Кваліфікаційна робота

на здобуття ступеня бакалавра

за спеціальністю 113 Прикладна
математикана тему:

Аналіз тональності тексту та обробка природної мови за допомогою штучного інтелекту для розуміння суспільних настроїв щодо повномасштабного вторгнення Російської Федерації в Україну

Виконала студентка 4-го
курсу
Гашева Аліна Михайлівна

Науковий
керівник:асистент
Денисов Сергій Вікторович

Засвідчую, що в цій роботі немає запозичень з праць інших авторів без відповідних посилань.

Студентка

Роботу розглянуто й допущено до захисту на засіданні кафедри обчислювальної математики
«29»травня 2023 р.,
протокол № 8
Завідувач кафедри

С. І. Ляшко

Київ-2023

РЕФЕРАТ

Обсяг роботи 42 сторінки, 12 ілюстрацій, 1 таблиця, 30 використаних джерел.

Об’єктом дослідження є методи наївного баєса та алгоритми на основі опорних векторів для тонального аналізу тексту, пов’язаного із повномасштабним вторгненням Російської Федерації в Україну. Було також розглянуто алгоритми тонального аналізу тексту без попереднього навчання, сфери використання аналізу тональності тексту та обробки природної мови.

Метою роботи є тренування 10 моделей для подальшого використання в цілях аналізу тональності тексту, що найкраще підходять для аналізу тексту, отриманого із соціальних мереж. Також метою роботи було дослідити на емоційне забарвлення коментарі під двома відео, що вийшли на самому початку повномасштабного вторгнення: відео зі зверненням президента Російської Федерації Володимира Путіна до жителів Донбасу та відповідь президента України Володимира Зеленського на погрозу В. Путіна розпочати спеціальну воєнну операцію на території Донбасу.

У роботі було виконане теоретичне дослідження існуючих та вибір двох основних методів на основі машинного навчання для проведення дослідження. Було проведене навчання моделей та було застосовано найкраще натреновану модель для аналізу суспільних настроїв щодо повномасштабного вторгнення за допомогою методів та бібліотек мови програмування Python. Результати дослідження було представлено у графічному вигляді.

ЗМІСТ

ВСТУП

Володіючи бізнесом, створюючи продукт, продаючи послуги або керуючи державою, займаючись будь-чим, що включає у себе важливість врахування суспільної думки, думки великої групи людей, цільової авдиторії, людина неодмінно зіштовхується із питанням збору відгуків, думок. Поява всесвітньої комп'ютерної мережі інтернет надала користувачам можливість висловлювати свою думку стосовно будь-якої речі, події, людини або продукту у текстовій формі або за допомогою картинок, коротких відео, емоджі тощо. Закономірно, зі зростанням кількості відгуків користувачів, зростала необхідність автоматизації їх збору, обробки та отримання з них корисної інформації, що підкаже подальший вектор розвитку зацікавлених сторін. Аналіз великої кількості повідомлень вручну займає занадто багато часу та трудових ресурсів і є невигідним з багатьох точок зору.

Обробка природної мови, зокрема аналіз тональності тексту, дозволяє автоматизувати збір цінної інформації з текстових повідомлень, вловлюючи їх емоційне забарвлення, а таким чином і відношення автора повідомлення до досліджуваного предмету, явища. Аналіз тональності тексту має набагато ширше коло використання, ніж просто збір інформації про продукт або явище, зокрема його використовують для покращення якості машинного перекладу тексту, для покращення комп'ютером розуміння людської природної мови і навчання комп'ютера генерувати адекватні, змістовні відповіді та реакції на написаний людиною текст (наприклад, у чат-ботах) і більш глобально для розвитку штучного інтелекту (надалі – ШІ) до того рівня, коли будь-який згенерований ШІ текст виглядав би настільки природньо, що його б неможливо було відрізнити від людської мови.

Останній найвідоміший на даний момент прорив у тренуванні ШІ роботі з обробкою природної мови та аналізом тональності тексту є чат-бот ChatGPT від компанії OpenAI. І хоча далеко не всі повідомлення та відповіді від цього

чат-бота є змістовними, правильними або такими, що їх складно відрізнити від відповідей або повідомлень живого співбесідника, він є дійсно потужним результатом у галузі ШІ, що працюють із природною мовою, а також є дуже наглядним відображенням швидкості розвитку технологій ШІ та росту їх популярності за останні роки.

Серед користувачів також все більшу популярність набуває ШІ, що можу імітувати діалог з персонажем книги, гри або фільму, підлаштовуючись під настрій живого користувача. Цей ШІ також було натреновано саме на роботу з аналізом тональності тексту, щоб зробити для користувачів комфортний, персоналізований простір для ведення діалогу. Деякі користувачі зазначали, що спілкування із таким ШІ має терапевтичний і заспокоюючий ефект, а не є лише розвагою.

У сучасному світі інформація є найціннішим активом. Компанії, державні органи, ЗМІ прагнуть якнайшвидше отримувати добре перевірену інформацію, якій можна довіряти. Моніторинг думок людей стосовно продуктів, новин, публічних осіб, явищ відбувається безперервно. Сам цей процес і якість його проведення має великий вплив на світ, що формується довкола нас.

Метою роботи є розробка веб-додатку на базі мови програмування Python та веб-фреймворків із можливістю завантаження файлів у текстовому форматі, подальшій обробці отриманого тексту за допомогою заздалегідь навченої моделі методами аналізу тональності тексту, і повернення користувачу результатів аналізу у графічному та текстовому вигляді.

Об’єктом дослідження є тональність написаного людиною тексту. Предметом дослідження виступають існуючі на даний момент методи аналізу тональності тексту та обробки природної мови.

Результатом проведеної роботи буде готовий до подальшого використання веб-додаток та API із навченими моделями для аналізу емоційного забарвлення природної мови.

РОЗДІЛ 1. Аналіз предметної галузі

1.1 Аналіз тональності тексту та обробка природної мови

Обробка природної мови (Natural language processing або скорочено NLP) є міждисциплінарною галуззю, що стоїть на перетині лінгвістики, комп’ютерних наук та ШІ, основною проблематикою якої є вивчення проблем комп'ютерного аналізу природної людської мови та її синтез. Під комп'ютерним аналізом розуміється здатність машини, що обробляє текст, отримувати корисну закладену у ньому інформацію, основну думку, яку намагається передати автор тексту. Під синтезом мови розуміється навчання машини генерувати наскільки можливо усвідомлені, змістовні, близькі до природної мови повідомлення.

Аналізом тональності тексту (Sentiment analysis) називають процес та набір методів комп’ютерної лінгвістики для аналізу тексту, основним призначенням яких є розпізнавання емоційного забарвлення тексту, визначення відношення автора до предмета, людини або явище, про які йдеться у тексті. Основними категоріями оцінки відношення автора тексту є позитивна, нейтральна та негативна оцінка. Емоційне забарвлення лексики підпадає під ті самі категорії.

Аналіз тональності тексту можна вважати підгалуззю обробки природної мови, однією з найбільш уживаних на практиці. Як обробка природної мови, так і аналіз тональності тексту використовують методи ШІ, глибокого та машинного навчання для розв’язання поставлених задач. Обробка природної мови, будучи більш загальною галуззю, ставить перед собою більший набір задач та має більше сфер застосування, ніж аналіз тональності тексту. Одними з основних застосувань обробки природної мови є:

- Розпізнавання усного мовлення та переведення усної мови у текстовий формат;
- Тегування мови – процес розпізнавання та виділення у тексті частин мови;
- Автокорекція та автодоповнення письмового тексту, що часто використовується у інтегрованих середовищах розробки, текстових редакторах (таких як Microsoft Word) тощо;
- Класифікація письмових повідомлень;
- Машинний переклад тексту;
- Написання коротких підсумків на великі об’єми тексту;
- Розпізнання контексту використання різних слів для позначення одного й того самого об’єкту у тексті;
- Генерація повідомлень, схожих на текст, написаний людиною.

Аналіз тональності тексту застосовують для виявлення емоційного забарвлення повідомлень і отримання корисних висновків з результатів проведеного аналізу.

1.2 Види тонального аналізу

Поділяти аналіз тональності тексту на різні види можна за різними критеріями. За критерієм розміру аналізованого тексту, кількості слів у тексті, тональний аналіз поділяють на крупнозернистий (coarse-grained) та дрібнозернистий (fine-grained). Як видно із назви, вхідними даними крупнозернистого аналізу є великі об’єми тексту: статті, документи, глави книг та іноді навіть речення вважаються об’єктами крупнозернистого аналізу. Крупнозернистий аналіз ставить перед собою задачу розпізнавання емоційного забарвлення тексту загалом, а не окремих його частин чи слів. Дрібнозернистий аналіз займається виділенням настрою та розпізнанням об’єкту чи явища, про які йде мова у реченнях, частинах речення. Дрібнозернистий аналіз часто використовується для збору цінних даних з відгуків на продукти на сайтах або з думок людей, висловлених

у дописах у соціальних мережах, таких як Twitter або Instagram, де для користувачів існує чіткий ліміт на кількість символів, які можна використовувати у дописах.

Дрібнозернистий аналіз, на відміну від крупнозернистого, допомагає сконцентрувати увагу аналітика на конкретній ознаці, функції або особливості об’єкта або явища, про які йде мова у тексті, а не на усьому об’єкті або явищі. Наприклад, якщо досліджуваним об’єктом є ноутбук, за допомогою дрібнозернистого аналізу можна буде дізнатися, що автор відгуку на ноутбук негативно описує ємність батареї ноутбуку, але крупнозернистий аналіз покаже, що загальне враження автора відгуку було позитивним.

Більшість сучасних систем, які автоматично визначають емоційний рейтинг тексту, використовують одновимірний емоційний простір: позитивний чи негативний. Однак відомі також успішні приклади використання багатовимірних просторів. Основне завдання тонального аналізу - класифікувати полярність тексту, тобто визначити, чи є думка, висловлена в тексті, позитивною, негативною або нейтральною. Більш детальна категоризація тональності виражається, наприклад, емоційними станами, такими як "гнів", "смуток" або "радість". Полярність документа можна визначити за бінарною шкалою. У цьому випадку для визначення полярності документа використовуються два класи оцінок - позитивні або негативні. Недоліком такого підходу є те, що не завжди можна однозначно визначити афективну складову документа, тобто документ може містити ознаки як позитивних, так і негативних оцінок.

До ранніх робіт у цій галузі належать роботи Терні та Панга, які застосували різні методи розпізнавання полярності. Це, наприклад, роботи на рівні документа: можна класифікувати полярність документа за багатосмуговою шкалою, як це зробили Панг і Снайдер. Вони розширили основне завдання класифікації рецензій на фільми з "позитивної чи негативної" оцінки до

прогнозування оцінки за три- або чотирибальною шкалою. Водночас Снайдер детально проаналізував відгуки про ресторани і спрогнозував оцінки за різними атрибутами, такими як їжа та атмосфера (за п'ятибальною шкалою). Ще одним із способів визначення тону - використання системи шкалювання для присвоєння чисел за шкалою від -10 до 10 (від найнегативнішого до найпозитивнішого) словам, які зазвичай асоціюються з негативним, нейтральним або позитивним тоном. Спочатку фрагменти неструктурованого тексту вивчаються за допомогою інструментів і алгоритмів обробки природної мови, а витягнуті з цього тексту об'єкти і терміни аналізуються, щоб зрозуміти значення цих слів.

Ще одна область досліджень - суб'єктивна/об'єктивна ідентифікація. Це завдання зазвичай визначається як віднесення даного тексту до одного з двох класів - суб'єктивного та об'єктивного. Текст, зокрема речення, відносять до об’єктивного класу, якщо у них описано загальний факт, новину, або взагалі без оцінки цього факту автором, або з оцінкою вираженою таким чином, що її було б важко швидко виділити і проаналізувати. До суб’єктивного класу найчастіше відносять невеликі частини тексту або частини речення, у яких автор виражає своє ставлення до об’єкту, явища, їх особливості або функції. Тексти суб’єктивного класу легша аналізувати за допомогою методів аналізу тональності тексту. Ця проблема може бути складнішою, ніж полярна класифікація. Суб'єктивність слова або фрази може залежати від контексту, а об'єктивний документ може містити суб'єктивний текст (наприклад, новинна стаття, у якій наведено думки людей). Крім того, результати сильно залежать від визначення суб'єктивності, використовуваного в текстовій анотації. Проте, Панг показав, що точність результатів можна підвищити, видаливши об'єктивні речення з документа до класифікації полярності.

