Шифр М32

ОСОБЛИВОСТІ СУЧАСНИХ МЕТОДИК АВТОМАТИЧНОГО СЕНТИМЕНТ-АНАЛІЗУ ТЕКСТУ

Зміст

Вступ

Актуальність праці. Сентимент-аналіз є важливим напрямком сучасних досліджень у прикладній лінгвістиці. Відомим фактом є те, що однією із найпоширеніших галузей прикладної лінгвістики є машинне опрацювання тексту, значення якого щодня зростає, як і збільшується кількість інформації, яку належить опрацьовувати. Із збільшенням кількості інформації, яку треба аналізувати, зростає актуальність відповідних досліджень на цю тему. Отже, дослідження сентимент-аналізу (аналізу тональності) є одними з набільш актуальних завдань прикладної лінгвістики із погляду сучасних потреб різних галузей.

Іншою надзвичайно важливою причиною актуальності досліджень сентимент-аналізу є практична цінність технологій, пов'язаних з ними. Належить відзначити широку сферу застосування сентимент-аналізу. Його використання доречне в бізнесі (для аналізу коментарів до продукції чи послуги), в мас-медіа (аналіз новин), політичній чи громадській діяльності (аналізування згадок партії чи організації в засобах масової інформації, чи соціальних мережах) та інші варіанти застосування, яких може бути величезна кількість; важливо також визначити, у який спосіб можна було б покращити роботу таких програм.

Ще однією важливою причиною актуальності таких досліджень є їхня важливість для розвитку штучного інтелекту. Дослідження та розуміння принципів машинного аналізу людського тексту є важливим із кількох причин. Сьогодні є помітними відмінності у підході до аналізу тексту спеціалізованим програмним забезпеченням та людиною. Текст, який і є безпосереднім об’єктом сентимент-аналізу, є писемною формою людського мовлення. Своєю чергою, загальновідомим є факт того, що є однією із найважливіших функцій мови є експресивна функція, тобто відображення внутрішнього світу людини, її емоцій. У цьому і є помітною відмінність між оцінкою тексту людиною, яка є добре обізнаною в емоційному складникові людського буття, та машиною, у комп’ютерних програмах для якої важко формалізувати людські емоції.

Дослідження штучного інтелекту ще від початку свого існування використовують як орієнтир єдиний відомий на цей момент тип інтелекту – людський, свідченням чого може слугувати тест Тюрінга, ідея якого зберігає актуальність і сьогодні, попри існування його критики. Важливим аспектом людської свідомості та підсвідомості є емоції та почуття, які у мовленні можуть бути як експліцитними, так і імпліцитними. Отже, виходячи із сукупності наведених фактів, можна твердити про те, що дослідження технології сентимент-аналізу є важливим не тільки для розвитку прикладної лінгвістики, а й для розвитку технологій штучного інтелекту. Окрім того, на сьогодні програмного забезпечення, яке б аналізувало українськомовні тексти створено недостатню кількість.

Об'єктом дослідження є аналіз технології сентимент-аналізу.

Предметом дослідження є аналіз найпоширеніших методів проведення сентимент-аналізу за допомогою спеціалізованого програмного забезпечення, порівняння результатів роботи таких програм з оцінкою тональності, зробленого людьми, визначення типових помилок у роботі комп'ютерних програм.

Мета цієї праці – проаналізувати алгоритми роботи програмного забезпечення, використовуваного для сентимент-аналізу, розглянути особливостей, пов'язаних із сентимент-аналізом за відповідними методами, визначити ефективність роботи таких методів за допомогою порівняння результатів з результатами оцінювання людиною; дати рекомендації для підвищення ефективності роботи програм.

Для досягнення мети належить виконати такі **завдання**:

- надати загальну характеристику сутності поняття сентимент-аналіз (аналіз тональності);
- розглянути алгоритми, що найчастіше використовуються спеціалізованими програмами для сентимент-аналізу;
- розглянути специфіку оцінки тональності тексту людьми;

- порівняти результати та методи сентимент-аналізу, здійсненого програмним забезпеченням, та людською оцінкою.

Для вирішення поставлених завдань використано такі методи, як індукція та дедукція, описовий метод, метод компонентного аналізу, експеримент.

Наукова новизна цієї полягає в порівнянні результатів оцінки тональності, наданої спеціалізованим програмним забезпеченням, та результатів оцінки тональності, що були надані людьми, а також порівняння методів, що використано для створення програмного забезпечення, та методів, що використовує людина для вирішення поставленого завдання.

Розділ 1. МЕТОДИКИ СЕНТИМЕНТ-АНАЛІЗУ В СУЧАСНІЙ ПРИКЛАДНІЙ ЛІНГВІСТИЦІ

1.1.Сутність сентимент-аналізу

Сентимент-аналіз (англ. *sentimental analysys*) – оцінювання тональності певного заданого тексту за допомогою методів опрацювання природньої мови (NLP), статистики та машинного навчання. У англійськомовних джерелах також вживають термін *opinion mining*, але обов'язково належить відзначити відмінність між поняттями, що стоять за цими двома термінами, хоча доволі часто їх вживають як синоніми. Згідно з визначенням авторитетного «Dictionary by Merriam-Webster», термін *sentiment* визначають як твердження, думку або судження засновані на почуттях [5], тоді як термін *oppinion* розуміють як погляд або судження сформовані у свідомості про конкретну справу або випадок [5]. Звісно, відмінність не є аж настільки великою, але теоретично вона може відіграти певну роль як у дослідженнях, так і в самому процесі аналізу, а також її усвідомлення може допомогти у вивченні цієї проблеми. Загалом сентимент-аналіз є ділянкою прикладної лінгвістики, пов'язаною з проблемою машинного опрацювання природньої мови.

Проблема сентимент-аналізу чітко відображає комплексність прикладної лінгвістики. Тут добре прослідкується перетин цієї галузі знань із дослідженнями штучного інтелекту, класичної психології та інформаційними технологіями.

Необхідність розвитку технології сентимент-аналізу в сучасному світі є доволі очевидною. Сентимент-аналіз може допомагати великій кількості людей з різними видами потреб. Особливого гостро ця необхідність стоїть перед автоматичною оцінкою новин, відгуків про найрізноманітніші товари та послуги, пошуком спаму в інформаційному потоці. Із комерційного погляду, класичним застосуванням цієї технології є збирання та сортування відгуків користувачів про певний продукт або послугу. Завдяки цьому, а саме співвідношенню позитивних та негативних відгуків, можна дізнатися, чи продукт впевнено почуває себе на ринку, чи потребує доопрацювання, чи його взагалі варто прибрати з ринку.

Згадане співвідношення є відносною оцінкою, адже люди у відгуках частіше діляться негативною інформацією, ніж позитивною, детальніше про цей аспект йтиметься нижче.

Сьогодні кількість уже наявної інформації в загальному доступі не може не вражати, при цьому вона збільшується щодня. Як вже було сказано вище, для багатьох цілей життєво необхідна швидка та якісна оцінка тексту. Зрозуміло, що за великих обсягах даних надати оцінку «від людини» вкрай важко, а здебільшого - неможливо. Технології автоматичного комп'ютерного оцінювання і є розв'язком задачі.

Проте проблеметика сентимент-аналізу ускладнюється ще одним дуже значущим чинником. Сучасні комп'ютерні програми без проблем здійснюють, для прикладу, граматичний аналіз тексту, використовуючи чіткі правила власне граматики. У випадку оцінювання тональності ми маємо справу з надзвичайно специфічним продуктом людської свідомості – емоціями, вербалізація яких у тексті є різноманітною, крім лексики на позначення емоцій та оцінної лексики може виражатися, наприклад, за допомогою метафор. На сьогодні для жодної із природних мов світу не створено надійного комп'ютерного інструмента для автоматичного визначення метафор у тексті. На цьому етапі розвитку штучного інтелекту коректний аналіз людських емоцій є справді складним завданням.

Попри таку непросту перешкоду, створено низку методів аналізу тональності тексту, на яких базовано спеціалізовані комп'ютерні програми. На мою думку, належить виокремити такі методи автоматичного визначення тональності тексту:

1. Використання правил із наперед заготовлених шаблонів (rule-based with patterns). Підхід полягає в генерації правил, на основі яких буде визначатися тональність тексту. Для цього текст розбивають на слова або послідовності слів (N-grams). Потім отримані дані використовують для виокремлення найпоширеніших шаблонів, які передають позитивну чи негативну оцінку. Такі шаблони застосовують у випадках, коли правило "ЯКЩО умова ТО висновок" дійсне [2].

