_sarc_and_not.csv` (синтетичні саркастичні дані GPT-4 Turbo та несаркастичні дані Telegram, збалансовані за класами).
6. Тека **5models_training**: тренування алгоритмів машинного навчання.

   Вхідні дані:

   1. `dataset_ready_for_models.csv `;

2. `synth_openai_sarc_and_not.csv`.

Вихідні дані:

1. `test_ready_for_models.csv`, `train_ready_for_models.csv` (тренувальна, тестувальна вибірки з `dataset_ready_for_models.csv`).