Детальнішою моделлю аналізу є аналіз на основі ознак/аспектів. Ця модель передбачає використання думок і почуттів, виражених різними

характеристиками або аспектами об'єкта, наприклад, мобільного телефону, цифрової камери або банку. Характеристики/аспекти - це атрибути або компоненти об'єкта, досліджуваного на предмет синхронності, наприклад, екран мобільного телефону або якість фотоапарата. Проблема передбачає розв'язання низки завдань, як-от ідентифікація відповідного об'єкта, вилучення його ознак/аспектів і визначення того, чи є думка, висловлена за кожною ознакою/аспектом, позитивною, негативною або нейтральною.

Існує безліч типів аналізу настрою, і інструменти аналізу тону варіюються від систем, що фокусуються на полярності (позитивний, негативний, нейтральний), до систем, що визначають емоції та почуття (наприклад, гнів, радість, смуток) і визначають наміри (наприклад, зацікавлений проти незацікавленого). У наступному розділі розглядаються найважливіші з них. Можливо, основною задачею аналітика стане більш точно визначити рівень полярності думки, тому замість просто позитивних, нейтральних і негативних думок необхідно буде розглянути такі категорії, як: Дуже позитивна - Позитивна - Нейтральна – Негативна - Дуже негативна. Це може бути, наприклад, присвоєння п'ятизіркового рейтингу у відгуку: дуже позитивний = п'ять зірок, дуже негативний = одна зірка тощо. Існують також системи, які забезпечують різні відтінки полярності, визначаючи, чи позитивні, чи негативні емоції пов'язані з певними почуттями, такими як гнів, смуток, занепокоєння (тобто негативні емоції), чи щастя, любов, ентузіазм тощо. На теперішній час є багато різних систем виявлення емоцій які покладаються на словники (списки слів і емоцій, які вони представляють) і складні алгоритми машинного навчання. Недоліком використання словника є те, що люди можуть виражати різні, та навіть полярні емоції одними й тими самими словами або конструкціями, тому для вирішення такої проблеми необхідно використовувати методи, що забезпечують роботу із контекстом повідомлень. Такі слова, як «жах» та «божевілля», часто виражають гнів, наприклад, «Ваш продукт жахливий», «Ціни у вашому магазині

божевільні», але також можуть виражати радість, наприклад, «Цей фільм настільки жахливий, що аж хороший», або «Виступ Хорватії на Євробаченні 2023 року показав справжнє європейське божевілля». Остання фраза має під собою позитивний підтекст, але зчитати його без додаткового контексту досить важко.

Аналіз природної мови і виділення у повідомленнях, автором яких є людина, емоційно забарвленої лексики є дуже важким завданням через складність і багатогранність людської мови, наявність у ній таких речей, як сарказм, іронія, віддзеркалювання, контекст та підтексти, які буває складно зчитати навіть живій людині. Людська мова розвивалася впродовж століть і продовжує розвиватись і сьогодні, з'являється все більше нових слів для позначення нових об'єктів та явищ, які з'явились лише у останні десятиріччя, багато мов доповнюються запозиченими з інших мов словами, мовними конструкціями та контекстом, пов'язаним із ними.

1.3 Індикатори аналізу тональності тексту

Є чимала кількість прийомів за допомогою яких можна отримати показники ефективності, щоб оцінити класифікатори та усвідомити точність моделі аналізу настроїв. Тепер розглянемо індикатори аналізу тональності тексту:

- Полярність настроїв: полярність настроїв стосується того, чи виражає текст позитивне, негативне чи нейтральне почуття. Це можна виміряти за допомогою різних підходів, таких як методи на основі правил, словників або машинного навчання. Отримана оцінка настрою може бути за шкалою від -1 до +1, де -1 означає найбільш негативний настрій, 0 вказує на нейтральний настрій, а +1 вказує на найбільш позитивний настрій.

- Суб'єктивність: суб'єктивність стосується ступеня, до якого текст виражає думку чи особисту точку зору. Його можна виміряти за шкалою

від 0 до 1, де 0 означає, що текст цілком об’єктивний, а 1 означає, що текст цілком суб’єктивний. Наприклад, новинна стаття, яка розповідає про певний факт, матиме низьку суб’єктивність, тоді як редакційна стаття, яка висловлює думку, матиме високу суб’єктивність.

- Емоційність: емоційність означає ступінь емоційної інтенсивності, яку передає текст. Його можна виміряти за шкалою від 0 до 1, де 0 означає, що текст позбавлений емоцій, а 1 означає, що текст дуже емоційний. Наприклад, текст, який передає гнів або радість, буде мати високу емоційність, тоді як текст, який передає інформацію без будь-якого емоційного вираження, матиме низьку емоційність.

- Тон: тон відноситься до загального ставлення або враження, яке передає текст, яке може бути позитивним, негативним або нейтральним. Його можна виміряти за допомогою різних методів, включаючи аналіз настроїв або ручну оцінку загальної тональності тексту.

- Інтенсивність: інтенсивність означає ступінь сили або інтенсивності почуттів або емоцій, переданих у тексті. Його можна виміряти, дивлячись на інтенсивність вживаних слів або інтенсивність мовних моделей у тексті.

- Контекст: контекст тексту може бути важливим показником тональності. Наприклад, у тексті можуть використовуватися позитивні слова, але загальний настрій може бути негативним через контекст тексту. Іронія та сарказм. Іронію та сарказм може бути складно виявити, і для цього можуть знадобитися вдосконалені методи обробки природної мови. Тексти, в яких використовується іронія чи сарказм, можуть виглядати як позитивними, так і негативними, але їх справжня тональність може бути протилежною до того, що пропонують слова.

- Зміни тону: зміни тону можуть вказувати на зміни у ставленні або перспективі автора та можуть бути важливим показником загальної тональності. Зміни тону можуть відбуватися з часом або в різних частинах тексту.

- Предметні знання: у деяких випадках для точного аналізу тональності тексту можуть знадобитися предметні знання. Наприклад, технічний жаргон або ідіоми, характерні для певної галузі, можуть вимагати спеціальних знань, щоб зрозуміти їхні почуття. Аналізуючи ці показники, можна отримати повне розуміння настроїв, емоцій і тону тексту, що дозволяє нам краще зрозуміти ставлення та думки, висловлені в тексті.

- Лексичні ознаки: Лексичні ознаки – це конкретні слова чи фрази, які зазвичай асоціюються з певним почуттям чи емоцією. Наприклад, слова «щасливий», «радісний» і «схвильований» зазвичай асоціюються з позитивними настроями, тоді як слова «сумний», «розчарований» і «злий» зазвичай асоціюються з негативними настроями. Аналіз наявності та частоти цих лексичних ознак може бути ефективним способом визначення тональності тексту.

- Граматичні моделі. Граматичні моделі стосуються способу структури та організації слів у реченні чи тексті. Певні граматичні моделі зазвичай асоціюються з позитивними чи негативними настроями, наприклад використання позитивних прикметників або негативних прислівників. Аналіз використання граматичних моделей може дати розуміння загальної тональності тексту.

- Стилістичні прийоми: стилістичні прийоми – це літературні прийоми, які використовуються для передачі значення, наприклад метафори, порівняння та аналогії.

- Культурні та соціальні фактори: Культурні та соціальні фактори можуть впливати на тональність тексту. Наприклад, певні культури можуть цінувати вираження емоцій більше, ніж інші, що може вплинути на емоційний зміст тексту.

Ці засоби можна використовувати для передачі позитивних або негативних настроїв, залежно від контексту та задуму автора. Аналіз використання стилістичних прийомів може дати розуміння емоційного та тонального змісту тексту.

Подібним чином соціальні чинники, такі як політичні чи ідеологічні уподобання, можуть впливати на тон і почуття тексту. Загалом, індикатори аналізу тональності тексту можна використовувати разом, щоб отримати всебічне розуміння настрою, емоцій і тону тексту. Аналізуючи ці показники, можна отримати цінну інформацію про ставлення та думки, висловлені в тексті, і зрозуміти емоційний вплив використаної мови. Таким чином, індикатори аналізу тональності тексту можна використовувати в поєднанні, щоб отримати повне розуміння настроїв, емоцій і тону тексту. Хоча деякі методи, як-от інструменти аналізу настроїв, можуть забезпечити швидкий і ефективний спосіб аналізу великих обсягів тексту, перевірка людиною все ще є важливою для точної інтерпретації тональності тексту.

1.4 Використання аналізу тональності тексту

Аналіз тональності тексту має широкий спектр застосувань у різних сферах. У дослідженні ринку його можна використовувати для аналізу відгуків клієнтів і оглядів, щоб отримати уявлення про їх ставлення та думки щодо продукту чи послуги. Потім компанії можуть використовувати цю інформацію для визначення областей для вдосконалення та розробки стратегій для кращого задоволення потреб і вподобань клієнтів. У бренд-менеджменті аналіз тональності тексту можна використовувати для моніторингу сприйняття бренду

суспільством. Аналізуючи розмови в соціальних мережах і огляди в Інтернеті, компанії можуть визначити сфери, де їх бренд сприймається позитивно чи негативно, і вжити заходів для покращення своєї репутації. Аналіз тональності тексту можна використовувати в політичному аналізі, щоб зрозуміти настрої та думку громадськості щодо політичних кандидатів або проблем. Політичні аналітики можуть аналізувати бесіди в соціальних мережах і новинні статті, щоб визначити тенденції та зміни в громадській думці, і використовувати цю інформацію для розробки стратегій перемоги на виборах або впливу на політику. В обслуговуванні клієнтів аналіз тональності тексту можна використовувати для виявлення та вирішення проблем і скарг клієнтів. Аналізуючи відгуки клієнтів і відгуки, компанії можуть швидко визначити сфери, де клієнти незадоволені, і вжити заходів для покращення їх досвіду. Для створення вмісту можна використовувати аналіз тональності тексту, щоб переконатися, що тон і настрої вмісту відповідають бажаному повідомленню та аудиторії. Творці контенту можуть аналізувати тональність наявного контенту та відгуки аудиторії, щоб приймати обґрунтовані рішення щодо мови та стилю свого контенту та максимізувати його вплив. Аналіз тональності тексту також може бути використаний у розробці продукту для визначення вподобань клієнтів і проблемних точок, в управлінні кризою для моніторингу негативних настроїв щодо компанії чи організації, в академічних дослідженнях для аналізу настроїв і емоцій, виражених у літературних творах і розмовах у соціальних мережах, у вивчення мови, щоб допомогти учням визначити тон і почуття різних типів тексту, а також у юридичному аналізі визначити почуття та емоції, виражені в юридичних документах і судових стенограмах. Загалом, аналіз тональності тексту є потужним інструментом для розуміння ставлень, емоцій і почуттів, виражених у тексті, і може використовуватися в різних контекстах для інформування щодо прийняття рішень. Аналіз тональності тексту також можна використовувати в аналізі настроїв для визначення емоційного змісту тексту. Аналіз настроїв особливо корисний у моніторингу соціальних медіа та аналізі

відгуків клієнтів, де його можна використовувати, щоб визначити, як клієнти ставляться до продукту, бренду чи послуги. Ще одним з переліків аналізу тональності тексту, можна використовувати в машинному навчанні та обробці природної мови для розробки алгоритмів, які можуть розуміти людську мову та реагувати на неї. Навчаючи моделі машинного навчання на великих наборах тексту, дослідники можуть розробляти системи, які можуть аналізувати, розуміти та реагувати на людську мову з високим ступенем точності. Це має важливі застосування в таких сферах, як чат-боти, віртуальні помічники та автоматизовані системи обслуговування клієнтів. Варто зазначити, що аналіз тональності тексту не позбавлений обмежень. Наприклад, може бути важко точно проаналізувати тональність текстів, які містять сарказм, іронію чи інші форми образної мови. Подібним чином може бути важко точно проаналізувати тональність текстів, написаних мовами зі складною граматичною структурою або в яких використовуються ідіоматичні вирази. Однак із прогресом у машинному навчанні та обробці природної мови ці обмеження дедалі менше викликають занепокоєння, і аналіз тональності тексту все більше стає цінним інструментом як для компаній, дослідників, так і для окремих людей.