2. Машинне навчання без вчителя (unsupervised learning). Такий метод заснований на ідеї того, що найважливішою в тексті є низка термінів, які в цьому тексті найпоширеніші. Якщо визначити тональність цих термінів, то відповідно можна визначити тональність усього тексту [2].
3. Машинне навчання з учителем (supervised learning). За використання цього методу необхідна наявність в базі певної кількості текстів, які є розібрані за емотивним принципом. На базі цих текстів власне й функціонує класифікатор [3].
4. Гібридний метод (hybrid method). Він поєднує всі або кілька вищезгаданих методів. Суть методу полягає в застосуванні класифікаторів на основі інших методів в певній послідовності [7, с. 45].
5. Метод, заснований на теоретико-графових моделях. В основі цього методу лежить припущення про те, що слова в тексті не рівнозначні. Згідно з цією ідеєю певні слова мають більшу вагу, і тому вони більше впливають на тональність тексту. У випадку застосування такого методу аналіз емотивного забарвлення містить кілька етапів:
   1) побудова графа на основі досліджуваного тексту;
   2) рангування його вершин;
   3) класифікація знайдених слів;
   4) обчислення результату [7, с. 46].

Історично склалося так, що традиційний підхід до сентимент-аналізу за своєю суттю є задачею класифікації тексту на дві-три категорії (негативний, позитивний, нейтральний або просто негативний/позитивний). Саме з такого завдання почав свій розвиток аналіз тональності: оцінити тональність відгуків будь-якої тематики.

Об'єкт, щодо якого виражається емоційна оцінка, прийнято називати об'єктом тональності [14, с. 47].

Отже, тональність висловлювання визначається трьома компонентами: суб'єктом тональності (хто висловив оцінку), об'єктом тональності (про кого або про що висловлена оцінка) і власне тональної оцінкою (як оцінили).

Також методи машинного аналізу тональності тексту класифікують так:

1. Наївний Байєсовий класифікатор – класифікує слова в тексті, використовуючи теорему Байєса, тобто визначає тональність спираючись на обставини, що могли бути пов’язані з цією подією.
2. Використання ідеї ентропії – коли елемент, що має меншу ймовірність появи в тексті, ніж інші найбільш поширені елементи, містить більше інформації, має більшу цінність.
3. Метод опорних векторів – розташовує елементи тексту на різних сторонах умовного вектору, цим самим розділяючи їх [6, с. 1061].

1.2. Оцінка результатів сентимент-аналізу

Не менш важливою є оцінка якості сентимент-аналізу. У ній наявні дві найважливіші характеристики: повнота та точність.

Точність доволі очевидно оцінюють з огляду .на те, наскільки результати проведеного аналізу за допомогою спеціалізованої програми збігаються з емоційним забарвленням, наданим самим автором тексту.

Розроблено методики для обчислення цих значень. Варто пам’ятати, що здебільшого отримані числові результати перетворюють у відсоткові значення.

$$\text{Точність} = \frac{\text{кількість правильно розпізнаних думок}}{\text{загальна кількість думок знайдених системою}}$$

Отже, таке поняття як “точність” - це відношення елементів, які були визначені аналітичною програмою так само, як і було у свідомому чи несвідомому задумі автора. Уважають, що здебільшого програма для аналізу є точною, якщо відсоткове визначення точності дорівнює 70% та більше [15, с. 869].

Іншою важливою характеристикою є повнота, яку визначають за такою формулою:

$$\text{Повнота} = \frac{\text{кількість правильно розпізнаних думок}}{\text{загальна кількість думок (як знайдених, так і не знайдених)}}$$

Із вище поданих формул та інформації зрозуміло, що точність та повнота в оцінюванні результатів сентимент-аналізу є доволі схожими, на перший погляд, проте означають різні за своєю суттю речі. Логічно, що переважно число, отримане під час обчислення точності, буде більшим, ніж число, отримане під час обчислення повноти.

Безпосереднім об’єктом сентимент-аналізу є текст, написаний людиною. Як відомо, текст є письмовою формою людського мовлення. Однією із найважливіших функцій мовлення є емотивна функція – вираження емоцій. Виходячи із сукупності наведених фактів, стає зрозуміло, що проблема аналізу тональності пов’язана із психологією набагато більше, ніж це може здаватися на перший погляд, адже об’єктом аналізу є виявлення емоцій [4]. З іншого боку, стає помітним зв’язок з дослідженнями штучного інтелекту. Для абсолютної більшості людей не виникне жодних проблем визначити, чи відгук про певний товар чи послугу є за своєю суттю позитивним чи негативним на підсвідомому рівні, не докладаючи зусиль. Своєю чергою, більшість аналітичних програм досліджують текст як комбінацію окремих слів, проте цей підхід дає не зовсім правильні результати.

Тут добре стає помітним те, що за допомогою аналізу емотивної лексики можна дізнатися основну оцінність, яку вклав автор. На сучасному етапі методи, що використовуються в сучасній прикладній лінгвістиці для аналізу тональності, певною мірою виправдовують себе, особливо завдяки своїй простоті та доволі високій ефективності. Особливо вартісним та цікавим, на мою думку, є використання ідеї ентропії у сентимент-аналізі [9, с. 65]. Звісно, у тексті ми не знайдемо ентропію в звичному її вигляді – прагнення всіх компонентів системи до хаосу, але дослідження кожного окремого слова в тексті використовує таку

ідею. Особливістю цього методу аналізу є те, що найбільше уваги приділяють окремим словам, які містять найбільше емотивого змісту.

Для прикладу можна навести таке речення: *“Після прочитання я можу з впевненістю сказати, що ця книга чудова”.* Якщо відкинути останнє слово в реченні, то стане зрозуміло, що всі інші слова не містять емоційного забарвлення, тоді як слово *чудовий* передає позитивну оцінку.

Натомість у текстах, створених людиною, оцінка може бути імпліцитною, переданою за допомогою різноманітних тропів, стилістичних прийомів, глибинних структур. Прикладом цього може послугувати таке речення: *“Книга була просто надзвичайно цікавою – я заснув після того, як прочитав перший розділ”.* Іншими словами, для інтерпретації цього речення людина використовує знання, а не окремі значення слів: *Якщо книга зовсім не цікава, то людина, читаючи її може заснути*. Експліцитне позитивне емоційне забарвлення в цьому реченні має слово *цікавий*, тоді як інші слова не мають такого забарвлення, з огляду на це речення буде розпізнане як позитивно забарвлене, хоча будь-яка людина скаже, що це не так.

Як вже було згадано, кількість негативних відгуків, відносно позитивних доволі значна, навіть у товарів та послуг найвищої якості. Психологи вказують на цю особливість людської психіки: люди частіше діляться негативною інформацією й не бачать потреби поширювати позитивну, оскільки позитивну інформацію ми сприймаємо як «норму», а негативну як відхилення від «норми». Таку специфіку треба враховувати, створюючи програми для аналізу, особливо під час висновування.

На мою думку, важливість сентимент-аналізу полягає не тільки в його практичній цінності для людей, а й в тому, що дослідження аналізу тональності має велике значення для розвитку штучного інтелекту загалом. Штучний інтелект намагаються створити за подобою людського – єдиної відомої форми інтелекту (згадаємо хоча б знаменитий тест Тюрінга). І на сьогоднішній день людські емоції та різноманітні їх прояви залишаються не надто зрозумілими для

машин. Саме тому технологія аналізу тональності є такою важливою для цієї галузі.

Отже, сентимент-аналіз – оцінка тональності певного заданого тексту за допомогою методів обробки природньої мови, статистики та машинного навчання. Іншою поширеною назвою для цього терміна є аналіз тональності. Проблема сентимент-аналізу прекрасно відображає комплексність прикладної лінгвістики, добре прослідкується дотичність із дослідженнями штучного інтелекту, класичною психологією та інформаційними технологіями. Сентимент-аналіз є надзвичайно актуальним сьогодні, адже має як і велику практичнну цінність для широкої категорії людей, так і цінність для прикладної лінгвістики та досліджень і розроблення штучного інтелекту. Попри своєрідність людського мовлення, поданого в текстовому вигляді, та певну обмеженість аналітичних програм на цьому етапі розвитку, існують доволі надійні та результативні методи оцінювання тональності тексту.

Розділ 2. МАШИННЕ ОЦІНЮВАННЯ ТЕКСТУ

2.1.Аналіз найбільш поширених типів текстів

Як вже було згадано, алгоритми, за якими спеціалізовані програми визначають тональність тексту, та алгоритми, за якими люди можуть визначити тональність тексту, для прикладу, коментаря-відгука, відрізняються. Програми для сентимент-аналізу, здебільшого, вишукують окремі ключові слова, які мають яскраво виражене емоційне забарвлення, на основі чого робиться висновок про забарвлення тексту загалом [12, с. 99]. Звісно, щодо такого методу виникають певні питання, зокрема щодо його надійності. Для перевірки надійності роботи програм спочатку було проаналізовано коментарі, щоб виявити типові тексти. З'ясовано, що в коментарях користувачі часто застосовують, наприклад, сарказм. З огляду на це, було сконструйовано «типові коментарі».

