Технічні науки

УДК 004.912

Орел Антон Валерійович
аспірант кафедри автоматизованих систем
обробки інформації та управління
Факультету інформатики та обчислювальної техніки
Національного технічного університету України
«Київський політехнічний інститут імені Ігоря Сікорського»
Орел Антон Валерьевич
аспирант кафедры автоматизированных систем
обработки информации и управления
Факультета информатики и вычислительной техники
Национального технического университета Украины
«Киевский политехнический институт имени Игоря Сикорского»
Orel Anton
PhD Student of Department of Computer-Aided
Management and Data Processing Systems of the
Faculty of Informatics and Computer Science of the
National Technical University of Ukraine
"Igor Sikorsky Kyiv Polytechnic Institute"

АНАЛІЗ СОЦІАЛЬНИХ МЕДІА ДЛЯ ВИЗНАЧЕННЯ ОЦІНОЧНИХ СУДЖЕНЬ НА ОСНОВІ АНАЛІЗУ ТОНАЛЬНОСТІ ТЕКСТУ

АНАЛИЗ СОЦИАЛЬНЫХ МЕДИА ДЛЯ ОПРЕДЕЛЕНИЯ ОЦЕНОЧНЫХ МНЕНИЙ НА ОСНОВЕ АНАЛИЗА ТОНАЛЬНОСТИ ТЕКСТА

SOCIAL MEDIA ANALYZING FOR EVALUATION OPINIONS DETERMINATION BASED ON SENTIMENT ANALYSIS

Анотація. *Стаття присвячена питанням розробки системи збору та аналізу думок про досвід користувачів товарів на основі аналізу тональності тексту. В роботі розглянуті прикладні задачі текстового аналізу та сентимент аналізу Facebook-коментарів та аналіз тональності текстів, отриманих із Twitter. Описана бізнес-логіка процесу обробки оціночних суджень та окремі стадії цього процесу. Поданий опис методу розробки лексикону для моделювання класифікатора машинного навчання при прогнозуванні настроїв користувачів щодо товарів. Приведені результати дослідження аналізу тональності відгуків про косметичні марки та косметичні товари.*

Ключові слова: *тональність тексту, видобуток тексту, аналіз оціночних суджень, соціальні мережі, машинне навчання.*

Аннотация. *Статья посвящена вопросам разработки системы сбора и анализа мнений об опыте пользователей товаров на основе анализа тональности текста. В работе рассмотрены прикладные задачи текстового анализа и сентимент анализа Facebook-комментариев и анализ тональности текстов, полученных из Twitter. Описана бизнес-логика процесса обработки оценочных мнений и отдельные стадии этого процесса. Поданное описание метода разработки лексикона для моделирования классификатора машинного обучения при прогнозировании настроений пользователей относительно товаров. Приведенные результаты исследования анализа тональности отзывов про косметологические марки и товары.*

Ключевые слова: *тональность текста, извлечение текста, анализ оценочных мнений, социальные сети, машинное обучение.*

Summary. *The article is devoted to the development of a system for collecting and analyzing opinions on the experience of users of goods based on the sentiment analysis.*

Key words: *sentiment analysis, text mining, evaluation opinions, social media, machine learning.*

Вступ. Зростаючі випадки контрафакції та пов'язані з ними негативні наслідки для економіки та здоров'я потребують розробки систем активного нагляду, здатних надавати своєчасну і достовірну інформацію для всіх учасників боротьби з контрафакцією. Однією із стратегій боротьби з контрафактною продукцією є ефективна комунікація та відстеження попереджень щодо вад продукції, побічні чи несприятливі ефекти тощо. Соціальні медіа тепер є платформою для обміну практично всіма видами інформації. Дослідження даних Twitter свідчать про те, що об'єднання мільйонів повідомлень може дати корисну інформацію для населення [2], отже, виробники продукції повинні відслідковувати думку користувачів про свою продукцію для прийняття рішень [3]. Користувачі, перш, ніж вибрати товар або компанію для регулярного обслуговування, шукають відгуки та рекомендації. Вони допомагають виявити недоліки продукту або сервісу, і своєчасно їх усунути. Відгуки в соціальних мережах в більшості своїй незалежні і відверті, тому викликають довіру. Такі оціночні судження можуть, як згенерувати цільовий трафік, так і зруйнувати репутацію компанії. Не всі клієнти готові прямо висловити своє невдоволення компанії і пояснити, чому вирішили піти. Але в Інтернеті така інформація часто буває доступною мільйонам користувачів.

Постановка проблеми. Для усунення загрози підробки товарів необхідна система моніторингу, здатна використовувати та відстежувати

онлайн відгуки та досвід застосування користувачами цих товарів. У статті запропоновані задачі дослідження для визначення:

- громадських настроїв щодо певної марки товарів;
- прогнозу появи небажаних наслідків та можливості підробки товарів через моніторинг відгуків в соціальних мережах як засобу контролю думки громадськості.