1.5 Постановка проблеми

Мета цієї дипломної роботи — надати огляд аналізу тональності тексту та його різноманітних застосувань у різних сферах за допомогою методів штучного інтелекту. Він має на меті пояснити концепцію тональності тексту, різні методи, що використовуються для аналізу тональності, і різні показники, які використовуються для визначення тональності тексту. Крім того, він досліджує переваги та обмеження аналізу тональності тексту та надає приклади того, як його можна використовувати в таких сферах, як дослідження ринку, управління брендом, політичний аналіз, обслуговування клієнтів, створення контенту та розробка продукту. Мета полягає в тому, щоб допомогти читачам зрозуміти, як можна використовувати аналіз тональності тексту, щоб отримати цінну

інформацію про ставлення, емоції та почуття, виражені в тексті, показати цінність отриманої інформації для бізнесу та надихнути читачів досліджувати способи використання цього потужного інструменту для інформування та прийняття рішень і покращення результатів у різних контекстах. Тому, буде проведено дослідження можливостей використання за допомогою методів машинного навчання. Результатами дослідження буде деяка кількість натренованих моделей машинного навчання на основі алгоритмів наївного баєса та опорних векторів. Тренування буде проводитись на даних текстів постів та коментарів із різних соціальних мереж. Результати тренування та оцінки якості моделей буде представлено у порівняльній таблиці. Натреновані моделі буде використано для отримання інформації про суспільні настрої в англомовному сегменті соціальних мереж щодо повномасштабного вторгнення Російської Федерації в Україну 24 лютого 2022 року.

Для того, щоб провести дослідження були використані саме такі джерела та застосунки:

- інтегрована середа розробки VSCode;

- засоби та бібліотеки мови програмування Python для аналізу тональності тексту, формування таблиць та графіків з отриманих висновків;

- засоби та бібліотеки мови програмування Python для роботи із різними текстовими форматами даних, перетворенням тексту;

- засоби та бібліотеки мови програмування Python для роботи із моделями машинного навчання із учителем та моделями обробки природного тексту без попереднього навчання.

Отримані результати дослідження порівняються із перевіркою емоційно забарвлених текстових даних людиною. Для оцінки емоційного забарвлення

відгуків будуть використовуватись наступні категорії: позитивні, нейтральні та негативні.

Висновки до першого розділу

Підсумовуючи, аналіз предметних областей за допомогою штучного інтелекту, зокрема аналіз тональності тексту, стає все більш важливим інструментом як для компаній, дослідників, так і для окремих людей. Завдяки використанню машинного навчання та обробки природної мови аналіз тональності тексту може надати цінну інформацію про ставлення, емоції та настрої, виражені в тексті, дозволяючи компаніям краще розуміти своїх клієнтів, дослідникам отримати нове розуміння людської поведінки та емоцій, і окремих людей для кращого спілкування та зв'язку з іншими. Хоча існують обмеження щодо точності аналізу тональності тексту, переваги, які він надає, значні та, ймовірно, продовжуватимуть зростати з розвитком технології ШІ. Загалом, використання штучного інтелекту в аналізі предметних областей є захоплюючою подією, яка може змінити спосіб розуміння навколишнього світу та оскільки технології штучного інтелекту продовжують розвиватися, ми можемо очікувати появи ще більш складних інструментів для аналізу предметної області. Наприклад, можуть з'явитися нові методи аналізу тексту, які краще вловлюють відтінки сарказму, іронії та інших форм образної мови. Так само ми можемо очікувати нових застосувань для аналізу тональності тексту в таких сферах, як освіта, охорона здоров'я та державна політика. Наприклад, аналіз тональності тексту можна використовувати для аналізу настроїв онлайн-розмов, пов'язаних із проблемами громадського здоров'я, допомагаючи посадовим особам у сфері охорони здоров'я краще зрозуміти громадське ставлення та занепокоєння. Однак, як і з будь-якою іншою технологією, також існують занепокоєння щодо етичних і соціальних наслідків аналізу предметної області за допомогою ШІ. Наприклад, існує занепокоєння щодо потенціалу штучного інтелекту для посилення існуючих упереджень і нерівності, особливо якщо дані, що

аналізуються, є упередженими або неповними. Також є занепокоєння щодо конфіденційності та захисту даних, особливо якщо особисті дані збираються та аналізуються без згоди окремих осіб. Загалом очевидно, що використання ШІ в аналізі предметної області має як величезний потенціал, так і значні проблеми. Оскільки ми продовжуємо розробляти та вдосконалювати ці інструменти, важливо, щоб ми робили це відповідально та етично, забезпечуючи справедливий розподіл переваг цих технологій і мінімізуючи їх потенційні ризики взаємодії з ним.

РОЗДІЛ 2. Аналіз методів дослідження

2.1 Алгоритм проведення аналізу тональності тексту

Процес аналізу тональності тексту можна розділити на декілька послідовних кроків. Першим кроком є збір та попередня обробка тексту, що буде аналізуватись. Під попередньою обробкою розуміють набір наступних процесів:

- Прибирання з тексту слів, що не будуть мати смислового навантаження, часто повторюваних слів (наприклад, «як», «і», «привіт» тощо);
- Прибирання з тексту знаків пунктуації;
- Розкриття та розшифровка абревіатур, скорочень слів, переписування їх у повній формі;
- Іноді до попередньої обробки тексту відносять також процес заміни слів з їх форму у тексті на їх корінь.

Оброблений текст матиме набагато менше слів та лексем, що могли б погіршити якість аналізу, легший для аналізу і навіть просто менший за вхідний текст за кількістю слів та символів, а отже аналізуватиметься швидше.

Наступним кроком є вилучення ознак або рис із обробленого тексту, на основі яких надалі буде проводитись негативна та позитивна полярна класифікація, за

результатами якої проводитиметься визначення думки автора тексту щодо досліджуваного предмету або явища. Деякими прикладами таких ознак є:

- Послідовності повторюваних змістовних слів;
- N-грами: послідовності з N елементів з вибірки тексту, пов'язаних між собою конкретним змістовним словом або набором слів.
- Групи слів, особливо дієслів та іменників, які допоможуть при класифікації тексту або суб'єктивного або об'єктивного. Також важливими є слова, що індикують початок/закінчення прямої мови або цитати;
- Слова, групи слів та фрази, що виражають емоції та думки, емоційно забарвлені ідіоми та сталі вирази;
- Місцезнаходження слів, що позначають досліджуваний об'єкт або явище у тексті;
- Слова, групи слів та фрази, що використовуються для позначення заперечення. Ця ознака є важливою для дослідження, оскільки полярно змінює емоцію, виражену у тексті, що йде після заперечення;

Методи вибору ознак поділяються на дві великі загальні групи: статистичні методи та методи на основі лексики. Методи на основі лексики не є повністю автоматизованими і потребують перевірки людиною вручну та ручного внесення правок. Статистичні методи є повністю автоматизованими, тому використовуються частіше, ніж методи на основі лексики.

Одними з найбільш уживаних статистичних методів вибору ознак є:

- Використання поточкової взаємної інформації (ПВІ, pointwise mutual information, PMI). ПВІ – особлива міра, що з'явилося у теорії інформації, і означає міру взаємної інформації між ознаками та класами ознак. ПВІ використовується для оцінки зв'язку між одночасно виникаючими подіями (однаковими словами,

конструкціями, наборами слів) у тексті. Формула ПВІ для визначення зв'язку між двома подіями А та В записується у наступному вигляді:

$$PMI(A;B) = log_2 \frac{p(B|A)}{p(B)} = log_2 \frac{p(A,B)}{p(A) * p(B)}$$

Чисельник дробу відображає ймовірність одночасної появи подій А та В, знаменник дробу – очікувані індивідуальні ймовірності появи подій А та В за припущенням їх взаємної незалежності. Додатня ПВІ індикує що поява однієї з подій підвищує ймовірність появи іншою події із пари аналізованих. Негативна ПВІ – навпаки. Якщо значення ПВІ дорівнює 0, це вказує на взаємну незалежність досліджуваних подій.

- Тест хі-квадрат. Формула для хі-квадрата записується як

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

У формулі: O_i - кількість слів у документі, відгуку тощо, що підпадають під заздалегідь визначену категорію для оцінки емоційного забарвлення слів/повідомлень; E_i – очікувана кількість слів у документі, відгуку тощо, що підпадають під заздалегідь визначену категорію для оцінки емоційного забарвлення слів/повідомлень за припущенням їх незалежності. Сумування відбувається по всім заздалегідь обраним категоріям. Тест хі-квадрат вважається кращим за ПВІ, оскільки значення χ^2 є нормалізованим, на відміну від значення ПВІ.

- Латентно-семантична індексація або латенто-семантичний аналіз (ЛСІ або ЛСА) – метод, основна ідея якого полягає у аналізі зв'язку між документами/відгуками та термінами, які у них зустрічаються. Метод будується на припущенні, що близькі за значенням, синонімічні слова будуть зустрічатися у подібних один до одного фрагментах тексту. Першим кроком ЛСІ є створення терм-документної матриці – матриці, в якій у кожний рядок записується унікальне слово з усіх документів, а у кожен стовпець – документ. Таким чином аналізується частота появи кожного унікального слова у документі. Наступним кроком

проводиться процес сингулярного розкладу матриці і матриця, створена на першому кроці, розкладається на 3 матриці: матриця слів, діагональна матриця терм-документної матриці, з якої далі формується матриця кількості появу кожного зі слів у документі, та матриця документів. Далі проводять зменшення розмірності матриць, щоб залишити лише найважливішу приховану семантичну інформацію, таким чином зменшуючи шум – кількість маловпливових та малозмістовних слів – та фіксуючи основні семантичні моделі документів. Далі вираховується семантична схожість слів за допомогою обчислення косинуса між двома векторами, де чим число ближче до 1, тим більше слова та документи схожі одне на одного.

Існує багато інших статистичних методів вибору ознак, таких як, наприклад, прихована марківська модель та прихований розподіл Діріхлє.

Визначення іронії є надзвичайно складним завданням у витягуванні ознак. Рейєс і Россо запропонували роботу, мета якої полягала у тому, щоб знайти відгуки, які містять іронію. Вони намагалися визначити модель ознак для представлення частини суб'єктивних знань, які лежать в основі таких відгуків, і намагалися описати важливі характеристики іронії. Вони створили модель, яка використовує шість категорій ознак, щоб показати вербальну іронію: n-грами, POS-грами, профілювання смішного, профілювання позитивного або негативного, профілювання емоцій і профілювання приємності. Вони створили безкоштовний набір даних, який містить іронічні, сатиричні та новинні відгуки, зібрані з сайту amazon.com.

2.2 Методи класифікації настрою

Наступним і основним кроком аналізу тональності тексту є класифікація настроїв.

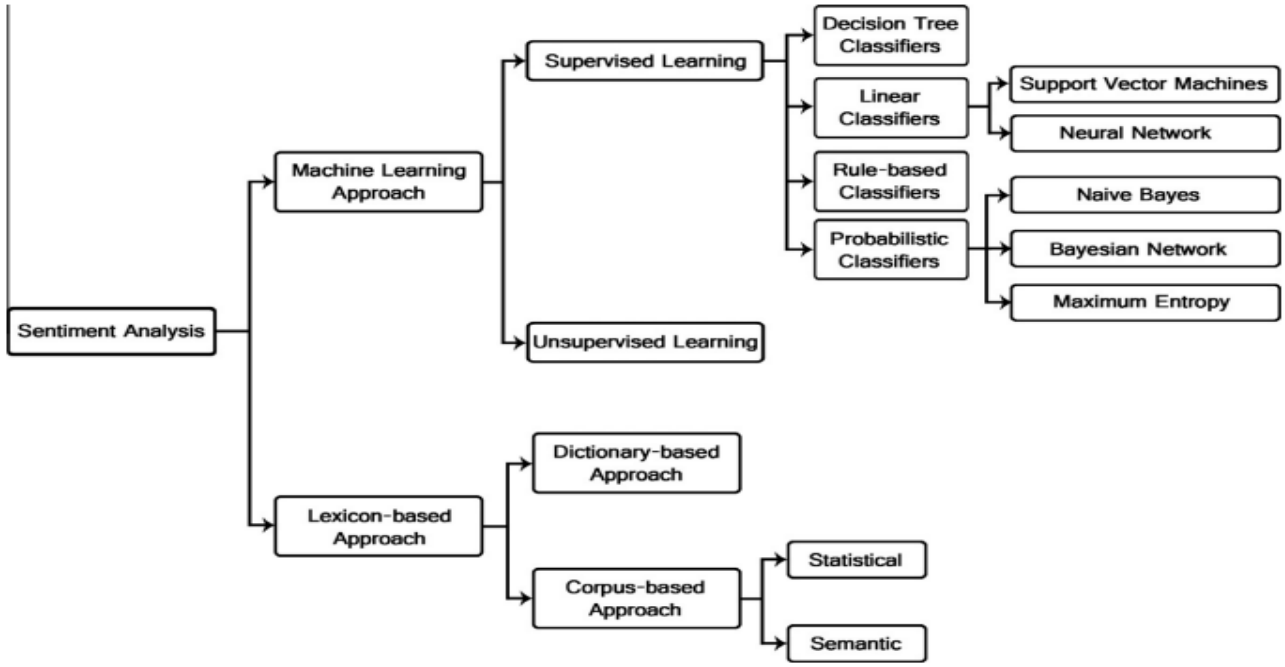


Рис. 2.2.1 Методи класифікації настрою

Три основні категорії методів класифікації настрою включають методи машинного навчання, методи засновані на лексиконі та гібридні підходи. Метод машинного навчання (ML) використовують лінгвістичні характеристики та відомі методи машинного навчання. Методи, засновані на лексиконі базуються на наявності словника заздалегідь скомпільованих і відомих термінів настрою. Вони поділяються на словниковий метод і корпусний підхід, які використовують статистичні або семантичні методи для визначення настрою. Гібридний підхід, що поєднує ці два підходи, є також досить популярним, комбінуючи частини найкращих підходів із двох описаних вище наборів методів.