Особливо підступним для таких програм є *сарказм*. Для прикладу я провів аналіз речення: *"Книга була просто надзвичайно цікавою – я заснув після того, як прочитав перший розділ"*.

Аналіз цього речення, як й інших, здійснено за допомогою вільно розповсюджуваної версії програми Text Analysys API. У результаті аналізу (див. рис. 1) речення було розпізнано як позитивно забарвлене.

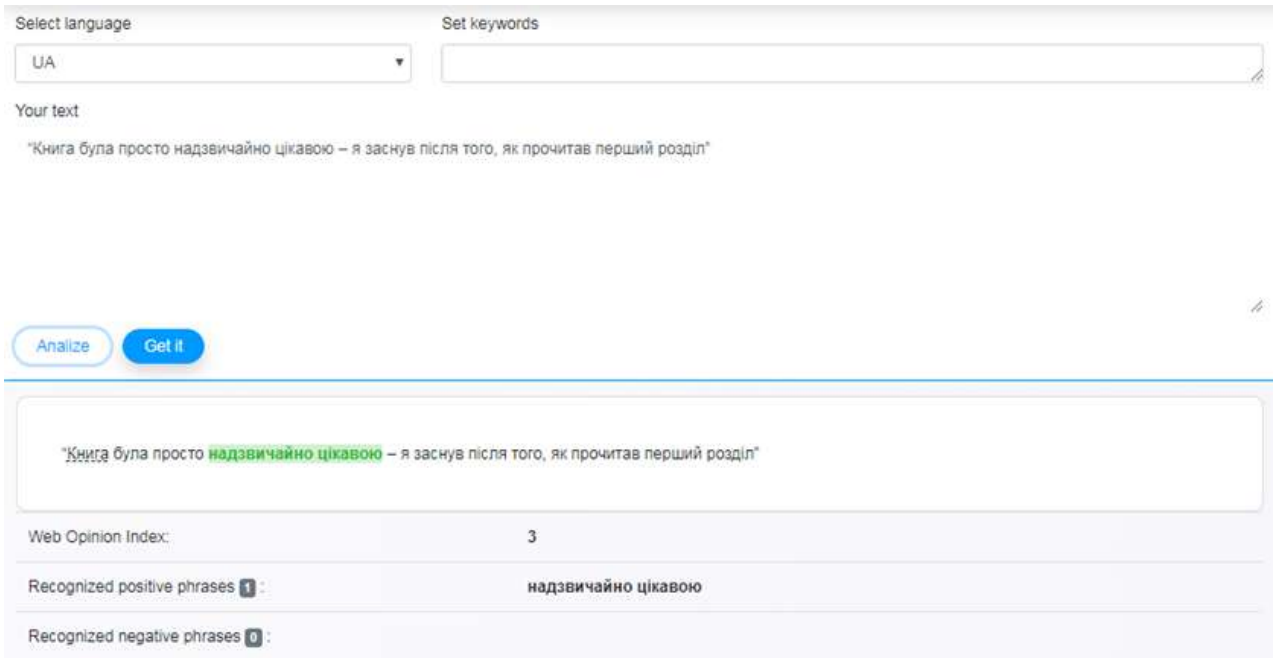

Рис. 1. Результати сентимент-аналізу, здійсненого Text Analysys API

Темою речення було визначено слово *“книга”*, що є цілком правильним результатом. Однак як позитивно забарвлену фразу було визначено *“надзвичайно цікавою”*, а як нам відомо – саме в цьому елементі і прихований сарказм. Параметр “Web opinion index”, що й відповідає за полярну оцінку тональності, показав результат “3”, що є позитивним результатом.

Для порівняння проналізовано таке речення: *“Це найкраща книга, яку я читав”*.

Аналіз цього речення (див. рис. 2) дав нам результат “3”, де позитивно забарвленим був визначений елемент *“найкраща”*. Результат аналізу цього однозначно позитивно забарвленого речення був абсолютно ідентичний результату аналізу попереднього речення. Звісно, аналіз та порівняння тільки двох речень не дає змоги дійти якихось певних висновків, але може вказати на певну тенденцію.

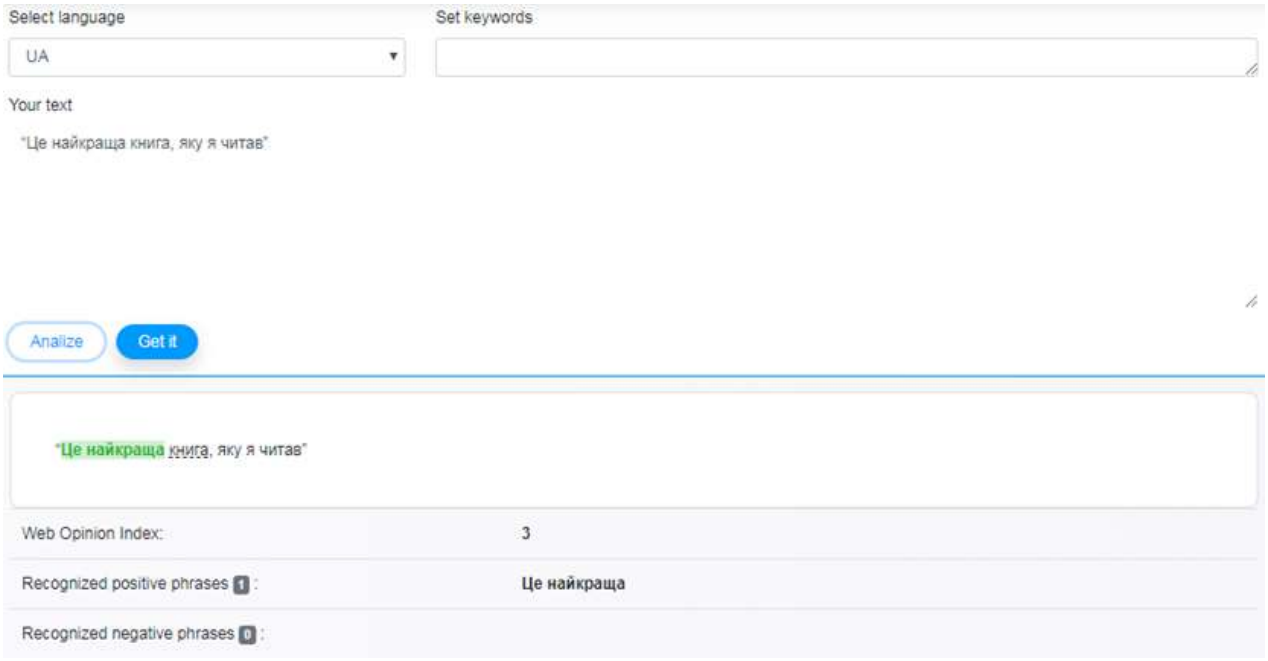

Рис. 2. Результати сентимент-аналізу, здійсненого Text Analysys API

Обов'язково належить відзначити один важливий аспект, пов’язаний із застосуванням програм для сентимент-аналізу на практиці. Із попереднього дослідження речення з використанням сарказму може виникнути враження, що це може вплинути на достовірність відображення точності та повноти результатів аналізу. І справді, за малої наявної кількості коментарів, навіть один

неправильно визначений коментар може негативно вплинути на результати аналізу. Проте існує одне “але”. На практиці такі програми застосовують для аналізу великої кількості коментарів (сотні, тисячі) [11, с.70]. За наявності малої кількості відгуків, власнику продукту немає жодної потреби застосовувати спеціалізовані програми для визначення тональності текстів, адже набагато доцільніше та точніше буде проаналізувати текст власними силами. Загалом така помилка навряд чи матиме значний вплив на загальну статистику проведеного аналізу.

Також у цьому дослідженні було проведено аналіз кількох різних типових речень, схожих на типові коментарі-відгуки на товари.

Розгляньмо таке речення: *"Коли я тільки почав читати цю книгу, то вона мені сподобалася, але під кінець я зрозумів, що ця книга жахлива"*.

Спеціалізована програма визначила рівень оцінки тональності значенням “0” (див. рис. 3).