Підхід до цих проблем оснований на класі методів контент-аналізу в комп'ютерній лінгвістиці, призначений для автоматизованого виявлення в текстах емоційно забарвленої лексики та емоційної оцінки авторів, тобто думок по відношенню до об'єктів, мова про які йде в тексті.

Аналіз останніх досліджень та публікацій. Видобуток тексту (Text Mining) є спеціалізованою галуззю, що застосовує методи видобування даних у тексті. Деякі спроби аналізу досліджуваних текстових даних наведені у [4]. Аналіз тональності тексту, що спрямований на виявлення думок, настроїв та ставлення людей та громад розглядають автори [5]. Коли методика аналізу тексту та аналізу тональності текстів поєднуються в проекті з даними соціальних мереж, результат часто є потужним описовим або попереджувальним інструментом. В роботі [6] видобуток тексту успішно застосовується для збору повідомлень у Facebook для класифікації тональності тексту під час Арабської весни (серія масових вуличних протестів, революції та внутрішніх військових конфліктів у низці арабських країн, що почалися наприкінці 2010 року у Тунісі й тривають у деяких країнах до нині).

Автори [7] розглядають застосування аналізу тональності тексту для спостереження за злочинністю, зокрема проаналізовані соціальні мережі та запропоновані основні елементи для системи їх моніторингу. Авторами [8] було запропоновано структуру для вивчення реакцій, настроїв та комунікації цивільних осіб у відповідь на терористичні напади. В роботі [9] запропоновані

обчислювальні способи оцінки тональності у соціальних мережах і наведені дані про кращу їх продуктивність у порівнянні зі стандартними методами.

Аналіз тональності тексту даних соціальних мереж також застосовується для відстеження спалахів захворювань. Так автори [10] описали спосіб збору твітів для виявлення раннього попередження та спалаху пандемії свинячого грипу. Розглянутий спосіб демонструє значний внесок у попередження учасників зацікавленої сторони для їх швидкого реагування.

Дані Twitter дають змогу провести географічний і просторовий аналіз, а у [11] описано структуру візуалізації варіантів настроїв громадськості з використанням твітів, зібраних з округів Великої Британії під час події 2013 року. Ця програма особливо корисна для відстеження тенденції висхідної або зниженої алергії на продукти в певному регіоні протягом певного періоду часу. Автори [12] використали технології виявлення тексту, щоб досліджувати ставлення споживачів до світових брендів застосувавши соціальну мережу Twitter.

Фармакологічний нагляд представляє особливий інтерес, він включає в себе моніторинг несприятливих наслідків фармацевтичних продуктів. Як повідомлялося у [13], користувачі Інтернету можуть надати ранні попередження про побічні ефекти за допомогою даних журналу (логів) інтернет-серфінгу. Візуалізація тональності текстів чату форуму CHFpatients.com була використана для вимірювання ефективності препарату через кількісну оцінку його побічних ефектів, особливо на користь членів форуму та їх лікарів [14].

Популярними підходами для аналізу тональності тексту є лексичний аналіз та машинне навчання. Лексичний підхід, описаний у [5], [15] та [16], використовує словники для слів, що анотовані з їх семантичними орієнтаціями. Підхід, оснований на машинному навчанні, описаний у [5] та

[17], вимагає створення моделі шляхом навчання класифікатора з позначеними прикладами. Використання комбінації лексичного аналізу та машинного навчання описано в [17] та [18].

Аналіз підходів до аналізу тональності тексту виявляє проблеми продуктивності:

- як визначати контекстно-залежні слова;
- як звертатися до кількох об'єктів з різними семантичними орієнтаціями в межах одного речення.

Одним із запропонованих підходів до вирішення цих питань є використання цілісної лексики, описаної в [15], що передбачає використання зовнішніх доказів та лінгвістичних звичаїв природних мовних виразів. Як повідомляється, підхід до машинного навчання перевершує лексичний підхід, але все ж страждає загальний недолік маркування великих навчальних даних.

Дослідники [17] класифікаторів на базі машинного навчання Naive-Bayes, максимальної ентропії та SVM (Support Vector Machines - підтримуючі векторні машини) окреслили низку питань, які необхідно вирішити перед використанням цих методів для класифікації тональності тексту. Визначені недоліки класифікаторів Naive-Bayes та максимальної ентропії. Недолік першого полягає в припущенні, що функції не залежать одна від одного; класифікатор максимальної ентропії зазнає надмірного пристосування у разі нестабільних даних. Проте повідомляється, що її продуктивність може бути покращена шляхом апріорі для кожної функції. Підтримка векторних машин перевершила інші методи, головним недоліком якої є складність визначення важливих слів, які вплинули на процес класифікації через його «чорний ящик». У огляді [17] описано метод, що називається розповсюдженням марки, який покращує точність процесу класифікації за допомогою графіка Twitter.