Для класифікації тексту за допомогою методів машинного навчання можуть використовуватися методи навчання без учителя та методи навчання із вчителем. У методах навчання із вчителем використовується велика кількість

заздалегідь сформованих тренувальних документів. У випадках, коли важко знайти ці тренувальні документи, застосовуються методи навчання без учителя.

Методи засновані на лексиконі ставлять перед собою задачу знайти слова, мовні конструкції та набори слів, в яких виражалась би думка автора. Такі методи використовують два різні підходи. Словниковий метод ґрунтується на отриманні "початкових слів відгуків" і подальшому пошуку синонімів і антонімів цих слів у словнику. Корпусний підхід, який має список початкових слів відгуків, шукає додаткові слова відгуків у великому корпусі, що допомагають виявити слова з відповідним контекстом. Для цього можуть застосовувати різні семантичні та статистичні методи.

Для класифікаторів лінійного типу зробимо наступні припущення: нехай вектор X – вектор нормалізованої частоти слів (кількість появ конкретного слова у тексті поділена на загальну кількість слів у тексті), A – вектор лінійних коефіцієнтів тієї ж розмірності, що і простір ознак слів, b – скаляр. Вихідним результатом лінійного класифікатора є наступний вираз

$$c = A * X + b$$

c є поділяючою гіперплощиною між різними класами.

Класифікатор на основі правил моделює простір даних із набором даних. Даний класифікатор можна уявити як диз’юнктивну нормальну форму. З правого боку матимемо набір класів, з лівого – умови на набір ознак слів. Умовою є наявність або відсутність слова, набору слів або мовної конструкції, але найчастіше використовують саме умову на наявність перерахованих вище об’єктів.

Класифікатори на основі лексики поділяються на класифікатор на основі словників та на класифікатор на основі корпусів. Ці класифікатори є автоматизованими, тобто не потребують втручання людини під час своєї

роботи, але вони вимагають попереднього збору емоційно забарвлених слів та наборів слів людиною вручну.

2.3 Опис обраних алгоритмів

Наївний баєсів класифікатор (НБ) є ймовірнісним класифікатором, що в своїй основі використовує теорему Баєса для визначення ймовірності приналежності досліджуваного спостереження – в випадку аналізу тональності тексту слова, мовної конструкції або групи слів – до одного із заздалегідь обраних класів (лейблів), враховуючи значення ознак спостережень, при наївному припущені незалежності ознак одна від одної, тобто вважаючи, що кожна ознака внесе власний незалежний внесок до вирішення задачі. Основними перевагами цього класифікатора є простота реалізації та велика швидкість обробки великого обсягу даних. Недоліком класифікатора є те, що його результати будуть адекватними лише в тому випадку, якщо припущення про незалежність ознак спостережень буде виконуватись. Але на практиці було з'ясовано, що навіть при наявності слабкої, або навіть істотної залежності між ознаками, наївний баєсів класифікатор зможе видати результат, що корелюватиме за справжнім розподіл слів та ознак за класами. Формально даний класифікатор записується як

$$P(label|features) = \frac{P(label) * P(features|label)}{P(features)}$$

Нехай маємо вектор точок даних $x = (x_1, \dots, x_n)$ та n ознак. Ймовірність точки належати класу y_kобчислюється як

$$P(y_k|x) = \frac{P(y_k) * P(x|y_k)}{P(x)} = \frac{P(y_k) * P(x_1, \dots, x_n|y_k)}{P(x_1, \dots, x_n)}, \qquad k = \overline{1, K}$$

За припущенням про умовну попарну незалежність між ознаками у векторі даних з ознаками, отримаємо

$$P(x_i|x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n|y_k) = P(x_i|y_k) =>$$

$$P(x_1, \dots, x_n|y_k) = \prod_{i=1}^{n} P(x_i|y_k) =>$$

$$P(y_k|x) = \frac{P(y_k) * \prod_{i=1}^{n} P(x_i|y_k)}{P(x_1, \dots, x_n)} =>$$

$$P(y_k|x_1, \dots, x_n) \propto P(y_k, x_1, \dots, x_n) \propto P(y_k)P(x_1, \dots, x_n|y_k)$$

$$\propto P(y_k)P(x_1|y_k) \dots P(x_n|y_k) \propto P(y_k) \prod_{i=1}^{n} P(x_i|y_k)$$

НБ виводить ймовірність точки x належати класу y_k як пропорційну до добутку апріорної ймовірності $P(y_k)$ на добуток умовних ймовірностей $\prod_{i=1}^{n} P(x_i|y_k)$. Формула вибору класу для точки даних виглядає як

$$\hat{C} = argmax_{k \in \{1, \dots K\}} P(y_k) \prod_{i=1}^{n} P(x_i|y_k)$$

Багатовидовий НБ є одним із класичних видів наївного баєсового класифікатора і застосовується до текстових даних, перетворених на вектори кількості появ слів у документах або реченнях. При цьому Багатовидовий НБ не накладає на досліджувані дані обмежень щодо їх розподілу, на відміну від Гаусового НБ, для коректної роботи якого необхідно, щоб частота появи слів у документах або реченнях мала нормальний розподіл.

Розподіл частоти появи слів у документах параметризується векторами $\theta_y = (\theta_{y1}, \dots, \theta_{yn})$, де y відповідає класам емоційного забарвлення, а n – кількість ознак токенів. θ_{yi} - $P(x_i|y)$, тобто ймовірність появу ознаки i у частині тексту, що відноситься до класу y. Параметр θ_y оцінюється за допомогою згладженої версії методу максимальної правдоподобності

$$\hat{\theta}_{yi} = \frac{N_{yi} + \alpha}{N_y + \alpha n}$$

$$N_{yi} = \sum_{x \in T} x_i$$

$$N_y = \sum_{i=1}^{n} N_{yi}$$

α – згладжувальний параметр.

Згладжувальний параметр необхідний для того, щоб запобігти появі ймовірностей, що дорівнювали б нулю. У Багатовидовому НБ може виникнути ситуація під час оцінки ймовірностей різних ознак для деякого класу, що у тренувальному наборі даних не зустрічалася певна ознака для цього класу. Тоді ймовірність появу цієї ознаки для досліджуваного класу буде рівна нулю. Згладжувальний параметр надають мале позитивне значення для кількості появи кожної ознаки, що не з’являлась у тренувальному наборі даних для певного класу.

Комплементарний НБ є розширенням Багатовидового НБ на випадок незбалансованих розподілів класів: такий випадок, коли кількість прикладів на один із тегів емоційного забарвлення у тренувальному набору даних сильно переважає над кількістю прикладів для інших тегів. Комплементарний НБ використовує поняття комплементу ознаки – спеціального признаку, що вказує на відсутність ознаки (яка може бути, наприклад, наявністю або відсутністю деякого слова у токені, характеристикою слова), а не на її наявність, як це відбувається у звичайному наївному баєсі. Комплемент ознаки допомагає вирішити проблему незбалансованих розподілів класів, фокусуючись саме на відсутності ознаки, враховуючи, що присутність ознаки може бути менш інформативною для класів, що мають невелике представлення у наборі даних, для подій, що відбуваються рідше. Комплемент ознаки є способом врахування незбалансованості класів, де більшість даних

належить одному класу, а меншість – іншим. Використання комплементу ознаки допомагає збалансувати вплив ознак з різних класів, надаючи більшу вагу більш рідкісним ознакам, що є характерними для меншості класів.

Формула для розрахунку ваг Комплементарного НБ має наступний вигляд

$$\hat{\theta}_{ci} = \frac{\alpha_i + \sum_{j:y_j \neq c} d_{ij}}{\alpha + \sum_{j:y_j \neq c} \sum_k d_{kj}}$$

$$w_{ci} = log\hat{\theta}_{ci}$$

$$w_{ci} = \frac{w_{ci}}{\sum_j |w_{cj}|}$$

j – документи, речення, що знаходяться у тренувальному наборі даних

c – класи, по яких проводиться класифікація

d_{ij} – частота, з якою токен i зустрічаються у документах, реченнях j

α_i – згладжуючий гіперпараметр

Класифікаційне правило виглядає як

$$\hat{c} = \arg min_c \sum_i t_i w_{ci}$$

Тобто точка даних отримує той клас, для якого комплемент буде найгіршим.

Класифікатор на основі опорних векторів (support vector machines – SVM) є одним з найбільш вживаних алгоритмів у задачі аналізу тональності тексту. Основною метою цього алгоритму є пошук такої гіперплощини, що максимізує відстань між двома найближчими до неї класами, а точніше розділяє ознаки, що належать цим класам. Для лінійно нероздільних даних даний алгоритм створює нелінійний простір виборів в заданій площині ознак, застосовуючи нелінійні функції переводу даних у нову площину, в якій вони можуть бути лінійно

розділені. Лінійні алгоритми використовуються як у задачах класифікації, так і у задачах лінійної регресії.

Основне застосування SVM – вирішення задачі бінарної класифікації. Вхідні дані представляються у якості вектору ознак $x = (x_1, \dots, x_n)$. Задача класифікатора – знайти таку гіперплощину, яка б розділяла точки вектору ознак за двома класами, тобто знайти таке рівняння $w \cdot x + b = 0$, що б розділяло точки двох класів найоптимальнішим способом. Таку гіперплощину ще називають роздільною або гіперплощиною прийняття рішень. Вектор $w = (w_1, \dots, w_n)$ є вектором нормалі до роздільної гіперплощини. Оптимальним вважається таке розділення, при якому відстань між двома найближчими одна до одної точками двох різних класів була б максимальною. Після проведеного навчання алгоритму необхідно вивести деяку функцію $F(x) = sign(w \cdot x + b)$, y – мітка одного з класів, цільове значення, $y_k \in \{-1, 1\}$.

У контексті SVM вводиться декілька важливих понять:

- Опорні вектори – ознаки, що знаходяться найближче до роздільної гіперплощини. Головна задача SVM – максимізувати відстань між опорними векторами;
- Відступ – відстань між опорними векторами двох класів. Обчислюється як перпендикуляр від опорного вектору класу до гіперплощини. Саме відступ необхідно максимізувати.