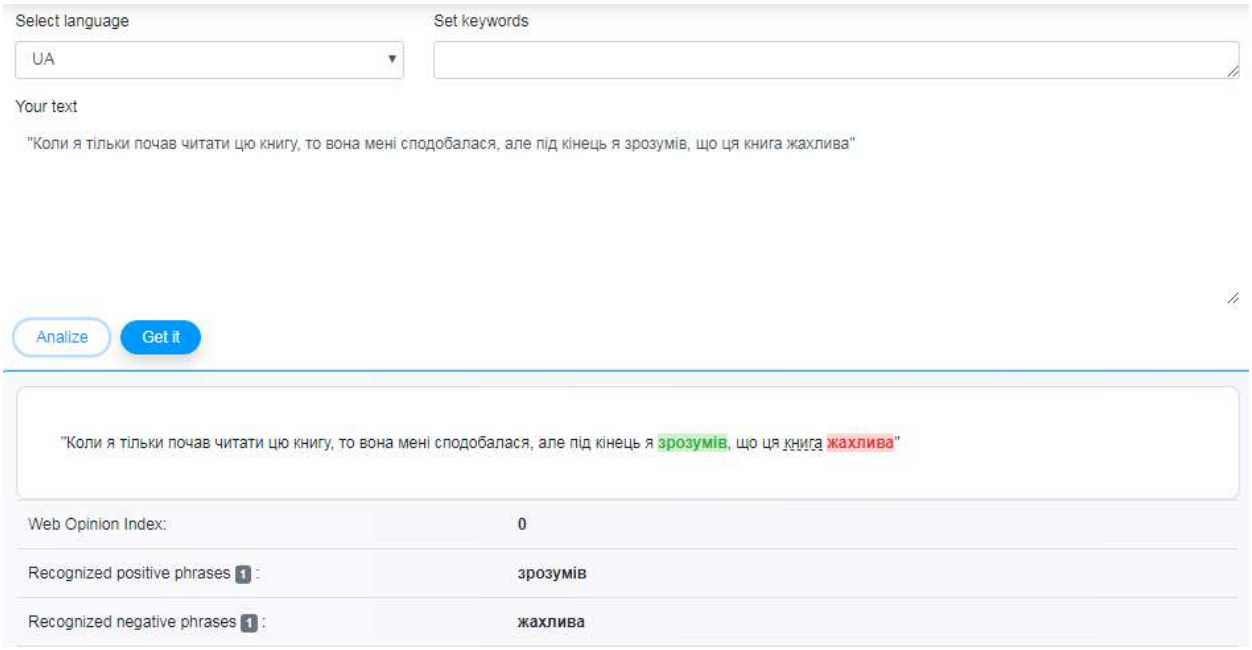

Рис. 3. Результати сентимент-аналізу, здійсненого Text Analysys API

Було виокремлено два емоційно забарвлені елементи: позитивний – *“сподобалася”*, та негативний – *“жахлива”*. У підсумку комбінація протилежних за емоційним забарвленням компонентів знівелювала їхні числові значення, урівноваживши їх. Отриманий висновок є доволі логічним, адже

сукупність протилежних значень ускладнює однозначне трактування речення [10, с. 555], тобто надання нейтрального значення ускладненому неоднозначному реченню є найбільш доцільним рішенням, через складність однозначного трактування в такому випадку. Зрозуміло, що неоднозначні речення, тобто речення, які мають кілька елементів, які дисонують між собою можуть набувати якогось конкретного значення, попри те, що це не очевидно з першого погляду. За своєю суттю це речення є негативно забарвленим, адже останній елемент речення, який є негативно забарвленим, повністю нівелює попередній позитивний елемент. Речення такого типу є доволі поширеними, тому доцільно розуміти їх суть для повноцінних досліджень та для розвитку цієї технології.

Доволі поширеними також є випадки, коли в тексті немає будь-яких слів, за якими можна було б визначити його тональність. Зрозуміло, що спеціалізоване програмне забезпечення не виокремить слів, на яких можна було би ґрунтувати аналіз, тому в таких випадках оцінка сентимент-аналізу дорівнюватиме нульовому (нейтральному) значенню.

Як приклад розгляньмо таке речення: *“Ця книга дуже своєрідна”*. Спеціалізована програма надала цьому реченню нейтральне нульове значення, що й очікувалося (див. рис. 4).

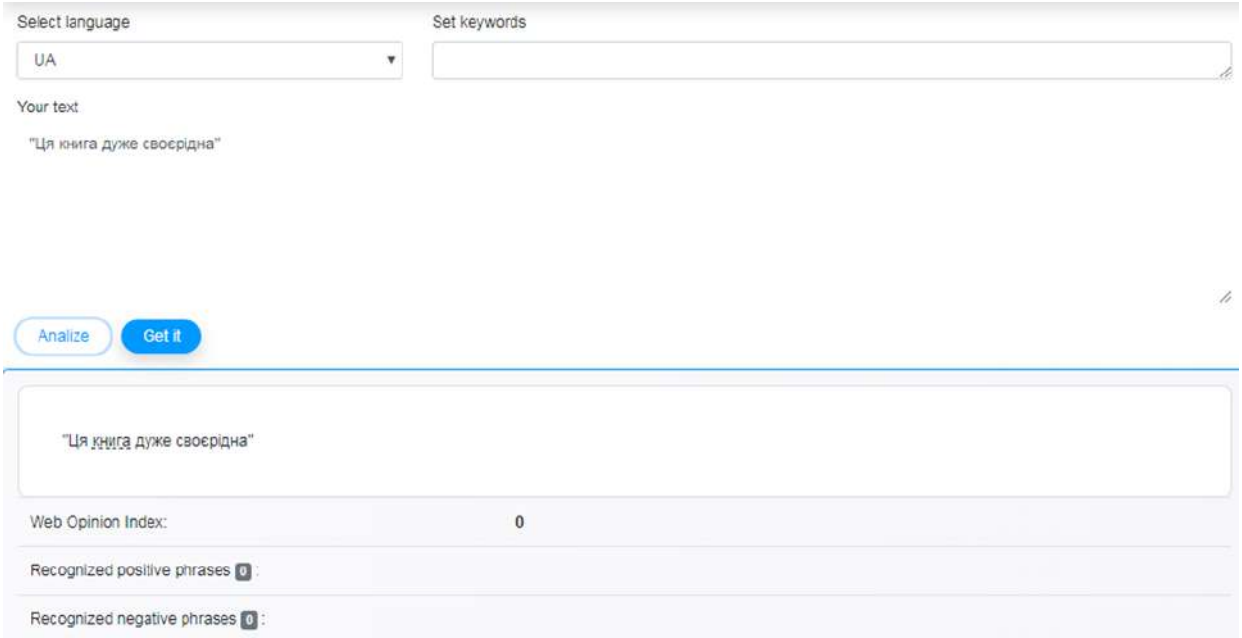

Рис. 4. Результати сентимент-аналізу, здійсненого Text Analysys API

Жодний складник речення не отримав оцінки завдяки аналізу тональності. Це відбулося через те, що в цьому тексті немає складників, які б транслювали емоції, надані автором [13, с. 81]. Надання нейтрального значення такому тексту є найбільш оптимальним вирішенням такої проблеми.

Своєрідність таких текстів також полягає в неоднозначності їхнього трактування. Відсутність емоційного забарвлення всіх елементів і передбачає невизначеність та нейтральність з боку мовця [13, с. 83]. Особливо чітко це помітно на прикладі коментарів та відгуків, схожих на розглянутий.

Варто розуміти, що такі коментарі становлять меншість у загальній масі, тому проблема в тому, що їх не можна класифікувати на позитивні та негативні, зовсім не є критичною.

2.2.Аналіз ненормативних мовних форм

Загальновідомим є факт того, що стандартизовану літературну мову не використовують у «чистому» вигляді в повсякденному житті. Розмовний стиль, соціолекти містять різноманітні мовні одиниці, які можуть бути ще не відображеними навіть у спеціалізованих словниках. В українській мові наявні численні ненормативні конструкції, якими активно користуються звичайні мовці, численні сленгові слова, ненормативні мовні запозичення та інше. В інтернет-мовленні також використовують діалектну лексику і фразеологію, інколи навіть трапляються діалектні граматичні конструкції.

Складності додає також те, що попри високий рівень грамотності населення в Україні, у всесвітній мережі Інтернет можна натрапити на величезну кількість граматичних, орфографічних та інших помилок у текстах, оскільки комунікації в Інтернеті властива спонтанність, інтернавти не надають великого значення мовним огріхам. Більшість людей без особливих зможе дати оцінку тексту, який містить вислови-неологізми чи слова з орфографічними помилками чи описками, тоді як комп’ютерні програми не зможуть визначити емоційно забарвлені слова, які написані з порушенням мовних правил, і речення може бути розпізнаним як нейтральне, або якщо у реченні наявні два або більше

елементів з різним емоційним забарвленням, елемент, який заперечує оцінне значення попереднього, може бути не розпізнаний, і як наслідок, для прикладу, позитивне речення може бути розпізнане як негативне. Сутність проблем, розглянутих у цій главі дуже схожа [2]. Особливу актуальність має проблема орфографічних та граматичних помилок, які спотворюють нормативну форму слів та мовних конструкцій.