Спираючись на ці мотивації, особливо на успіхи аналізу тексту та аналізу тональності тексту даних соціальних мереж для спостережень [7], [8], [9], [10] та моніторингу репутації бренда [12], а також дизайну структури класифікації тональності тексту [8], [11] та [17], автори цих робіт запропонували нову основу для використання та обробки висловлювань та опису досвіду використання споживачів популярних марок лікарських та косметичних засобів, що повідомляються як оновлення статусу, твітів або коментарів на платформах соціальних мереж. Також існує теорія адаптації комбінації цих підходів та обчислювальних інтелектуальних методів, таких як нечітке оцінювання настрою, в рамках порівняння результатів з традиційною лексикою та іншими методами.

Детальний опис бізнес-логіки аналізу тональності тексту наведено у наступному розділі.

1. Бізнес-логіка процесу обробки оціночних суджень

Схема бізнес-логіки процесу обробки думок та відгуків про досвід клієнтів популярних марок лікарських та косметичних продуктів подана на рисунку 1. Схема включає чотири етапи: збору та очищення тексту, попередньої обробки, сентимент (тональність тексту) аналізу і, нарешті, оцінки. Нижче наведено короткий опис цих етапів.

Рис. 1. Схема бізнес-логіки процесу обробки думок та відгуків про досвід користувачів

Порядок збору та очищення тексту. Більшість організацій та компаній, включаючи соціальні медіа-платформи, створюють прикладний програмний інтерфейс (API) для обміну даними. На етапі збору та очищення тексту API-інтерфейси соціальних мереж здійснюють API-виклик для аутентифікації та збору даних. Розглянемо детальніше API-інтерфейси Twitter і Facebook. API Twitter складається з передачі репрезентативного стану (REST) та потокового API [19]. REST API пропонує методи перевірки аутентифікації застосувань, обробки запитів, обробки встановлених обмежень, тощо. Потокові API надають клієнтські додатки з глобальним потоком (публічним, користувацьким і сайту) даних Twitter. API Facebook Graph являє собою засоби для отримання даних у соціальному трафіку Facebook та виходу з нього. У фреймворку використовуються як REST, так і потоковий API для пошуку та отримання твітів, у той час як API Facebook Graph використовується для отримання сторінок, оновлення статусу та коментарів,

що пропонують опис користувальницького досвіду і судження користувачів стосовно лікарських та косметичних товарів.

Рисунок 2 описує робочий процес збору даних. Зібраний текст є зашумленним в текстовому форматі обробки даних (JSON) і методах очищення та розбору даних для формування структури збірки текстів, тобто кола коментарів та твітів, що включені для подальшої обробки.

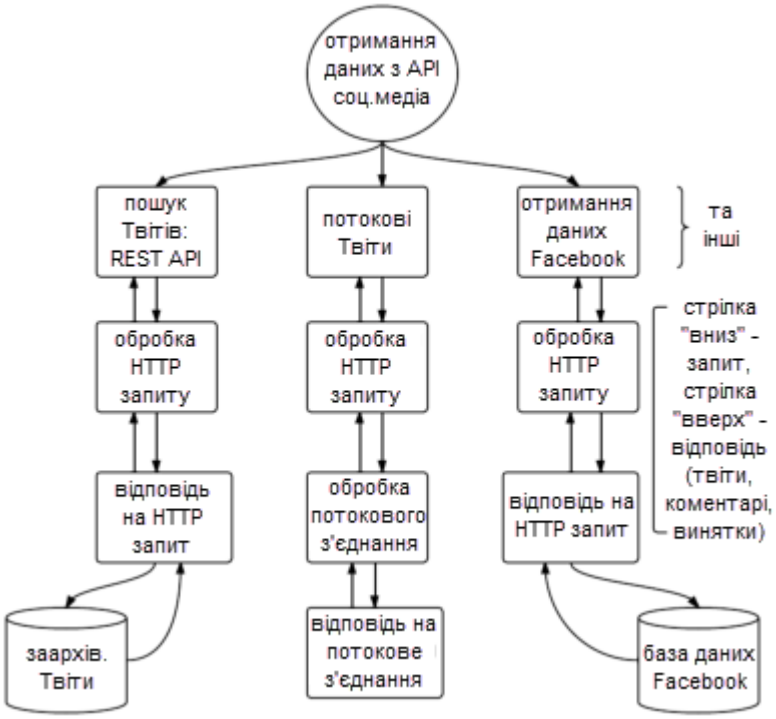

Рис. 2. Процес збору даних

Стадія попередньої обробки. На цьому етапі збірка текстів перетворюється на функціональні вектори. Відповідно до описаних у [20] та [21] функцій проведення цієї роботи адаптується «мішком слів» ("bag of words"). Наступна послідовність завдань включає адаптацію простого методу вибору, попередню обробку функції для перетворення або лексичного аналізу текстового потоку у слова, видалення обмежувачів, перетворення всіх слів на нижній регістр, видалення чисел та «стоп» слів [21], виведення слів на їхню основу та деяких перетворень певних програм або доменів. Токени

представлені як реплікарна матриця «мішка слів», використовуючи термін «частотно-зворотня частота документів» (*tf-idf*). Схема зважування, описана у [20] та [21].