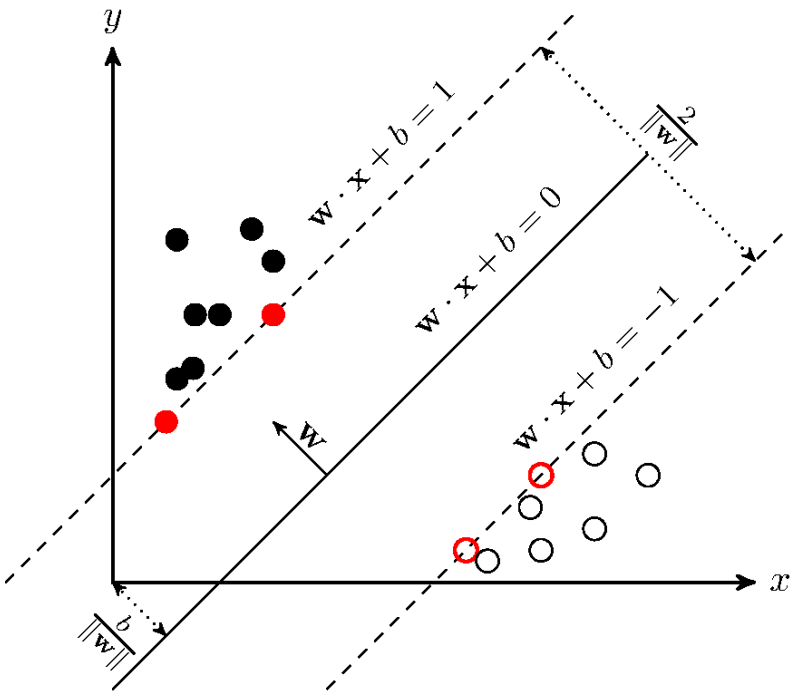


Рисунок 2.3.5 Ілюстрація алгоритму SVM на двовимірній площині

Вагами, зміна значень яких призведе до зміни результатів роботи класифікатора, тут виступають вектор w та значення b. Знайдемо проекцію вектору різниці між опорними векторами на вектор w. Таким чином можна буде вивести формулу для обчислення відстані між опорними векторами. За формулою проекції вектору матимемо:

$$\frac{(w \cdot x_+ - w \cdot x_-)}{||w||} = \frac{(1-b)+(1+b)}{||w||} = \frac{2}{||w||}$$

Формула для обчислення відступу записується як $M = y_k(w * x_k + b)$. Якщо $M \in (0,1)$, то досліджувана точка потрапляє роздільної площини, позначеної пунктиром на рисунку 2.3.5. Тоді необхідно накласти наступну умову: $y_k(w \cdot x_k + b) \geq 1$ (1). Таким чином усі досліджувані точки будуть потрапляти у один із класів, а не до роздільної площини. Цю умову називають канонічною репрезентацією роздільної гіперплощини. Таким чином, проблема пошуку роздільної гіперплощини зводиться до задачі оптимізації:

$$\begin{cases} \frac{1}{2}\left|\left|w\right|\right|^2 \to min \\ y_k(w \cdot x_k + b) \geq 1 \end{cases}$$

За теоремою Куна-Таккера ця задача оптимізації є еквівалентною до двоїстої задачі пошуку сідлової точки функції Лагранжа. Щоб мати змогу розв'язати цю задачу аналітично, вводяться множники Лагранжа $a_n \geq 0$ та вектор $a = (a_1, \dots, a_n)$. Вводиться також функція Лагранжа:

$$L(w, b, a) = \frac{1}{2}||w||^2 - \sum_{k=1}^{n} a_k\{y_k(w \cdot x_k + b) - 1\}$$

Покладемо похідні функції Лагранжа по змінних w та b рівними нулю. Тоді

$$w = \sum_{k=1}^{n} a_k \cdot y_k \cdot x_k$$

$$0 = \sum_{k=1}^{n} a_k \cdot y_k$$

Описану вище задачу можна звести до еквівалентної задачі квадратичного програмування, що не буде містити змінних w та b

$$\begin{cases} \tilde{L}(a) = \sum_{k=1}^{n} a_k - \frac{1}{2}\sum_{k=1}^{n}\sum_{m=1}^{n} a_k a_m y_k y_m k(x_k, x_m), \qquad k(x, x') = (x \cdot x') \\ a_k \geq 0 \\ \sum_{k=1}^{n} a_k \cdot y_k = 0 \end{cases}$$

Маючи задачу у такому вигляді, можемо переписати задачу виділення класів як

$$F(x) = \sum_{k=1}^{n} a_k y_k k(x, x_k) + b$$

$$\begin{cases} a_k \geq 0 \\ y_k F(x_k) - 1 \geq 0 \\ a_k\{y_k F(x_k) - 1\} \geq 0 \end{cases}$$

Із умов бачимо, що або $a_k = 0$, або $y_k F(x_k) = 1$. Будь-яка точка, для якої $a_k = 0$ не буде включатись у задачу. Таким чином ми зменшуємо час та обчислювальну складність задачі, виключаючи деякі точки з розгляду.

Введемо також функцію помилки

$$\sum_{k=1}^{n} E_{inf}(y_k F(x_k) - 1) + \lambda||w||^2,\ E_{\text{inf}}(z) = 0\ if\ z \geq 0, else\ E_{\text{inf}}(z) = inf$$

Описаний алгоритм називають алгоритмом із жорстким відступом.

Описаний вище алгоритм працює для випадку, коли точки можуть бути розділені лінійно. Але іноді стається так, що точки вхідних даних можуть перекривати одне одного, тобто дані будуть лінійно нероздільними, і у такому випадку описаний алгоритм буде давати невірний результат. Щоб отримати адекватну класифікацію необхідно дозволити алгоритму іноді робити помилки, а не перераховувати значення $F(x)$ кожного разу, як точку було невірно класифіковано. Необхідно ввести деяке «покарання» для точок, що було класифіковано невірно. Чим далі точки від області свого класу та від роздільної гіперплощини, тим більше значення «покарання». Для цього вводяться також слабкі змінні $\xi_k \geq 0, k = \overline{1,n}$

$$\xi_k = \begin{cases} 0, & \text{точка лежить в області свого класу} \\ |y_k - F(x_k)| \end{cases}$$

Умова (1) у такому випадку зміниться на наступний вираз: $y_k F(x_k) \geq 1 - \xi_k(2)$.

$$\xi_k = \begin{cases} 0, & \text{точка в області свого класу або на границі роздільної області} \\ (0,1], & \text{точка всередині відступу, але на правильній його частині} \\ (1,\infty), & \text{точка на невірній стороні роздільної області} \end{cases}$$

Тепер необхідно мінімізувати функцію за умови (2):

$$C\sum_{k=1}^{n}\xi_k + \frac{1}{2}||w||^2, \qquad C > 0$$

C – параметр, що дозволяє регулювати відношення між максимізацією ширини роздільної області та мінімізацією помилки.

Надалі всі роздуми та перетворення залишаються такими самими, що були у попередньому випадку. В кінці отримаємо наступну систему

$$\begin{cases} \tilde{L}(a) = \sum_{k=1}^{n} a_k - \frac{1}{2}\sum_{k=1}^{n}\sum_{m=1}^{n} a_k a_m y_k y_m k(x_k, x_m) \\ 0 \le a_k \le C \\ \sum_{k=1}^{n} a_k \cdot y_k = 0 \end{cases}$$

Цей алгоритм називають алгоритмом із м’яким відступом.

Функція $k(x, x')$ називається ядровою функцією. Основна задача ядрової функції – обчислити ступінь близькості вхідних точок одна до одної у заданій площині. Ядрові функції, або ядра, використовуються у SVM у контексті ядерного трюку (kernel trick). Ядерний трюк було запропоновано як спосіб вирішення проблеми нелінійної класифікації вхідних точок. Основна ідея ядерного трюку – переведення вхідних даних у площину вищої розмірності без перерахування у ній кожної точки та вектору. Замість цього у ядерному трюці використовують ядерну функцію. Результуючий алгоритм виходить дуже схожим на звичайний лінійний алгоритм, лише кожний скалярний добуток у алгоритмі замінюється на нелінійну функцію ядра – скалярний добуток у площині вищої розмірності. Деякими найпоширенішими ядрами є:

- Однорідне поліноміальне – $k(x, x') = (x \cdot x')^d$
- Неоднорідне поліноміальне – $k(x, x') = (x \cdot x' + 1)^d$

- Гаусова радіально-базисна функція – $k(x, x') = \exp\left(-\gamma\left||x - x'|\right|^2\right), \gamma > 0$
- Сигмоїдне – $k(x, x') = \tanh\left(\delta x \cdot x' + c\right)$ для деяких $\delta > 0,\ c > 0$

SVM є бінарним класифікатором, тобто може створити роздільну гіперплощину у випадку, якщо класів, до яких можуть належати точки вхідних даних, усього два. У задачі аналізу тональності тексту найчастіше виділяють 3 або 5 класів. У випадках небінарної класифікації використовуються два основні підходи: One-vs-One та One-vs-Rest.

One-vs-Rest підхід використовує ідею перетворення задачі полікласифікації на N задач бінарної класифікації, де N – кількість класів у задачі. Кожен клас протиставляється об'єднаному набору з усіх інших N-1 класів, і вже на таких переведених у бінарний вигляд даних відбувається процес бінарної класифікації.

Під час процесу полікласифікації для кожної точки вхідних даних обчислюється ймовірність ймовірність належати одному з N класів. Відповідним точці класом стає той, для якого пов'язана із ним ймовірність для точки буде максимальною.

Основною проблемою цього підходу є необхідність створення та тренування окремої моделі для кожного з N класів. Це може призвести до збільшення часу тренування та збільшення обчислювальної складності, якщо у задачі буде дуже багато класів (наприклад, сто), якщо тренувальний набір даних буде великим (мільйони строк) тощо.

One-vs-One підхід базується на ідеї розбиття набору класів на N(N-1)/2 пар, для яких потім проводиться бінарна класифікація.

SVM частіше використовує One-vs-One підхід. Причинами на те є більш легкі обчислення, ніж у One-vs-Rest підході, оскільки, хоча кількість проведених

алгоритмів бінарної класифікації більша у One-vs-One, кількість точок вхідних даних для кожного з алгоритмів буде меншою. Також One-vs-One підхід проводить бінарну класифікацію на більш збалансованих наборах даних, ніж One-vs-Rest підхід, а отже використання його моделі даватиме більш адекватні результати.

Висновки до другого розділу

Існує дуже велика кількість підходів і методів розв'язання задачі класифікації слів за їх емоційним забарвленням. У кожного із підходів є свої переваги та недоліки, а вибір підходу повинен робитись в залежності від конкретного завдання, наявних обмежень по часу виконання аналізу, пам'яті комп'ютера, наявності або відсутності великої кількості специфічної для деякої галузі лексики, наявності або відсутності іронії та сарказму у тексті, розміру тексту, кількості у ньому заперечних слів, що полярно змінюють тональність слова, набору слів.

Найчастіше жоден класифікатор не застосовується окремо. Стандартним підходом є комбінація декількох класифікаторів задля досягнення найкращого, найбільш точного результату.

Найбільше зараз розвивають методи на основі машинного навчання, як з учителем, так і без вчителя, оскільки вони майже не потребують втручання людини у процес аналізу, а також не потребують багато часу та людського ресурсу для збору даних для тренування.

Після вивчення найчастіше вживаних методів класифікації слів/наборів слів для аналізу тональності тексту було вирішено обрати наївний баєсів класифікатор та класифікатор на основі опорних векторів.

РОЗДІЛ 3. Проведення дослідження

3.1 Інструменти дослідження та дані

Мовою програмування для виконання дослідження було обрано мову Python. Python є інтерпретованою мовою програмування, створена Гвідо ван Россумом у 1990 році. Python набув популярності завдяки легкому для читання та розуміння синтаксису та великою кількістю бібліотек із відкритим кодом. Python надає можливість будь-якій людині створити бібліотеку із будь-яким функціоналом, і будь-якому програмісту використати цю бібліотеку у своїй роботі. Основними, але не єдиними сферами застосування цієї мови програмування є веб-програмування, здебільшого backend, та робота із даними: аналіз даних, робота із реляційними базами даних, інженерія даних, розробка та використання нейронних мереж, ШІ, моделей машинного навчання.

Даними дослідження обрано коментарі під відео на платформі YouTube. Для тренування моделі було обрано два набори даних: перший набір даних містить текст твітів соціальної мережі Twitter із відповідними тегами класифікації емоційного забарвлення; інший набір даних містить текст коментарів найпопулярніших відео в англомовному сегменті платформи YouTube. Дані з другого набору не мають тегів емоційного забарвлення, тому їх буде додано вручну за допомогою використання класифікатора VADER. Для оцінки емоційного забарвлення тексту українською, російською та будь-якими іншими мовами необхідно обирати набори даних для тренування, що містять дані саме цими мовами, або використовувати класифікатори, що не базуються на машинному навчанні, а мають у своїй основі ідею використання корпусу тексту чи словника. Оскільки найпопулярнішою мовою великих соціальних мереж як Twitter, Facebook, YouTube та просто найпопулярнішою мовою для міжнародного спілкування є саме англійська, було обрано набори даних для тренування моделей на опрацювання тексту саме англійською мовою.

Для виконання дослідження будуть застосовані деякі бібліотеки із відкритим кодом мови програмування Python для роботи із візуалізацією даних, машинним навчанням та алгоритмами аналізу тональності тексту, зокрема:

- Pandas – бібліотека, завдяки якій можна переводити дані у формат датафреймів та працювати із ними надалі як з матрицями або таблицями;

- NLTK – потужна бібліотека, що зберігає велику кількість алгоритмів та методів для обробки природної мови та тонального аналізу тексту. У цій бібліотеці також зберігаються словники для використання класифікаторів на основі словників та корпусів та словники слів різними мовами, що не несуть ніякої важливої інформації у тексті;
- Sklearn – бібліотека для роботи із методами машинного навчання. Надає можливості для створення, тренування та перевірки алгоритмів машинного навчання, класифікації, регресії, кластеризації тощо;
- Matplotlib – бібліотека для візуалізації статистичних даних у якості гістограм, графіків тощо.