Ще однією задачею у комп'ютерному сентимент-аналізі є розпізнавання слів, у яких використано некоректні символи. Наприклад, інтернавти використовують цифру "0" замість літери "о". Найчастонішим охріхом є використання літери "ы" з російської розкладки клавіатури замість української літери "і". Помилки такого типу здебільшого через неуважність та є доволі поширеними у Всесвітній мережі Інтернет.

Існує ще один аспект, який обов'язково належить розглянути. Дві найбільш поширені галузі застосування сентимент-аналізу: мас-медіа та бізнес (відгуки і коментарі про товари та послуги) [14, с. 46]. Згадані огріхи часто трапляються в коментарях та відгуках у мережі інтернет, але доволі рідко в засобах масової інформації в мережі. Це пояснюється більшим рівнем уважності до статей в електронних виданнях та рівнем освіченості авторів, які працюють у таких виданнях. Звісно, хоча ця тема набагато більш актуальна для тих, хто зацікавлений у аналізі коментарів до товарів та послуг, люди, які зацікавлені в аналізі масових медіа також не повинні забувати про це.

Для дослідження ефективності комп'ютерних програм сентимент-аналізу протестовано низку речень, які містять орфографічні помилки (див. рис. 5 та 6).

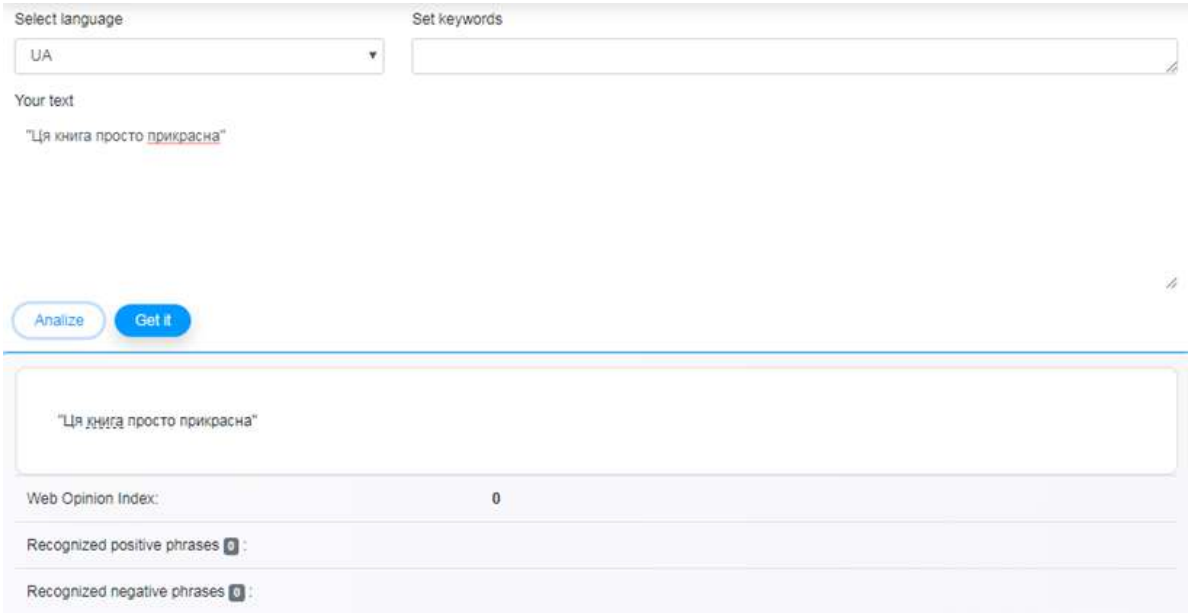

Рис. 5. Результати сентимент-аналізу, здійсненого Text Analysys API

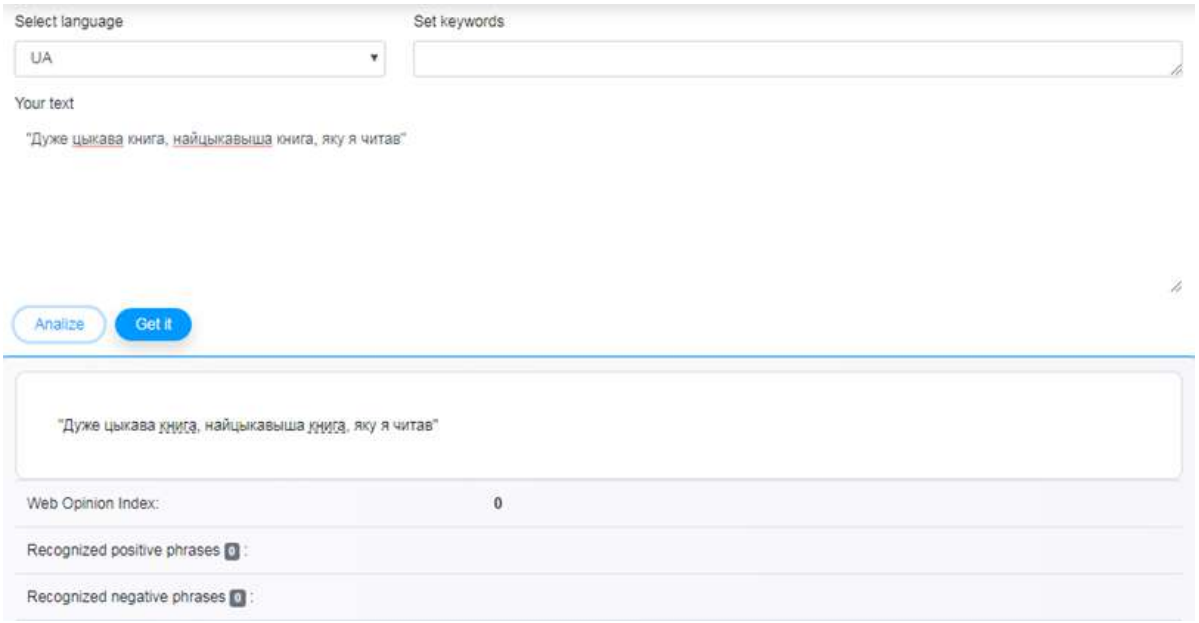

Рис. 6. Результати сентимент-аналізу, здійсненого Text Analysys API

Перше речення має такий вигляд: *"Ця книга просто прикрасна"*. У цьому реченні ми бачимо спеціально зроблену помилку у слові *“прекрасна”*, що є доволі поширеним на практиці. Спеціалізована комп’ютерна програма не виявила жодного емоційно забарвленого слова, що й логічно, адже слів із офографічними помилками у базі даних цієї програми немає. Як наслідок, речення, якому мовець надав позитивного забарвлення отримало нейтральну оцінку від програми, через одне неправильно записане слово.

Друге речення має такий вигляд: *"Дуже цыкава книга, найцыкавыша книга, яку я читав"*. Цей приклад ілюструє вищезгадану проблему - використання літери “ы” з російської розкладки клавіатури замість української літери “і”. Спеціалізована програма для сентимент-аналізу визначила тональність цього речення як нейтральну. Це відбулося внаслідок того, що емоційний складник не був розпізнаний у тексті, оскільки у словах використано літери, які містять інше кодування [16, с. 431].

Отже, алгоритми, за якими спеціалізовані програми визначають тональність тексту, та алгоритми, за якими люди можуть визначити тональність, відрізняються. Програми для сентимент-аналізу здебільшого шукають окремі ключові слова, які виражають емоційно-експресивну оцінку, на основі чого робиться висновок про забарвлення тексту загалом. Такий метод виправдовує себе у доволі значній кількості випадків, але є ситуації, з якими програми не справляються, використовуючи цей підхід. Таке вираження людських емоцій як сарказм, зазвичай не розпізнається машиною, або інтерпретується нею некоректно. Належить зауважити, що такі технічні засоби використовуються під час опрацювання великої кількості інформації, за неможливості таких дій людиною. Ще двома своєрідними аспектами у цій задачі є наявність у реченні складників, що транслюють протилежну емоційно-експресивну оцінку, що «врівноважують», «нейтралізують» один одного, внаслідок чого ми отримуємо нейтральний результат, та тексти, де відсутні будь-які слова, що підлягають такій оцінці, що унеможливлює їх трактування машиною. Варто також окремо виділити проблеми, які стосуються не структури речення, як попередні, а структури слів: використання ненормативного написання слів (за допомогою символів із інших абеток), використання ненормативних слів та орфографічні помилки у словах. Треба також відзначити те, що програм для сентимент-аналізу, які підтримують українську мову, є не надто багато на цей момент.