Позначимо збірку текстів як C, що містить N документів, визначених як d_i, де $i = 1 \ldots N$, і твіти марковані як слова або терміни t. Схема вагової оцінки *tf-idf* враховує відносну важливість слова у документі та присвоює терміну t_j, вага у документі d_i дається:

$$tf - idf(t_j, d_i) = tf(t_j, d_i) * idf(t_j)$$

де $tf(t_j, d_i)$ означає частоту терміну, кількість введених слів у документі;

$idf(t_j) = log2 \frac{N}{df t_j}$ означає частоту зворотного документа, причому $df(t_j)$представляє кількість документів, що містять слово.

На рисунку 3 описаний процес попередньої обробки.

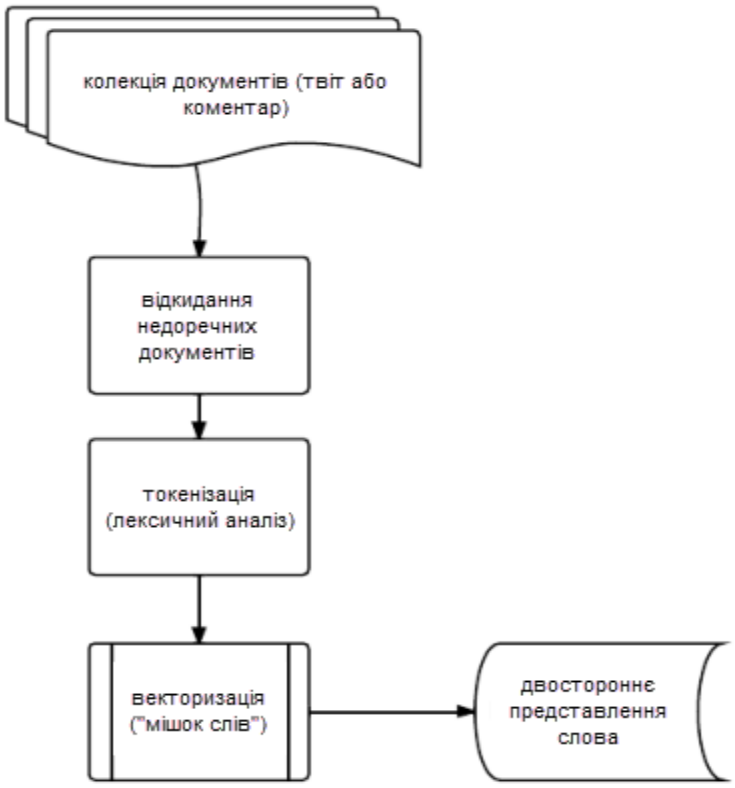

Рис. 3. Процес попередньої обробки

Подальша розрідженність оброблюється шляхом вибору термінів, які відображаються в мінімальній кількості документів. Результуюче двостороннє представництво є входом, на якому виконуються інші завдання.

Стадія сентимент аналізу (аналізу тональності тексту). На цьому етапі розглядаються вимірювання полярності, сентимент класифікація та кластеризація всієї збірки текстів, а також для деяких цільових об'єктів. Ми підходимо до цих завдань, використовуючи як лексику, так і методи навчання.

Розглянемо класифікацію тональності тексту на основі лексики. Під лексиконізованим підходом розуміється попередньо позначений список слів або полярність лексики. Для покращеного класифікаційного результату, описаного в [22], фреймворк об'єднує дві лексики (спеціально зібрану лексику та загальну англомовну лексику), розроблену та підтримувану авторами [23]. Інша вимога до класифікатора на основі лексики - це реалізація функції оцінювання тональності тексту (рис. 4). Одна з найпростіших обчислювальних схем поляризації описана у [24] і передбачає, що усі слова у збірці текстів або цільовій колекції порівнюються зі словами у лексиконі, а загальна оцінка тональності тексту або підмножини буде тоді різницею між кількістю позитивно і негативно позначених слів. Тому відповідна оцінка полярності для кожного коментаря чи твіту у збірці текстів визначається:

$$Score = \sum_i^n pw - \sum_j^m nw$$

де pw і nw – це позитиво та негативно позначені слова відповідно;

Коментар чи твіт має загальний позитивний сентимент, якщо Score > 0. Коментар чи твіт має загальний нейтральний сентимент, якщо Score = 0. Коментар чи твіт має загальний негативний сентимент, якщо Score < 0,