Оцінка натренованих моделей буде проводитись по 4 критеріях: accuracy, precision, recall, f1-score.

Precision відповідає на питання: яка кількість позитивних передбачень була дійсно вірною? Розглянемо логіку обрахунку precision на прикладі бінарного класифікатору.

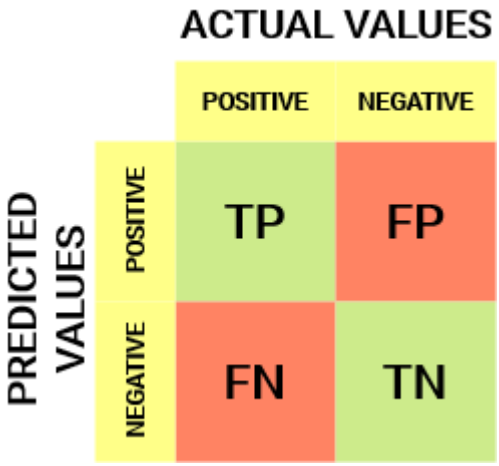


Рисунок 3.1.1 Матриця невідповідностей для бінарного класифікатору

TP – True Positive – значення було передбачене як позитивне, воно і було позитивне

TN – True Negative – значення було передбачене як негативне, воно і було негативне

FP – False Positive – значення є негативним, але було класифіковане як позитивне

FN – False Negative – значення є позитивним, але було передбачене як негативне

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

Accuracy відображає відношення кількості правильних передбачень до загальної кількості зроблених передбачень.

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

Recall відображає, яка частина дійсно позитивних передбачень було вірно ідентифікована.

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

Recall та precision мають обернену залежність, тобто спроби покращити одну з цих метрик призведуть до погіршення іншої.

F1-score – гармонічне середнє між recall та precision. Ця метрика дає загальне уявлення про якість про recall та precision одночасно. Досягає свого максимального значення, коли precision дорівнює recall.

$$F1 - score = \frac{2}{\frac{1}{Recall} + \frac{1}{Precision}}$$

3.2 Опис та результат роботи програми

Першим важливим кроком для роботи із класифікатором на основі машинного навчання було знайти набори даних, на яких цей класифікатор міг би навчатись. Один з обраних наборів даних підходив для навчання одразу, інший же набір необхідно перетворити, додавши у нього теги емоційного забарвлення. Для цього було використано класифікатор на основі лексики та правил, що не потребує попереднього навчання – VADER.

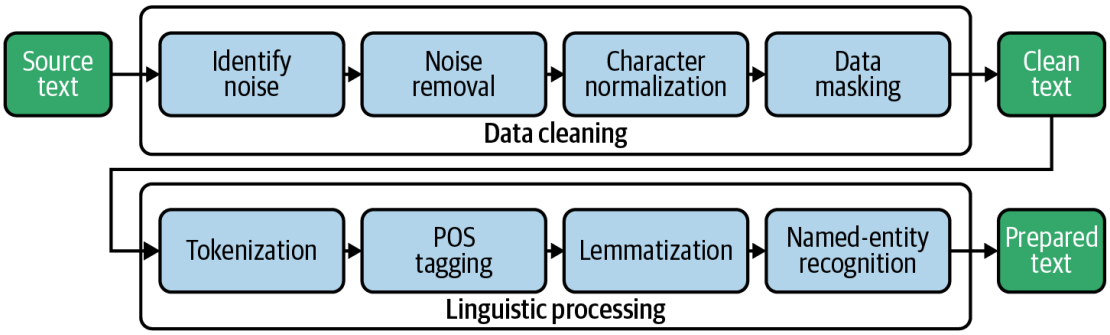
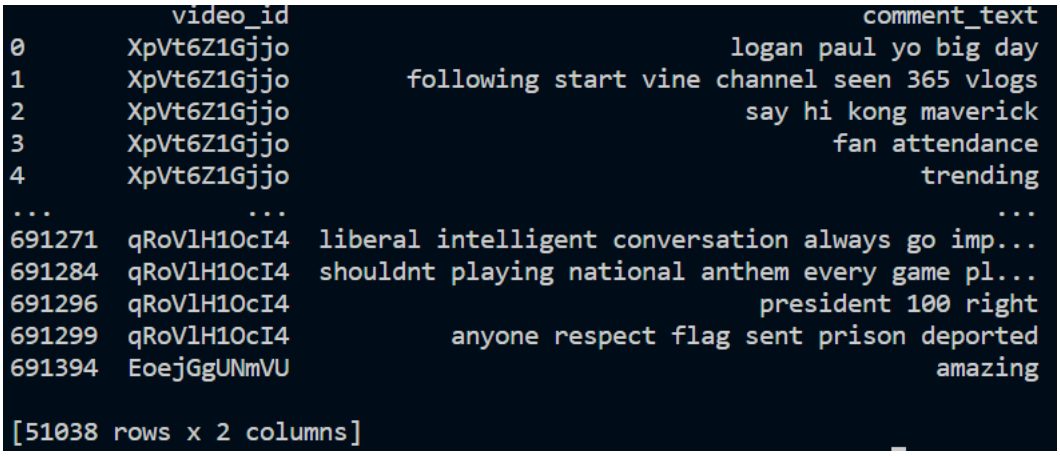


Рисунок 3.2.1 Процес попередньої підготовки тексту

Для попередньої підготовки та очищення тексту застосуємо методи та словники word_tokenize із nltk.tokenize, stopwords із nltk.corpus та WordNetLemmatizer із nltk.stem. Із набору даних з коментарями з платформи YouTube було прибрано усі нетекстові символи (наприклад, емоджі, знаки відсотку тощо). Після цього було проведено токенізацію коментарів і приведено всі слова у коментарях до нижнього регістру. Токенізацією називається процес розбиття тексту на невеликі частини (слова, сполучення слів), об’єднаних за якоюсь ознакою. Процес токенізації є важливим для роботи із обробкою природної мови, оскільки дозволяє працювати із текстом та аналізувати його на більш гранульованому рівні. Наступним кроком було прибрано усі слова, що не несуть змістовного навантаження (такі слова як a, the, and, for). Після цього усі слова було приведено до їх базової форми, тобто слова замінювались на свої корені. Таким чином було зменшено кількість тексту у коментарях, прибрано слова, що не були б корисними і лише заважали б визначенню емоційного забарвлення частини тексту та прибрано зайві символи.



	video_id	comment_text
0	XpVt6Z1Gjjo	logan paul yo big day
1	XpVt6Z1Gjjo	following start vine channel seen 365 vlogs
2	XpVt6Z1Gjjo	say hi kong maverick
3	XpVt6Z1Gjjo	fan attendance
4	XpVt6Z1Gjjo	trending
...	...	...
691271	qRoVlH1OcI4	liberal intelligent conversation always go imp...
691284	qRoVlH1OcI4	shouldnt playing national anthem every game pl...
691296	qRoVlH1OcI4	president 100 right
691299	qRoVlH1OcI4	anyone respect flag sent prison deported
691394	EoejGgUNmVU	amazing

[51038 rows x 2 columns]

Рисунок 3.2.2 Результат попередньої обробки набору даних з коментарями з платформи YouTube

Наступним кроком підготовки даних для тренування моделі машинного навчання було виділення емоційного забарвлення коментарів за допомогою SentimentIntensityAnalyzer модуля nltk, що використовує технологію VADER. VADER(Valence Aware Dictionary and sEntiment Reasoner) є класифікатором на основі лексикону та правил, що не потребує попереднього навчання. Він також є одним з найбільш уживаних класифікаторів для роботи із текстами, отриманими з соціальних мереж. Основною роботи VADER є заздалегідь створений словник, що містить більше ніж 7000 слів та фраз із відповідними тегами емоційного забарвлення, розділених на п’ять категорій: негативні, дещо негативні, нейтральні, дещо позитивні, позитивні. Перевагами VADER є простота використання, легкість у класифікації неформальної мови, що зазвичай використовується користувачами під час створення текстів у соціальних мережах, швидкість роботи. Недоліками цього класифікатору є обмеженість у можливостях роботи із іронією та сарказмом, та можлива генерація неякісних результатів під час аналізу великої кількості тексту, оскільки VADER показує найкращі результати на класифікації емоційного забарвлення слів або невеликих сукупностей слів, але показує гірші результати під час аналізу емоційного забарвлення тексту загалом.

VADER повертає словник із 4 дробових значень в діапазоні від -1 до 1:

- Compound: загальна оцінка емоційного забарвлення
- Neg: оцінка ступеня негативності висловлювання
- Pos: оцінка ступеня позитивності висловлювання
- Neu: оцінка ступеня нейтральності висловлювання

Для проведення класифікації було використано значення compound. Числові значення було переведено у текстові наступним чином: ‘positive’ якщо compound > 0, ‘negative’ якщо compound < 0, ‘neutral’ якщо compound = 0.

	video_id	comment_text	scores
0	XpVt6Z1Gjjo	logan paul yo big day	neutral
1	XpVt6Z1Gjjo	following start vine channel seen 365 vlogs	neutral
2	XpVt6Z1Gjjo	say hi kong maverick	neutral
3	XpVt6Z1Gjjo	fan attendance	positive
4	XpVt6Z1Gjjo	trending	neutral
...	...	...	...
691271	qRoVlH1OcI4	liberal intelligent conversation always go imp...	negative
691284	qRoVlH1OcI4	shouldnt playing national anthem every game pl...	positive
691296	qRoVlH1OcI4	president 100 right	neutral
691299	qRoVlH1OcI4	anyone respect flag sent prison deported	negative
691394	EoejGgUNmVU	amazing	positive

Рисунок 3.2.3 Результат роботи VADER на наборі даних з коментарів з платформи YouTube

Надалі необхідно було створити та натренувати модель машинного навчання. Для цього необхідно було перетворити існуючий набір тренувального тексту на спеціальну матрицю ознак TF-IDF. TF-IDF є статистичним показником, що відображає вагу або важливість кожного слова у документі. Вага слова визначається як кількість вживань цього слова у документі/реченні, обернено пропорційна до частоти вживання словах у інших документах/реченнях. Формула TF-IDF складається з двох частин: TF та IDF. TF – term frequency – відображає відношення частоти появи обраного слова до загальної кількості слів у документі. Формула для обчислення TF має наступний вигляд:

$$TF = \frac{n_i}{\sum_k n_k}$$

В чисельнику дробу стоїть кількість входжень слова у документ, у знаменнику – загальна кількість різних слів у документі.

IDF – inverse document frequency – інверсія значення частоти, із яким обране слово зустрічається у документах. Формула для обчислення IDF виглядає як:

$$IDF = log \frac{|D|}{|(d_i \supset t_i)|}$$

У чисельнику дробу стоїть загальна кількість документів у наборі досліджуваних даних. У знаменнику дробу – кількість документів, у яких зустрічається слово t_i. Фінальною формулою є $TF - IDF = TF * IDF$.

Для виконання завдання було обрано наївний баєсів класифікатор. НБ розподіляється на декілька видів: Гаусів, Багатовидовий, Комплементарний, Бернулі та Категоріальний. Класичним вважається Гаусів НБ, для проведення роботи було обрано Гаусів, Багатовидовий та Комплементарний. В ході проведення роботи було виявлено, що найкраще на обраних наборах даних для тренування та перевірки показують себе саме Комплементарний та Багатовидовий НБ.

```
Multinomial Naive Bayes
Accuracy: 0.744515843119876
Precision: 0.7900028731156358
Recall: 0.744515843119876
F1-score: 0.7378328701355829
**********
Complement Naive Bayes
Accuracy: 0.7869488145357855
Precision: 0.7944064614525789
Recall: 0.7869488145357855
F1-score: 0.7848205735972291
**********
```

Рисунок 3.2.4 Результат тренування та тестування натренованої моделі Багатовидового наївного баєса та Комплементарного наївного баєса

Другим алгоритмом було обрано алгоритм SVM. Дослідження проводилося на 4 видах ядерних функцій: лінійних, поліноміальних, сигмоїдних та на радіальній функції Гауса.