Розділ 3. ПОРІВНЯННЯ ОЦІНКИ ТЕКСТУ, ЗРОБЛЕНОГО ЛЮДИНОЮ, З РЕЗУЛЬТАТОМ АНАЛІЗУ ПРОГРАМАМИ

3.1.Людська оцінка поширених текстів та її порівняння з результатом програмного аналізу

Ефективність комп'ютерних програм сентимент-аналізу визначають, порівнюючи результати роботи програми із результатами оцінювання, здійссненого людьми [1]. Звісно, дослідження методів, за якими люди оцінюють прочитаний текст, стосуються низки галузей знань – когнітології, психології, лінгвістики. Варто пам'ятати, що принципи, за якими відбувається розуміння та усвідомлення емоцій, вербалізованих автором тексту, у людей, які цей текст читають, та у спеціалізованих комп'ютерних програм, для аналізу тональності тексту, відрізняються. Програмне забезпечення запрограмоване ділити речення на окремі слова, тоді як люди радше сприймають речення загалом.

Обов'язково треба відзначити, що сприйняття тексту, з погляду людини, доволі суб'єктивне, і залежить від великої кількості найрізноманітніших чинників [7, с. 46]. Особливо важливу роль грає суб'єктивність сприйняття людиною дійсності. На сприйняття людиною чогось дуже сильний вплив також має культурне середовище, у якому вона живе. Інтерпретація тексту може бути різною навіть у людей, які належать до одного «культурного» прошарку, до однієї соціальної групи. Не кожний зможе зрозуміти сарказм, ужитий в певному тексті.

Для прикладу розгляньмо результати опитування, проведеного серед користувачів інтернету. В опитуванні взяло участь 35 осіб (80% з яких віком 17-25 років). Респондентам було запропоновано оцінити тональність речення, яке за своєю суттю є типовим коментарем-відгуком. Треба було вибрати одну із оцінок: "Позитивний", "Нейтральний (важко визначити однозначно)" "Негативний". Такі відповіді дають змогу чітко (хоча не дуже точно) виявити оцінність тексту.

Це речення було розпізнане спеціалізованою програмою як позитивне: *“Книга була просто надзвичайно цікавою – я заснув після того, як прочитав перший розділ”*. На рис. 7 уміщено відповіді респондентів.

Рис. 7. Результати оцінювання респондентами речення *“Книга була просто надзвичайно цікавою – я заснув після того, як прочитав перший розділ”*

Більшість опитаних схилилася до думки, що відгук є негативним за оцінністю. Як ми можемо зрозуміти з цього, на відміну від спеціалізованого програмного забезпечення, більшість людей (74,3%) здатні розпізнати сарказм у писаному тексті [8, с. 21]. 17,1% опитаних визначили текст як нейтральний, лише 8,6% визначили його як позитивний.

Належить відзначити, що трактування сарказму є достатньо складною когнітивною процедурою, що й помітно за результатами опитування: хтось навіть не помітив сарказму і сприйняв речення буквально. Надзвичайно важливо розуміти те, що при оцінці тексту людиною, необхідно враховувати людський фактор. Людина може не зрозуміти до кінця умови тексту, чи питання, бути неуважною під час проведення аналізу, втратити концентрацію під час проведення аналізу та інше.

Спеціалізована програма визначила текст *“Це найкраща книга, яку я читав”* як позитивний. Щодо опитування, то його результати збігаються із результатами програми (рис. 8).

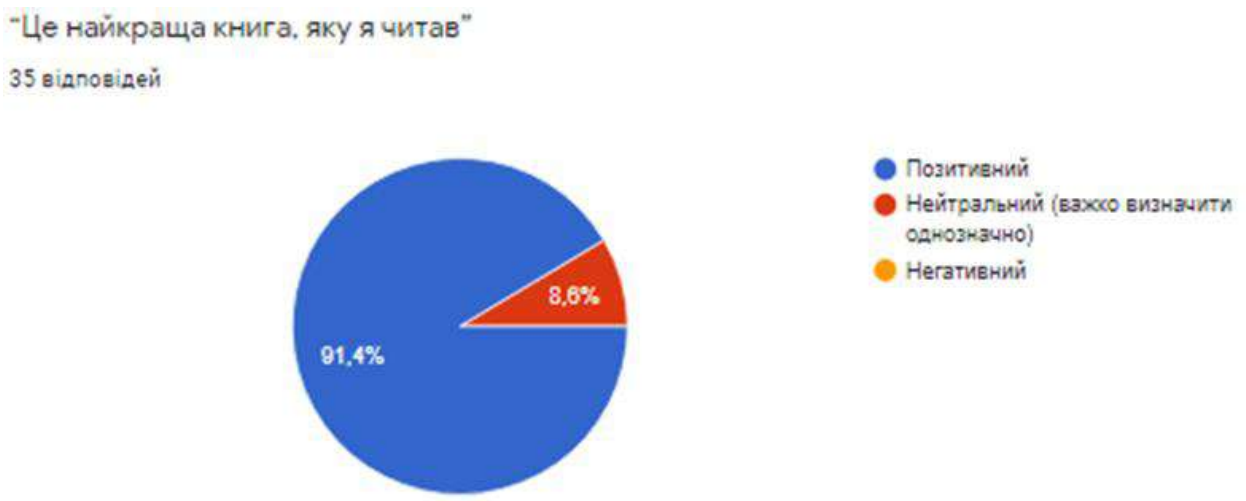

Рис. 8. Результати оцінювання респондентами речення *“Це найкраща книга, яку я читав”*

Лише 8,6% опитаних визначили текст як нейтральний, і жодний респондент не вважав його негативним. На перший погляд, подане речення є занадто простим як для цього опитування, але воно чудово ілюструє аргумент, наведений в попередньому абзаці. Люди суб’єктивно сприймають навіть простий текст, тим більше, що на їхнє сприйняття могли впливати різні чинники, тоді як програма справилася з аналізом цього речення на високому рівні, і на відміну від людини, може аналізувати надзвичайно великі обсяги тексту без «людської суб’єктивності».

Машина визначила речення *"Коли я тільки почав читати цю книгу, то вона мені сподобалася, але під кінець я зрозумів, що ця книга жахлива"* як нейтральне, адже в ньому наявні дві протилежні оцінки, сума числових значень яких утворила нульове значення. Більшість респондентів визначили негативне забарвлення речення. Чверть респондентів оцінила речення як нейтральне, або таке, забарвлення якого важко визначити, один респондент назвав подане речення позитивним (рис. 9).

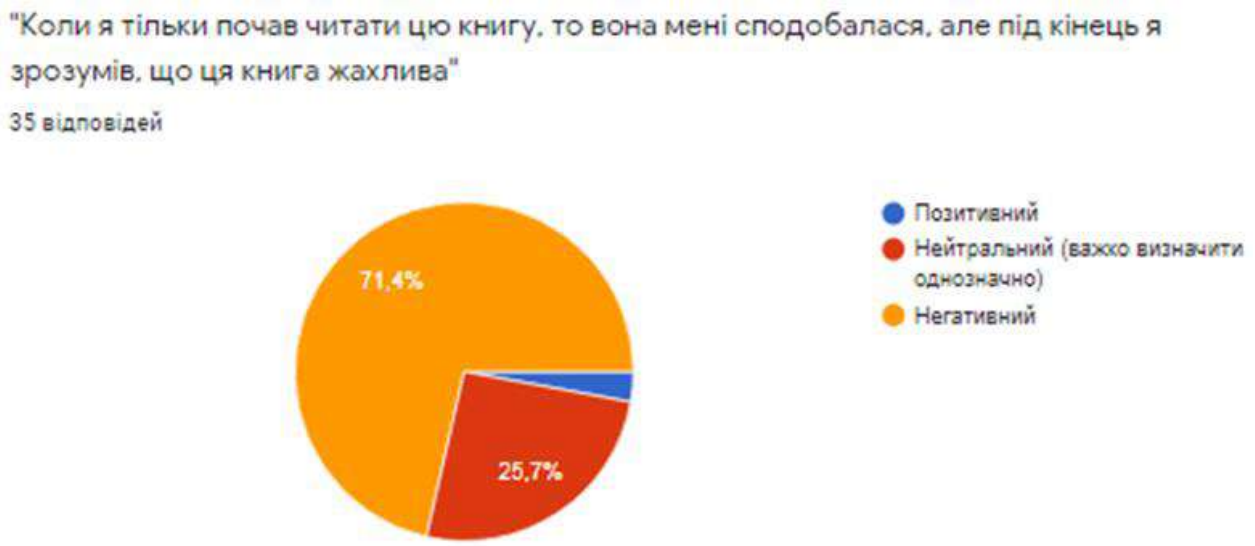

Рис. 9. Результати оцінювання респондентами речення *"Коли я тільки почав читати цю книгу, то вона мені сподобалася, але під кінець я зрозумів, що ця книга жахлива"*

На такому прикладі добре помітно те, що люди вміють надавати більше значення певному елементу речення, на відміну від програм, які підсумовують значення всіх забарвлених елементів [13, с. 86].