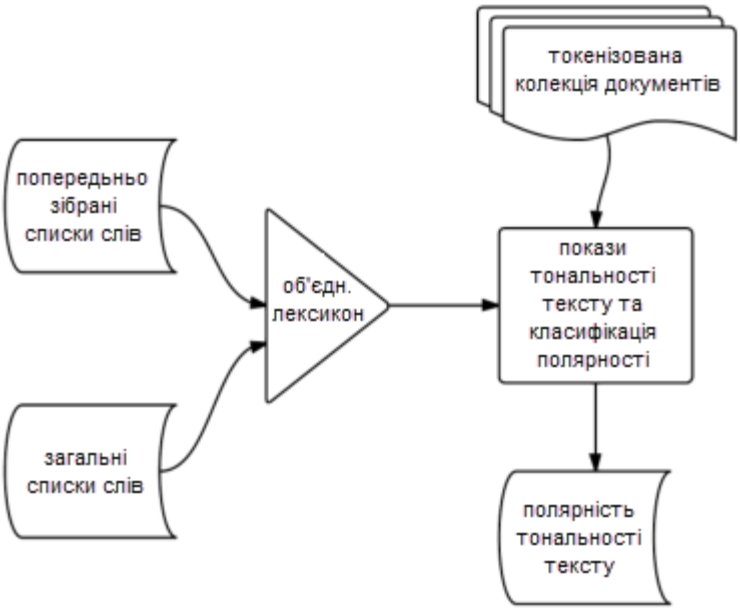

Рис. 4. Процес лексикону на основі класифікації сентименту

Загальна оцінка для процесу візуалізується та оцінюється за допомогою гістограми та статистичного графіку (діаграми). Більш прогресивна оціночна схема включає нечіткі міркування, що є методом обчислювального інтелекту, який може застосовуватися для покращення класифікації тексту та завдань кластеризації. Ця методика була використана у [25] для створення інтуїтивно зрозумілих нечітких чисел для 150 слів, сформованих зі списку функціональних слів для класифікації відгуків по керуванню готельними системами, з більшою точністю та швидкістю нагадування.

Особливістю класифікації сентиментів (тональності тексту) на основі машинного навчання є використання навчального набору даних, який вже закодований класами тональності текстів. Класифікатор навчається або моделюється маркованими даними, таким чином, що нові, але аналогічні документи тестуються з отриманою моделлю аби передбачати напрямок сентименту (тональності тексту) нових документів (рис. 5).

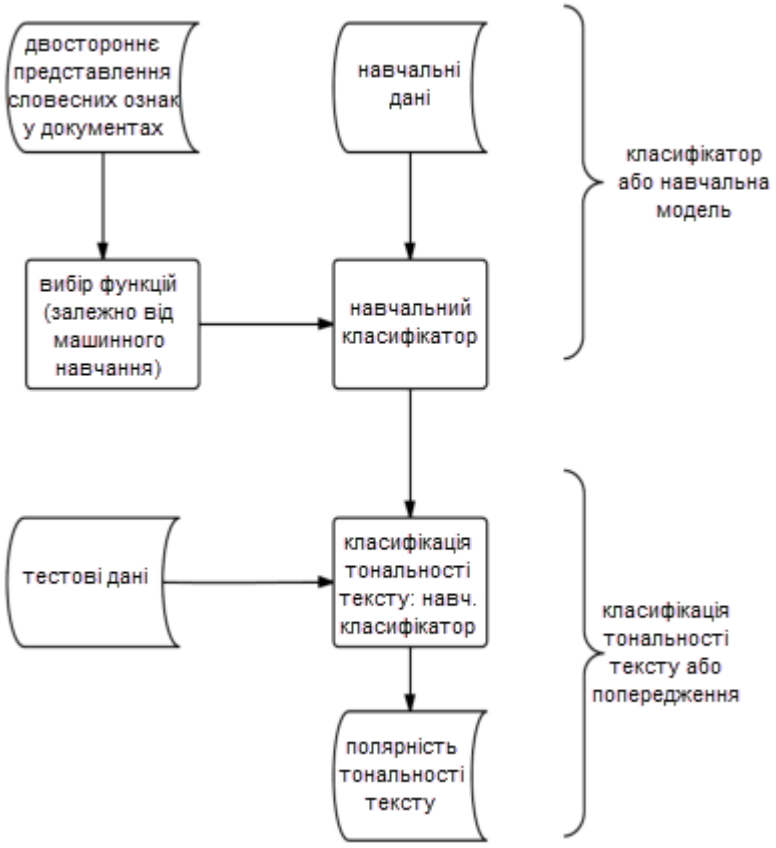

Рис. 5. Процес класифікації

В якості базового класифікатора використовуватиме Naive-Bayes через його ефективність згідно з [21]. Припускаємо, що майбутні слова є незалежними, а потім кожне входження використовується для класифікації твітів або коментарів у відповідний клас сентиментів. Такий підхід називається мультиноміальною моделлю подій.