Результат тренування та оцінки результатів тренування представлено у Таблиці 3.2.1

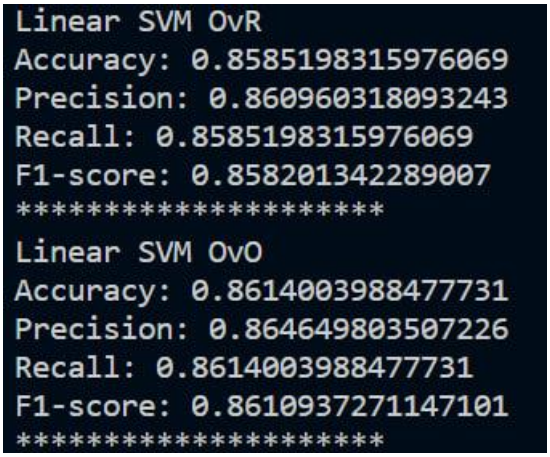


Рисунок 3.2.5 Результат тренування та тестування натренованої моделі SVM із лінійним ядром

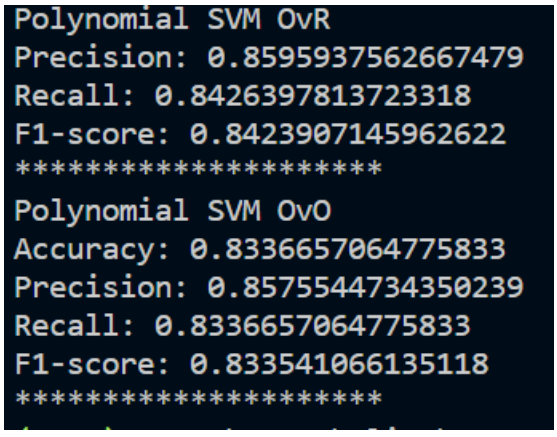


Рисунок 3.2.6 Результат тренування та тестування натренованої моделі SVM із поліноміальним ядром

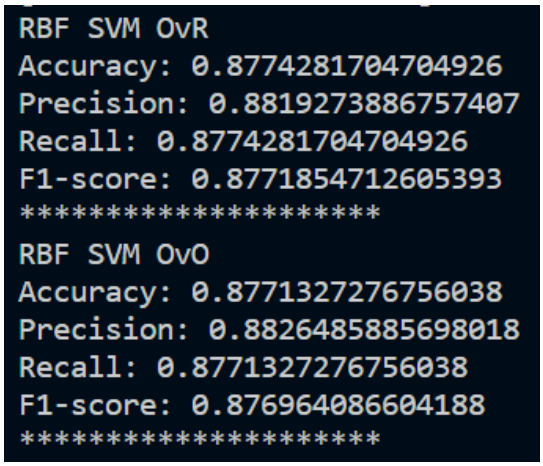


Рисунок 3.2.7 Результат тренування та тестування натренованої моделі SVM із ядром на основі радіально-базисної функції

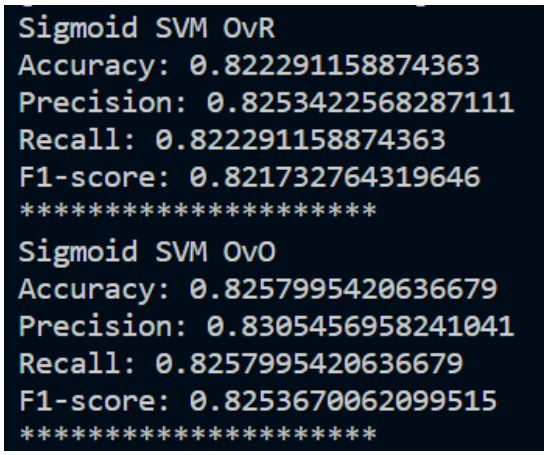


Рисунок 3.2.7 Результат тренування та тестування натренованої моделі SVM із сигмоїдною ядерною функцією

Алгоритм	Варіація алгоритму	Accuracy	Precision	Recall	F1-score
Наївний Баєс	Багатовидовий НБ	0.7445	0.7900	0.7445	0.7378
	Комплементарний НБ	0.7869	0.7944	0.7869	0.7848
One-vs-Rest SVM	Лінійна ядерна функція	0.8585	0.8609	0.8585	0.8582

	Поліноміальна ядерна функція	0.8426	0.8595	0.8426	0.8423
	Гаусова радіально-базисна функція	0.8774	0.8819	0.8774	0.8771
	Сигмоїдна ядерна функція	0.8222	0.8253	0.8222	0.8217
One-vs-One SVM	Лінійна ядерна функція	0.8614	0.8646	0.8614	0.8611
	Поліноміальна ядерна функція	0.8336	0.8575	0.8336	0.8335
	Гаусова радіально-базисна функція	0.8771	0.8826	0.8771	0.8769
	Сигмоїдна ядерна функція	0.8257	0.8305	0.8257	0.8253

Таблиця 3.2.1 Порівняння результатів тренування та тестування різних модифікацій обраних алгоритмів

Із Таблиці 3.2.1 бачимо, що найкращі результати було отримано після тренування моделі на основі опорних векторів із Гаусовою радіально-базисною функцією пі підходом One-vs-Rest. Отже саме цю модель було використано для дослідження даних щодо суспільних настроїв щодо повномасштабного вторгнення.

Дані для дослідження було зібрано за допомогою Google YouTube API. Для дослідження було обрано два відео: «Russia-Ukraine crisis: Putin order military operation in Ukraine», «”We will defend ourselves”, says Ukrainian president Volodymyr Zelensky». Було зібрано набір із 100 найбільш популярних коментарів під кожним відео і збережено у списку. Після проведення

класифікації результат роботи класифікатора було виведено у формі зведеного графіку

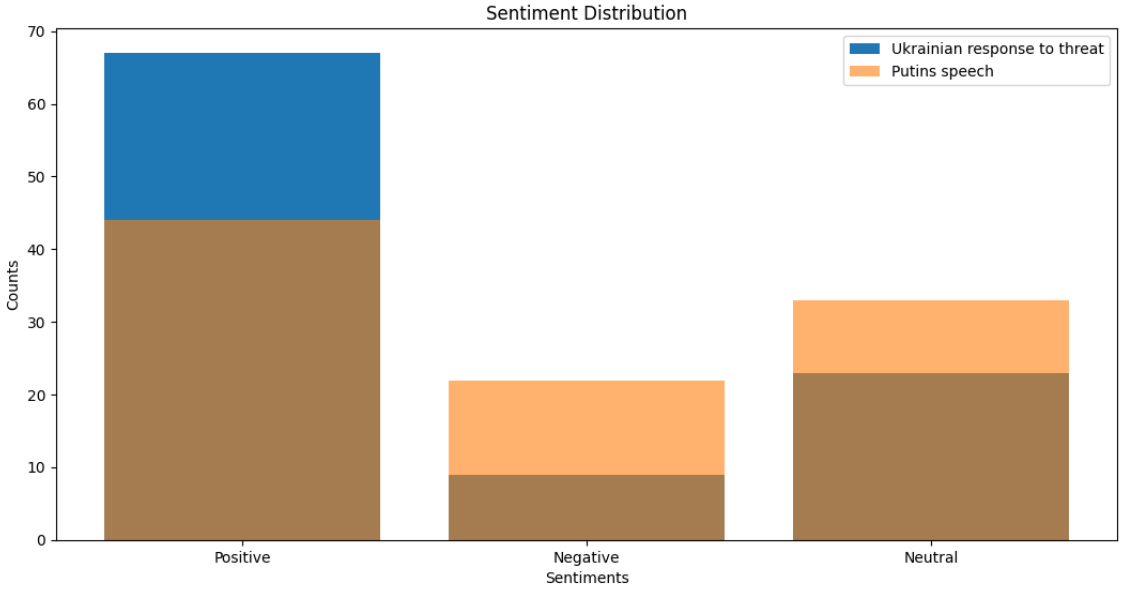


Рисунок 3.2.8 Результат обробки коментарів під відео про повномасштабне вторгнення

Легко побачити, що для обох відео більша частина коментарів була позитивною. Та необхідно враховувати, що навчена модель аналізу тональності тексту не ідеально працює з контекстом. Тож, коментарі під виступом Володимира Путіна із текстом, наприклад, «I am so glad, people of Donbass will finally be saved!» будуть оцінюватись як позитивні, а не негативні.

Висновки

Аналіз тональності тексту та обробка природної мови до сих пір лишаються одними з найскладніших та найважливіших задач машинного навчання та ШІ. Людська мова постійно змінюється, ускладняється новими словами, конструкціями, до звичайних слів додаються нові контексти та підтексти, іронія та сарказм стають все більш витонченими, настільки, що навіть людині буває важко їх вловити. У сучасному світі інформація є найважливішим та найдорожчим ресурсом. Урядам країн важливо знати думку своїх громадян щоб розуміти, чи залишаться вони головами країни ще на декілька років, чи схвалюється їхня політика, хто їх основними виборцями тощо. Аналіз тональності тексту є потужним інструментом, що дає можливість зрозуміти громадські настрої та у подальшому базувати свої рішення на релевантних даних про ці настрої.

У роботі було досліджено 2 різних алгоритми аналізу тональності тексту: наївний баєс та алгоритм на основі опорних векторів. Було розроблено та натреновано 10 моделей, по одній на кожну релевантну модифікацію алгоритмів, на даних із соціальної мережі Twitter та платформи Youtube. Було показано, що алгоритм на основі опорних векторів повертав кращі результати, ніж наївний баєс. Такий результат залежить, в першу чергу, від обраних для тренувального набору даних, їх кількості та якості. Якість роботи моделей наївного баєсу можна спробувати покращити, застосувавши до даних без тегів емоційного забарвлення нейронну мережу BERT або алгоритм TextBlob замість алгоритму VADER. Також можна спробувати змінити функцію попередньої обробки даних, виключити більше нерелевантних токенів з тренувального набору, або навпаки, залишити у тексті більше слів.

Також якість моделі для конкретного завдання даної роботи можна було підняти, створивши власний великий набір даних і надавши кожному з елементів набору даних тег емоційного забарвлення вручну. Таким чином

можна було б переконатися у тому, що модель буде вірно оцінювати контекст завдання.

За допомогою YouTube API було зібрано коментарі під двома відео, що вийшли на початку повномасштабного вторгнення Російської Федерації в Україну: звернення президента РФ Володимира Путіна щодо початку спеціальної воєнної операції на Донбасі, та відповідь на це президента України Володимира Зеленського. Коментарі було проаналізовано за допомогою попередньо навченої моделі на основі опорних векторів із Гаусовою радіально-базисною функцією пі підходом One-vs-Rest. Було виведено графік кількості негативних, нейтральних та позитивних коментарів під кожним із відео.

Розвиток технологій для роботи із природною мовою буде лише прискорюватись, що вже підтверджує популярність, наприклад, чат-ботів на основі ШІ. Такі технології будуть використовуватись все більше як для вирішення бізнес-проблем, так і для проведення досліджень у соціальній сфері життя людей. Можна навіть сказати, що моделі машинного навчання для обробки природної мови «розумнішають» разом із ускладненням людської мови. Що буде у майбутньому і до чого призведе такий розвиток, невідомо, але за цим дуже цікаво спостерігати.

Джерела

[1] ERNST, Michael D. Natural language is a programming language: Applying natural language processing to software development. In: *2nd Summit on Advances in Programming Languages (SNAPL 2017)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2017.

[2] CUNHA, Alexandre Ashade Lassance; COSTA, Melissa Carvalho; PACHECO, Marco Aurélio C. Sentiment analysis of youtube video comments using deep neural networks. In: *Artificial Intelligence and Soft Computing: 18th International Conference, ICAISC 2019, Zakopane, Poland, June 16–20, 2019, Proceedings, Part I 18*. Springer International Publishing, 2019. p. 561-570.

[3] BHUIYAN, Hanif, et al. Retrieving YouTube video by sentiment analysis on user comment. In: *2017 IEEE International Conference on Signal and Image Processing Applications (ICSIPA)*. IEEE, 2017. p. 474-478.

[4] ASGHAR, Muhammad Zubair, et al. Sentiment analysis on youtube: A brief survey. *arXiv preprint arXiv:1511.09142*, 2015.

[5] GO, Alec, et al. Twitter sentiment analysis. *Entropy*, 2009, 17: 252.

[6] GIACHANOU, Anastasia; CRESTANI, Fabio. Like it or not: A survey of twitter sentiment analysis methods. *ACM Computing Surveys (CSUR)*, 2016, 49.2: 1-41.

[7] SHARMA, Anuj; DEY, Shubhamoy. A document-level sentiment analysis approach using artificial neural network and sentiment lexicons. *ACM SIGAPP Applied Computing Review*, 2012, 12.4: 67-75.