Ще одним аспектом, розглянутим в попередньому розділі, був аналіз текстів, які не мають вираженого емоційного забарвлення. Розгляньмо приклад з попереднього розділу: *“Ця книга дуже своєрідна”*. Комп’ютерна програма не виявила в тексті жодного емоційно забарвленого елемента, тому надала йому нейтральну оцінку (рис. 10).

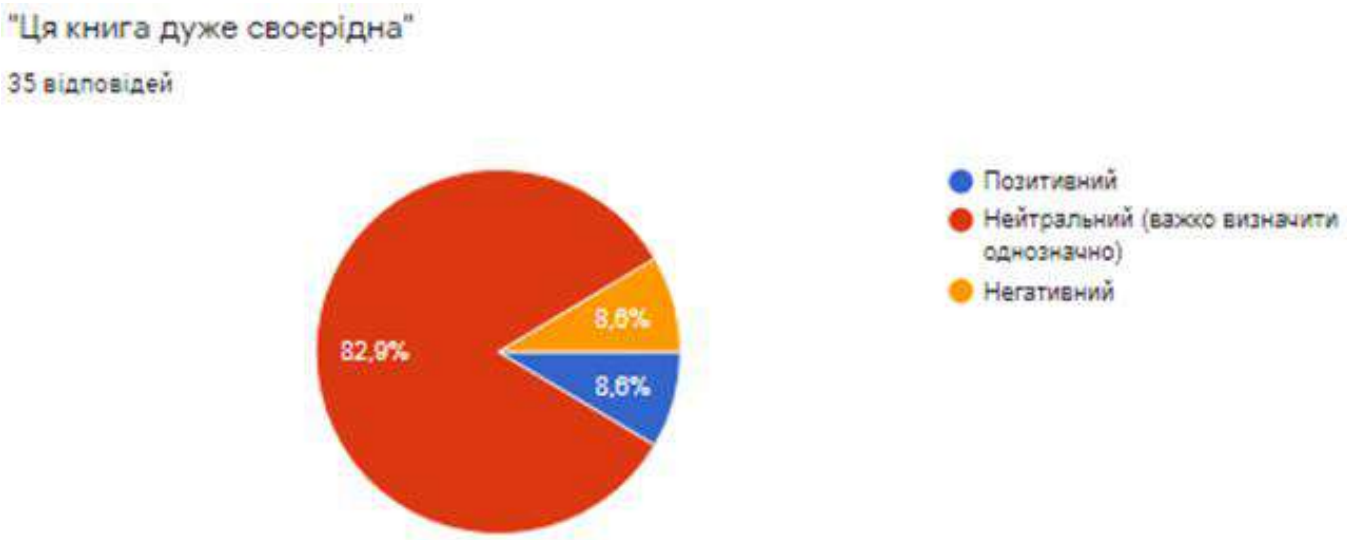

Рис. 10. Результати оцінювання респондентами речення *“Ця книга дуже своєрідна”*

Абсолютна більшість респондентів визначила цей текст як нейтральний або такий, забарвлення якого важко визначити. Невелика частина опитаних визначила забарвлення як позитивне, так і негативне, у рівній пропорції. Сприйняття речень такого типу є суб’єктивним, оскільки така риса, як своєрідність може бути оцінена і позитино (щось виняткове, цікаве, унікальне), і негативно (ймовірно, як евфемізм до слів дивний, дивакуватий), що залежить від психологічних характеристик респондента.

3.2.Оцінка ненормативних мовних форм

Комп’ютерні програми не здатні виявити емоційно забарвлені слова, якщо вони написані з помилками або з використанням некоректних символів. Варто відзначити, що оцінка таких текстів людьми є доволі своєрідною.

Перше речення, яке містило офографічну помилку, таке: *"Ця книга просто прикрасна"*. Реакція респондентів була неоднозначною (рис. 11).

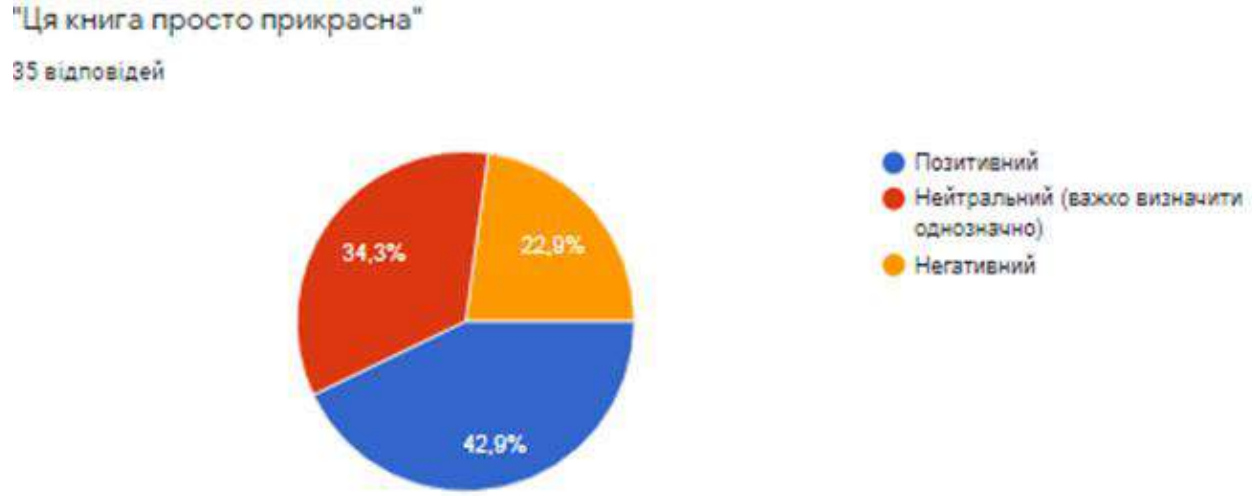

Рис. 11. Результати оцінювання респондентами речення *"Ця книга просто прикрасна"*

Трохи менше половини опитаних визначили речення як позитивне, трохи більше третини як нейтральне або таке, тональність якого важко визначити, решта як негативне. Важливо звернути увагу на те, що саме орфографічна помилка могла вплинути на оцінку тональності як негативну.

Використання некоректних символів. Половина респондентів визначили речення *"Дуже цыкава книга, найцыкавыша книга, яку я читав"* як позитивне, третина як нейтральне, або таке, яке не можна однозначно визначити, решта як негативне (додаток 12).

Рис. 12. Результати оцінювання респондентами речення *"Дуже цыкава книга, найцыкавыша книга, яку я читав"*

Люди можуть суб'єктивно сприймати речення з такими помилками, і часом важко передбачити, до якої оцінки тексту такі помилки приведуть. Іншими словами, такі тексти можуть відштовхувати людину, для прикладу, речення з великою кількістю помилок може викликати негативні емоції в людини, адже може асоціюватися з неосвіченістю автора коментаря [16, с 427].

Отже, протестовано на сконструйованих прикладах ефективність роботи спеціалізованої програми Text Analysys API. Результати «машинного» аналізу порівняно з думкою респондентів. Програмне забезпечення аналізує тексти як сукупність окремих слів, тоді як люди сприймають речення цілком, залучаючи до аналізу не лише значення слів, але й знання. Варто відзначити те, що сприйняття тексту людиною є доволі суб'єктивним і залежить від великої кількості найрізноманітніших чинників (на сприйняття може впливати досвід людини, який залежить від культурного середовища у якому сформувався респондент; навіть настрій). Як було показано, сарказм може трактуватися зовсім по-різному різними людьми. Сконструйований коментар-відгук із сарказмом був неоднозначно сприйнятий частиною респондентів, хоча загалом

люди, на відміну від спеціалізованих програм, вміють виявляти сарказм у тексті. Під час експерименту з’ясовано, що орфографічні помилки можуть впливати на оцінювання тексту людиною. Люди вміють визначати забарвлення тексту, що складається з різних елементів, на відміну від програм, які підсумовують числові значення емоційно забарвлених елементів тексту.