З [21] випливає, що цей класифікатор, який використовує максимальне правило апостеріорного рішення, може бути представлений таким чином:

$$c_{map} = \text{argmax}_{c \in C}(P(c|d)) = \text{argmax}_{c \in C}(P(c) \prod_{1 \le k \le n_d} P(t_k|c)$$

де t_k позначає слова в кожному твіті або коментарі і C набір класів, що використовуються в класифікації; $P(c|d)$ є умовною ймовірністю класу C даного документу d, $P(c)$ - попередня ймовірність класу C і $P(t_k|c)$ є умовною

ймовірністю слова t_k даного класу *C*. Щоб оцінити попередні параметри, це рівняння потім зводиться до:

$$c_{map} = \text{argmax}_{c \in C} \left(logP(c) + \sum_{1 \le k \le n_d} logP(t_k|c) \right.$$

Для обробки нульових імовірностей, які можуть виникнути, коли слово не зустрічається в певному класі, використовується *(tf-idf)* зважування або згладжування Лапласа шляхом додавання 1 до кожного підрахунку; із згладжуванням Лапласа, *P(t|c)* стає:

$$P(t|c) = \frac{T_{ct}+1}{\sum_{t' \in V}(T_{ct'}+1} + \frac{T_{ct}+1}{(\sum_{t' \in V} T_{ct'})+B'}$$

де *B'* означає кількість термінів, що містяться у словнику *V*.

Оціночна стадія. Результат класифікації на основі лексикону відображається як гістограма ступенів полярності, і результат оцінюється з посиланням на реальність або людське судження. Ми використовуємо таблиці непередбачених ситуацій або реальну таблицю, щоб відобразити вихід класифікатора та вихідний результат для порівняння продуктивності.

2. Тематичні дослідження

Текстовий аналіз та сентимент аналіз Facebook коментарів. Щоб продемонструвати корисність запропонованої схеми, автори [22] зібрали коментарі та думки користувачів зі сторінок 3-х популярних брендів лікарських та косметичних продуктів (*Avon, Dove* та *OralB*) з використанням Facebook API Graph. Для питань конфіденційності назви торгових марок будуть випадковим чином кодуватися як А, Б та В. Спочатку був видалений загальнодоступний вміст цільових сторінок, а потім вилучалися коментарі та відгуки користувачів із популярних публікацій, що пропонують рекламу продукту. Результати подані в таблиці 1

Таблиця 1

Резюме зібраних коментарів

Ім'я бренду	Загальна кількість зібраних постів	Загальна кількість вилучених коментарів
А	5000	654
Б	665	1747
В	5000	957

У всій збірці застосовано токенізацію (лексичний аналіз), призупиняємо видалення слів і перетворюємо у функції нижнього регістру, щоб отримати функції лише окремих слів. Потім функції були представлені у вигляді моделі «мішку слів» зі схемою зважування *(tf-idf)* для створення розрідженої матриці, обмежуючи довжини слова до трьох символів. Розріджена матриця оброблялася за допомогою функції, яка видаляє розріджені терміни, що мають принаймні 99% розріджених елементів. Потім виконується аналіз термінів для генерування популярних слів із попередньо зібраного лексикону, а також був адаптований метод, який використовується у [22], спираючись на англійський словник та тезаурус. Заздалегідь зібраний лексикон зливається із сленгами соціальних медіа та загальною лексикою для сентимент класифікації. Порівняння аналізу сентимент лексикону було здійснено на трьох різних брендах.

Цікаво побачити, що загальний сентимент на всіх трьох брендах був позитивно асиметричним із коефіцієнтами 1:42:175, 1:3:3 та 1:5:5 брендів А, Б та В відповідно, де цифри означають негативні, нейтральні та позитивні значення коефіцієнтів. Розподіл цих оцінок поданий у таблиці 2.

Таблиця 2

Розподілення результатів сентименту для 3-х брендів

Бренд	Сентимент оцінки			
	Негативний	Нейтральний	Позитивний	Загальна кількість у рядку
А	3	127	524	654
Б	282	742	723	1747
В	85	408	464	957
Загально	370	1277	1711	3358

Також опрацьовані 3 окремі набори даних у Бренді Б, кожен із трьох генеричних (непатентованих) товарів: мило, крем та дезодорант, а потім проведено аналіз тональності за сукупними даними. Розподіл цих балів представлений у таблиці 3. За результатами аналізу оцінки думок користувачів щодо крему, дезодоранту та мила виражені у співвідношенні 1:2:2, 1:1:2 і 1:2:5 відповідно, де співвідношення позначає негативне, нейтральне та позитивне сприйняття товару користувачами.

Таблиця 3

Розподілення результатів сентименту для 3-х продуктів

Продукт	Сентимент оцінки			
	Негативний	Нейтральний	Позитивний	Загальна кількість у рядку
Крем	5	11	11	27
Дезодорант	21	23	46	90
Мило	11	26	56	93
Загально	37	60	113	210

Розглянемо аналіз тональності всієї збірки текстів із класифікатором тональності Naive-Bayes, описаним у [26], моделюючи полярність та емоційну лексику. Класифікація методу полярності класифікує коментарі як позитивні, нейтральні або негативні. Результати класифікації по усій збірці текстів для обох методів порівнюються в таблиці 4.