[8] DOS SANTOS, Cicero; GATTI, Maira. Deep convolutional neural networks for sentiment analysis of short texts. In: *Proceedings of COLING 2014, the 25th international conference on computational linguistics: technical papers*. 2014. p. 69-78.

[9] LONDHE, Alka; RAO, PVRD Prasada. Aspect Based Sentiment Analysis–An Incremental Model Learning Approach Using LSTM-RNN. In: *Advances in Computing and Data Sciences: 5th International Conference, ICACDS 2021, Nashik,*

India, April 23–24, 2021, Revised Selected Papers, Part I 5. Springer International Publishing, 2021. p. 677-689.

[10] FELDMAN, Ronen. Techniques and applications for sentiment analysis. *Communications of the ACM*, 2013, 56.4: 82-89.

[11] THANAKI, Jalaj. *Python natural language processing*. Packt Publishing Ltd, 2017.

[12] PALMER, Martha. Natural language processing. 2003.

[13] LIDDY, Elizabeth D. Natural language processing. 2001.

[14] NADKARNI, Prakash M.; OHNO-MACHADO, Lucila; CHAPMAN, Wendy W. Natural language processing: an introduction. *Journal of the American Medical Informatics Association*, 2011, 18.5: 544-551.

[15] CHOWDHARY, KR1442; CHOWDHARY, K. R. Natural language processing. *Fundamentals of artificial intelligence*, 2020, 603-649.

[16] HARDENIYA, Tanvi; BORIKAR, Dilipkumar A. Dictionary based approach to sentiment analysis-a review. *International Journal of Advanced Engineering, Management and Science*, 2016, 2.5: 239438.

[17] ASTYA, Parmanand, et al. Sentiment analysis: approaches and open issues. In: *2017 International Conference on computing, Communication and automation (ICCCA)*. IEEE, 2017. p. 154-158.

[28] YUSOF, Nor Nadiah; MOHAMED, Azlinah; ABDUL-RAHMAN, Shuzlina. Reviewing classification approaches in sentiment analysis. In: *Soft Computing in Data Science: First International Conference, SCDS 2015, Putrajaya, Malaysia, September 2-3, 2015, Proceedings 1*. Springer Singapore, 2015. p. 43-53.

[19] ELBAGIR, Shihab; YANG, Jing. Twitter sentiment analysis using natural language toolkit and VADER sentiment. In: *Proceedings of the international multiconference of engineers and computer scientists*. 2019. p. 16.

[20] HUTTO, Clayton; GILBERT, Eric. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In: *Proceedings of the international AAAI conference on web and social media*. 2014. p. 216-225.

[21] ADARSH, R., et al. Comparison of VADER and LSTM for sentiment analysis. *International Journal of Recent Technology and Engineering*, 2019, 7.6: 540-543.

[22] ELBAGIR, Shihab; YANG, Jing. Sentiment analysis on Twitter with Python's natural language toolkit and VADER sentiment analyzer. In: *IAENG Transactions on Engineering Sciences: Special Issue for the International Association of Engineers Conferences 2019*. 2020. p. 63-80.

[23] WEBB, Geoffrey I.; KEOGH, Eamonn; MIIKKULAINEN, Risto. Naïve Bayes. *Encyclopedia of machine learning*, 2010, 15: 713-714.

[24] RISH, Irina, et al. An empirical study of the naive Bayes classifier. In: *IJCAI 2001 workshop on empirical methods in artificial intelligence*. 2001. p. 41-46.

[25] DEY, Lopamudra, et al. Sentiment analysis of review datasets using naive bayes and k-nn classifier. *arXiv preprint arXiv:1610.09982*, 2016.

[26] C. J. H. E. Gilbert, "Vader: A parsimonious rule-based model for sentiment analysis of social media text," in Eighth International Conference on Weblogs and Social Media (ICWSM-14). Available at (20/04/16)

[27] V. Kharde and P. Sonawane, "Sentiment analysis of twitter data: a survey of techniques," arXiv Prepr. arXiv1601.06971, 2016

[28] Lewis, D. (1998) Naive Bayes at forty: The independence assumption in information retrieval. In Machine Learning: ECML-98, Proceedings of the 10th European Conference on Machine Learning, Chemnitz, Germany, Springer, Berlin, pp 4-15.

[29] Andrew McCallum and Kamal Nigam (1998) A comparison of event models for Naive Bayes text classification. In AAAI-98 Workshop on Learning for Text Categorization, AAAI Press, pp. 41-48.

[30] Bishop, Pattern recognition and machine learning (2006)

Відгук

на кваліфікаційну роботу бакалавра на тему:

«Аналіз тональності тексту та обробка природної мови за допомогою

штучного інтелекту для розуміння суспільних настроїв щодо

повномасштабного вторгнення Російської Федерації в Україну»

студентки 4-го курсу факультету комп'ютерних наук та кібернетики Київського національного університету імені Тараса Шевченка Гашевої Аліни Михайлівни

Аналіз тональності тексту та обробка природної мови за допомогою штучного інтелекту стали актуальними та необхідними інструментами для розуміння суспільних настроїв. Використання цих інструментів стає особливо важливим для глибокого аналізу великих обсягів даних та виявлення патернів в суспільних реакціях на глобальні соціальні катаклізми, такі, як вторгнення РФ в Україну.

є своїй роботі авторка проводить глибокий аналіз поняття обробки природної мови та аналізу тональності тексту, використовуючи методи наївного Баєса та алгоритму на основі опорних векторів. Особливо цінним є застосування цих підходів до аналізу реальних ситуацій, що дозволяє не тільки розуміти настрої в суспільстві, але й передбачати можливі сценарії розвитку подій.

Авторка провела теоретичний огляд методів аналізу тональності тексту та обробки природної мови, та продемонструвала використання цих методів для створення моделі, що успішно аналізує дані з реального життя, з демонстрацією на відслідковуванні динаміки суспільних настроїв щодо вторгнення РФ в Україну.

Під час роботи над дипломом, студентка продемонструвала здатність досліджувати теоретичні основи алгоритмів, а також вміння застосовувати теоретичний матеріал в практичних задачах. Я вважаю, що робота відповідає всім вимогам, які висуваються до кваліфікаційних робіт, і заслуговує на оцінку «відмінно».

Асистент кафедри обчислювальної
математики факультету комп’ютерних

наук та кібернетики Київського

національного університету
імені Тараса Шевченка

Сергій ДЕНИСОВ

Рецензія

на кваліфікаційну роботу бакалавра на тему:

«Аналіз тональності тексту та обробка природної мови за допомогою

штучного інтелекту для розуміння суспільних настроїв щодо повномасштабного вторгнення Російської Федерації в Україну» студентки 4-го курсу факультету комп'ютерних наук та кібернетики Київського національного університету імені Тараса Шевченка

Гашевої Аліни Михайлівни

Студентка Гашева Аліна Михайлівна у дипломній кваліфікаційній роботі досліджує цікаву та достатньо складну тему обробки природної мови, що на сьогодні є одним з найважливіших напрямів у розвитку методів штучного інтелекту. Частиної цієї галузі є аналіз тональності тексту — який є критично важливим в сучасному світі, де інформація має реальний вплив на думки людей. Особливо цікавим це питання постає в контексті аналізу суспільних настроїв щодо важливих глобальних подій, таких як вторгнення РФ в Україну.

У своїй статті студентка проводить дослідження алгоритмів наївного баєса та алгоритму на основі опорних векторів для обробки природної мови і аналізу тональності тексту. Робота охоплює широкий спектр тем, від теоретичних аспектів до практичних застосувань. Сама проблематика, очевидно, має велике соціальне значення - зокрема у світлі подій останнього року в Україні .

В теоретичній частіні роботи Аліна Гашева показала глибоке розуміння алгоритмів, які вона досліджує. Було представлено і описано математичне підґрунтя, а також надано пояснення кожної з використаних модифікацій.

Проте, до роботи можна зробити і деякі зауваження. Так, недостатньо уваги було приділено проблемі визначення тональності в контекстуальній залежності. Крім того, авторка могла провести більше досліджень щодо ефективності різних алгоритмів в різних контекстах. Набір тренувальних даних, який використано, не є найкращим для висвітлення контексту соціальних настроїв щодо повномасштабного вторгнення.

Ці зауваження не зменшують загальної позитивної оцінки роботи. Вважаю, що дипломна робота відповідає вимогам, які висуваються до бакалаврських робіт, і заслуговує на оцінку відмінно, а її автор заслуговує на присвоєння кваліфікації бакалавра.

Рецензент:

Кандидат технічних наук,

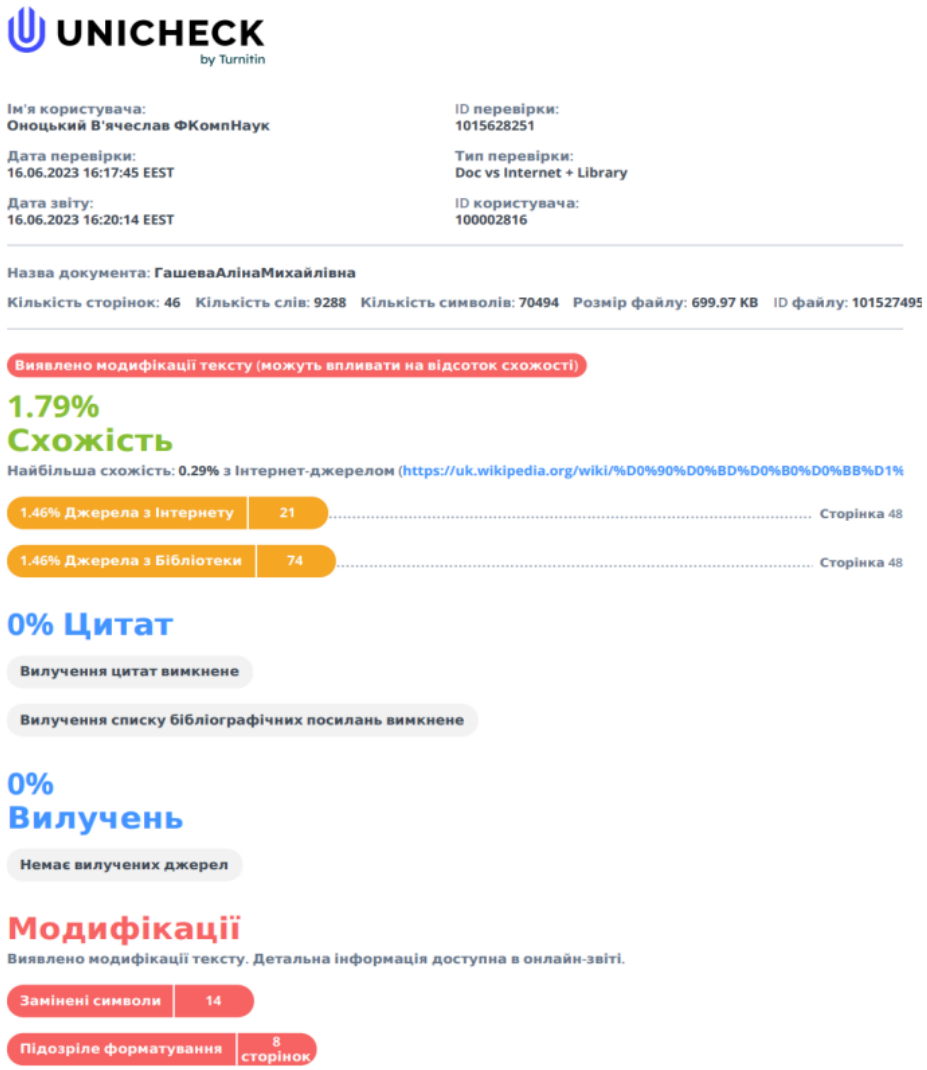
Катерина ГОЛУБЄВА

доцент

КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ ІМЕНІ ТАРАСА ШЕВЧЕНКА

СИСТЕМА ЗАПОБІГАННЯ ТА ВИЯВЛЕННЯ АКАДЕМІЧНОГО ПЛАГІАТУ

Довідка про оригінальність кваліфікаційної роботи за освітнім рівнем бакалавр



Експертна оцінка роботи науковим керівником :

Робота студентки 4-го курсу Гашевої Аліни Михайлівни «Аналіз тональності тексту та обробка природної мови за допомогою штучного інтелекту для розуміння суспільних настроїв щодо повномасштабного вторгнення Російської Федерації в Україну» виконана самостійно, при цьому обсяг цитувань та запозичень становить 1.79% та не перевищує норму.

`

Науковий керівник: _____ Денисов С.В.

(підпис) (ПІБ)

Оператор: _____ Оноцький В.В.

(підпис) (ПІБ)