Висновки

Сентимент-аналіз є оцінюванням тональності заданого тексту за допомогою методів опрацювання природньої мови, статистики та машинного навчання. Сентимент-аналіз є розділом прикладної лінгвістики, пов’язаним з проблемою машинного опрацювання природньої мови.

Проблема аналізу тональності якнайкраще відображає комплексність прикладної лінгвістики, та її зв’язок з іншими дисциплінами.

На сьогоднішній день чітко помітна важливість розвитку технологій сентимент-аналізу, з огляду на велетенську кількість інформації у вільному доступі, що постійно зростає. Розвиток технологій сентимент-аналізу є важливий для прикладної лінгвістики як науки, для досліджень штучного інтелекту та має неабияку практичну цінність для підприємців, журналістів та інших категорій людей. Проте на цьому етапі розвитку штучного інтелекту аналіз людських емоцій, які і є об’єктом аналізу тональності, є складним завданням для нього.

Зараз нам відомо чимало різних алгоритмів, за якими працює спеціальне програмне забезпечення, що проводить сентимент-аналіз тексту, та різні варіації групування та класифікації таких алгоритмів і методів.

Найпоширенішими методами, які використовує відповідне програмне забезпечення є: використання правил із наперед заготовлених шаблонів (rule-based with patterns), машинне навчання без вчителя (unsupervised learning), машинне навчання з учителем (supervised learning), гібридний метод (hybrid method) та метод, заснований на теоретико-графових моделях.

Традиційно прийнято вважати, що завданням сентимент-аналізу є класифікація тексту на позитивний, найтральний та негативний, або просто на позитивний та негативний. Тональність визначається трьома компонентами: суб'єктом тональності, об'єктом тональності, і власне тональною оцінкою.

Методи, що використовуються тут, також можна класифікувати таким чином: ті, що використовують наївний Баєсовий класифікатор, що використовують ідею ентропії, та такі, що використовують поняття опорних

векторів. Важливими є також характеристики, що оцінюють результат аналізу, такі як точність та повнота, які визначаються за спеціальними формулами.

Для вирішення поданих проблем було проведено аналіз штучно сконструйованих речень за допомогою спеціалізованої програми та були проаналізовані результати. Обов'язково необхідно розуміти, що опрацювання текстів тільки однією програмою мало на меті показати стандарт роботи та найпоширеніші випадки і проблеми.

Алгоритми, на яких ґрунтуються такі програми зазвичай не розпізнають сарказм, непогано розпізнають речення з одним забарвленим елементом. Особливої уваги потребує аналіз текст з наявністю ненормативних мовних висловів, орфографічних помилок чи некоректних символів у словах. Алгоритми програм не справляються із такими текстами належно. Отже, бази даних комп'ютерних програм сентимент-аналізу варто було би доповнити інформацією, яка стосується типових орфографічних помилок та одруківок; даними про евфемізацію написання пейоративної лексики, яку часто застосовують користувачі. Окрім того, важливим завданням сучасної прикладної лінгвістики є створення методик автоматичного семантичного аналізу, що дало б змогу комп'ютерним програмам «розуміти» образні вислови.

Було проведене опитування респондентів, більшість яких є студентами. Результати опитування зіставлено з результатами машинного аналізу. Загалом більшість оцінок людей відповідали задуму сконструйованих коментарів, попри особливості цих текстів, які перешкоджали коректному трактуванню програмами. Проте більшість опитаних доволі неоднозначно оцінила приклад із орфографічною помилкою та приклад з використання некоректних символів.

Отже, сентимент-аналіз є аспектом прикладної лінгвістики, який заслуговує на вивчення та розвиток з огляду на свою важливість для науки та суспільства. Методи, які лежать в основі алгоритмів роботи відповідних програм

треба вдосконалювати в напрямку покращеного розуміння особливостей людських думок та емоцій, виражених в текстах.

Список використаних джерел

1. Гаспаров Б. М. Язык, память, образ. Лингвистика языкового существования // М.: «Новое литературное обозрение», 1996. 352 с.

2. Якобсон Р. О. Лингвистика и поэтика // Структурализм: «за» и «против». М.:Прогресс, 1975. 473 с.

3. Bo Pang, Lillian Lee, Shivakumar Vaithyanathan Thumbs up? Sentiment Classification using Machine Learning Techniques // Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002). 2002. Pp. 79–86. Режим доступу: https://www.aclweb.org/anthology/W02-1011

4. Brants, Th. TnT — A Statistical Part-of-Speech Tagger // In Proceedings of the sixth conference on Applied natural language processing. Seattle, WA, 2000. Pp. 224-231.

5. Dictionary by Merriam-Webster. Режим доступу: merriam-webster.com

6. Spertus, Ellen. Smokey: Automatic recognition of hostile messages. // Proc. of Innovative Applications of Artificial Intelligence (IAAI), 1997. Pp. 1058–1065.

7. Mosteller, Frederick, Wallace, David L. Applied Bayesian and Classical Inference: The Case of the Federalist Papers // Springer Science & Business Media, 1984. Pp. 45-47.

8. García-Moya, L., Anaya-Sanchez, H., Berlanga-Llavori, R. Retrieving product features and opinions from customer reviews / IEEE Intelligent Systems 28(3), 2013. Pp. 19 – 27.

9. Toutanova K., Manning, C. D. Enriching the Knowledge Sources Used in a Maximum Entropy Part-of-Speech Tagger. // Proceedings of the Joint SIGDAT / Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-2000). Pp. 63–70.

10. Crammer, K., Dekel, O., Keshet, J., Shalev-Shwartz, S., & Singer, Y. Online passive-aggressive algorithms // Journal of Machine Learning Research, 7(Mar), 2006. Pp. 551-585.

11. Nasukawa, T., Yi, J. Sentiment analysis: capturing favorability using natural language processing // Proceedings of the 2nd international conference on Knowledge capture, Florida, USA, October 23–25, 2003. Pp. 70–77.

12. Pang, B., Lee, L. Opinion Mining and Sentiment Analysis // Foundations and Trends in Information Retrieval, v.2 n. 1–2, 2008. Pp. 1–135.

13. Pang, B., Lee, L., Vaithyanathan, S. Thumbs up? Sentiment classifi cation using machine learning techniques // Proceeding of the conference on empirical methods in natural language processing (EMNLP 2002), 2002. Pp. 79 – 86.

14. Chetviorkin, I., P. Braslavskiy, and N. Loukachevitch. "Sentiment Analysis Track // Компьютерная лингвистика и интеллектуальные технологии. «Диалог-2013». Сб. науч. ст., т. 2, 2013. С. 40 - 50.

15. Singla, P., Domingos, P. Discriminative training of Markov logic networks // Proceedings of the Twentieth National Conference on Artificial Intelligence. Pittsburgh, 2005. Pp. 868–873.

16. Yi, J., Nasukawa, T., Niblack, W., Bunescu, R. Sentiment analyzer: extracting sentiments about a given topic using natural language processing techniques // Proceedings of the 3rd IEEE international conference on data mining (ICDM 2003), 2003. Pp. 427–434.

АНОТАЦІЯ

Актуальність праці полягає у практичній цінності технологій сентимент-аналізу, оскільки він має широку сферу застосування. Мета цієї праці – проаналізувати алгоритми роботи програмного забезпечення, використовуваного для сентимент-аналізу, розглянути особливостей, пов’язаних із сентимент-аналізом за відповідними методами, визначити ефективність роботи таких методів за допомогою порівняння результатів з результатами оцінювання людиною; дати рекомендації для підвищення ефективності роботи програм. Для досягнення мети поставлено низку завдань: надати загальну характеристику сутності поняття сентимент-аналіз (аналіз тональності); розглянути алгоритми, що найчастіше використовуються спеціалізованими програмами для сентимент-аналізу; розглянути специфіку оцінки тональності тексту людьми; порівняти результати та методи сентимент-аналізу, здійсненого програмним забезпеченням, та людською оцінкою.

Для вирішення поставлених завдань використано такі методи, як індукція та дедукція, описовий метод, метод компонентного аналізу, експеримент.

Наукова новизна цієї полягає в порівнянні результатів оцінки тональності, наданої спеціалізованим програмним забезпеченням, та результатів оцінки тональності, що були надані людьми, а також порівняння методів, що використано для створення програмного забезпечення, та методів, що використовує людина для вирішення поставленого завдання.

Праця складається зі вступу, трьох розділів, , висновків та списку використаних джерел, а також додатків.