Таблиця 4

Порівняння оцінок тональності лексикону та машинного навчання на основі коментарів збірки текстів

Метод	Сентимент оцінки			
	Негативний	Нейтральний	Позитивний	Загальна кількість у рядку
Лексикон	539	1436	1383	3358
Naïve Bayes	554	368	2436	3358

Негативні оцінки для обох методів збігаються, хоча різко відрізняються як нейтральні, так і позитивні оцінки. Оцінка того, який метод забезпечує точний результат, залежить від декількох факторів і не входить в обсяг цієї роботи.

Аналіз тональності текстів, отриманих із Twitter. Для прогнозування тональності тексту методом машинного навчання, що класифікує дані з соціальних мереж, було зібрано близько 11431 твітів з наступними ключовими словами: медицина, рецепт, без рецепту, побічні ефекти, Інтернет-аптека, антибіотики. Спочатку твіти були очищені, а потім перетворені у єдину збірку текстів, при цьому кожен твіт був представлений як єдиний документ. Таким чином, є 11413 документів, що складають збірку текстів. Збірка текстів потім представлена як розріджена матриця для постійного аналізу термінів та генерації лексикону.

Проведений аналіз тональності лексикону по усій збірці 11413 твітів, таким чином, що отримані сентимент-оцінки класифіковані за відповідними оцінками (табл. 5), і служитиме початковим набором даних. Для простоти класифікуємо оцінки більше за 0, як позитивні, а оцінки нижчі за 0 – як негативні.

Таблиця 5

Випадковий перегляд оцінок полярності

Оцінка	Твіт
-1	Для всіх, хто знає мене, гадаю, що час мені підписати документи аби відмовитись від антибіотиків. Це дозволить померти мені в моєму домі.
2	Впевнений, що ми відвідали лікарню недарма і що антибіотики допоможуть якомога швидше.
0	Приймав свої антибіотики без їжі, і я перебував на роботі хворим.
-2	Це смішно, як часто я хворію, незалежно від того, чи в мене простуда, чи я на антибіотиках, моя імунна система - сміття

Набір даних був розділений на навчальний набір, який склад 75%, тобто 8573 позначених твітів, та на тестовий набір, що склав 25%, тобто 2858 позначених твітів, так що класифікатор може бути оцінений на даних, які він раніше не бачив. Застосовуємо алгоритм Naive Bayes як базовий класифікатор. Таблиця 6 – це подання результату як матриці помилок.

Таблиця 6

Результат класифікатору Naive Bayes

Передбачено	Фактично		
	Негативний	Позитивний	Загальна кількість у рядку
Негативний	531 0,639	180 0,089	711
Позитивний	300 0,361	1847 0,911	2147
Загально	831 0,291	2027 0,709	2858

З результатів таблиці 6 видно, що 300 із 831 позитивних повідомлень (36%) були невірно класифіковані як «негативні», тоді як 180 із 2027 негативних повідомлень (8,9%) були невірно класифіковані як «позитивні»;

загальна точність складає 83%. Ця продуктивність буде використовуватися як базис для оцінки інших класифікаторів.

Висновки. Моніторинг відгуків в соціальних мережах як контроль думки громадськості є актуальною задачею для компаній, що постачають на ринок різні товари, зокрема лікарські та косметичні засоби. Ця задача зведена до задачі аналізу тональності тексту. Для визначення тональності тексту, що вказує на думки та досвід використання товарів, запропонований метод машинного навчання. Для роботи з відгуками користувачів в соціальних мережах розглянуто сервіси, проведено дослідження аналізу тональності відгуків про три косметичні марки (А, Б та В), а також про три косметичні товари (мило, крем і дезодорант).

На прикладі аналізу поглядів і опису досвіду використання косметичних і лікарських препаратів у формі твітів та коментарів у Facebook за допомогою підходів лексикону та машинного навчання проведені дослідження. Продемонстровано, як розробляти власні лексикони та навчальні дані, а також моделювали класифікатор Naive Bayes для прогнозування поглядів та досвіду користувачів на прикладі популярних марок лікарських та косметичних продуктів.

У майбутньому робота розглядатиме спам коментування, порівняння різних характеристик класифікації тональності тексту машинного навчання, часовий аналіз для виявлення тенденцій тональності тексту певної марки чи продукту, а також групування твітів та настроїв користувачів за місцем розташування.

Література

1. K. Dégardina, Y. Roggoaand and P. Margot, "Understanding and fighting the medicine counterfeit market," Journal of Pharmaceutical and Biomedical Analysis, vol. 87, pp. 167-175, January 2014.

2. M.J. Paul and M. Dredze, "You are what you Tweet: Analyzing Twitter for Public Health," in 5th International AAAI Conference on Weblogs and Social Media (ICWSM 2011), Barcelona, 2011.
3. C. Kaiser and F. Bodendorf, "Mining Patient Experiences on Web 2.0 - A Case Study in the Pharmaceutical Industry," in SRII Global Conference (SRII), California, 2012, pp. 139-145.
4. A. M. Hearst, "Untangling text data mining," in Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics, Maryland, 1999, pp. 3-10.
5. B. Pang and L. Lee, "Opinion Mining and Sentiment Analysis," Foundations and Trends in Information Retrieval, vol. 2, no. 1-2, pp. 1-135, January 2008.
6. J. Akaichi, Z. Dhouioui, and M.J. Lopez-Huertas Perez, "Text mining facebook status updates for sentiment classification," in System Theory, Control and Computing (ICSTCC), 2013 17th International Conference, Sinaia, 2013, pp. 640-645.
7. M.D. Sykora, T.W. Jackson, A. O'Brien, and S. Elayan, "National Security and Social Media Monitoring: A Presentation of the EMOTIVE and Related Systems," in 2013 European Intelligence and Security Informatics Conference (EISIC), Uppsala, 2013, pp. 172-175.
8. M. Marc and C. Lee. Vincent, "A microblogging-based approach to terrorism informatics: Exploration and chronicling civilian sentiment and response to terrorism events via Twitter," Information Systems Frontiers, vol. 13, no. 1, pp. 45-59, March 2011.
9. K. Glass and R. Colbaugh, "Estimating the sentiment of social media content for security informatics applications," in IEEE International Conference on Intelligence and Security Informatics (ISI), Beijing, 2011, pp. 65-70.

10. E. de Quincey and P. Kostkova, "Early Warning and Outbreak Detection Using Social Networking Websites: The Potential of Twitter," in Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering. Istanbul, Turkey: Springer Berlin Heidelberg, 2010, ch. 3, pp. 21-24.
11. V. D. Nguyen, B. Varghese, and A. Barker, "The royal birth of 2013: Analysing and visualising public sentiment in the UK using Twitter," in IEEE International Conference on Big Data, California, 2013, pp. 46-54.
12. M. Mostafa Mohamed, "More than words: Social networks' text mining for consumer brand sentiments," Expert Systems with Applications, vol.40, no. 10, pp. 4241-4251, August 2013.
13. R. W. White, N. P. Tatonetti, N. H. Shah, R. B. Altman, and E. Horvitz, "Web-scale pharmacovigilance: listening to signals from the crowd," J Am Med Inform Assoc, March 2013.
14. B. Chee, K.G. Karahalios, and B. Schatz, "Social Visualization of Health Messages," in 42nd Hawaii International Conference on System Sciences, HICSS '09. , Big Island, 2009, pp. 1-10.
15. X. Ding, B. Liu and P. S. Yu, "A holistic lexicon-based approach to opinion mining," in WSDM '08 Proceedings of the 2008 International Conference on Web Search and Data Mining, New York, 2008, pp. 231-240.
16. M. Taboada, J. Brooke, M. Tofiloski, K. Voll and M. Stede, "Lexicon-Based Methods for Sentiment Analysis," Journal of Computational Linguistics, vol. 37, no. 2, pp. 267-307, June 2011.
17. S. Bhuta, A. Doshi, U. Doshi, and M. Narvekar, "A review of techniques for sentiment analysis Of Twitter data," in International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT), Ghaziabad, 2014,

pp. 583-591.

18. F. Balage, P. Pedro and T. A. S. Pardo, "NILC_USP: A Hybrid System for Sentiment Analysis in Twitter Messages," in Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013), Georgia, 2013, pp. 568--572.
19. M. Kevin Makice, Twitter API: Up and Running: O'Reilly Media, 2009.
20. M. Shafiei, S. Wang, R. Zhang, E. Milios, B. Tang, J. Tougas, R. Spiteri, "Document Representation and Dimension Reduction for Text Clustering," in IEEE 23rd International Conference on Data Engineering Workshop, Istanbul, 2007, pp. 770-779.
21. C. D. Manning, P. Raghavan, and H. Schtze, Introduction to Information Retrieval: Cambridge University Press, 2008.
22. R. Heimann and N. Danneman, Social Media Mining with R.: PACKT Publishing, 2014.
23. M. Hu and B. Liu, "Mining and summarizing customer reviews," in KDD '04: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, Seattle, 2004, pp. 168-177.
24. S. Gaston. (2014, June) Mining Twitter with R. [Online]. https://sites.google.com/site/miningtwitter/home
25. X. Feng Li and D. Li, "Sentiment Orientation Classification of Webpage Online Commentary Based on Intuitionistic Fuzzy Reasoning," Applied Mechanics and Materials, vol. 347 - 350, pp. 2369-2374, August 2013.
26. P. Jurka Timothy, Tools for Sentiment Analysis, 2012.