УДК: 004.85:004.912

М. В. Білецька
С. О. Жуков

СИСТЕМНИЙ АНАЛІЗ МЕТОДІВ ВЕКТОРИЗАЦІЇ ДЛЯ МОДЕЛЮВАННЯ ЕМОЦІЙНОЇ ТОНАЛЬНОСТІ ТЕКСТОВИХ ВІДГУКІВ

Вінницький національний технічний університет

Анотація

Робота присвячена системному аналізу методів векторизації текстових даних для задач класифікації емоційної тональності. Проведено дослідження ефективності статистичного підходу Bag-of-Words та лінійного методу TF-IDF у поєднанні з логістичною регресією та наївним баєсівським класифікатором. На основі аналізу відгуків Amazon Books визначено вплив методів зважування ознак на точність прогнозів та швидкість навчання моделей. Розроблений підхід демонструє високу ефективність у виявленні користувацьких настроїв завдяки оптимізації розріджених матриць даних.

Ключові слова: системний аналіз, аналіз тональності, векторизація тексту, TF-IDF, Bag-of-Words, логістична регресія, машинне навчання, обробка природної мови.

Abstract

The paper is devoted to the system analysis of text data vectorization methods for sentiment classification tasks. The effectiveness of the Bag-of-Words statistical approach and the TF-IDF linear method in combination with Logistic Regression and Naive Bayes classifier was investigated. Based on the analysis of Amazon Books reviews, the influence of feature weighting methods on prediction accuracy and model training speed was determined. The developed approach demonstrates high efficiency in detecting consumer sentiments by optimizing sparse data matrices.

Keywords: system analysis, sentiment analysis, text vectorization, TF-IDF, Bag-of-Words, logistic regression, machine learning, natural language processing.

Вступ

В умовах стрімкого зростання обсягів електронної комерції задача автоматизованого аналізу користувацьких відгуків набуває критичного значення для бізнесу. Сирі текстові дані є неструктурованими, тому ключовим етапом їх обробки є векторизація – перетворення тексту на числовий формат, придатний для алгоритмів машинного навчання. Оскільки природна мова створює високорозмірні та розріджені матриці даних, виникає потреба у системному аналізі методів формування простору ознак.

Метою даної роботи є порівняльне дослідження та системний аналіз базових методів векторизації текстів (TF-IDF та Bag-of-Words) для оптимізації моделей класифікації емоційної тональності відгуків..

Методологія дослідження

Для проведення дослідження використано відкритий набір даних «Amazon Books Reviews» з платформи Kaggle, що опублікований користувачем Mohamed Bekheet [1].

Велика кількість неструктурованих відгуків на онлайн-платформах, таких як Amazon, створює складне завдання для видавництв та авторів, які потребують швидкої та точної оцінки зворотного зв’язку від читачів.

У роботі проведено порівняльний аналіз лінійних та статистичних методів класифікації емоційної тональності текстів. Повна версія програмного коду та результати обчислювальних експериментів доступні за посиланням [2]. Даний набір інтегрований із каталогом метаданих літературних творів за ідентифікатором назви книги. Початковий масив даних характеризувався значним обсягом (понад 3 млн записів) та критичним дисбалансом класів: кількість позитивних відгуків суттєво переважала кількість негативних.

Цільова мітка тональності (sentiment) була сформована шляхом бінаризації числового рейтингу користувача (*review/score*), що варіюється від 1 до 5 зірок. Відгуки з оцінками 1–2 було класифіковано як negative, а 4–5 – як positive; відгуки з нейтральною оцінкою (3 зірки) було вилучено для уникнення семантичної невизначеності. Для нівелювання ефекту «оптимістичного зміщення» моделі застосовано стратегію undersampling, у результаті якої сформовано репрезентативну збалансовану вибірку обсягом 30 000 записів (по 15 000 для кожного класу).

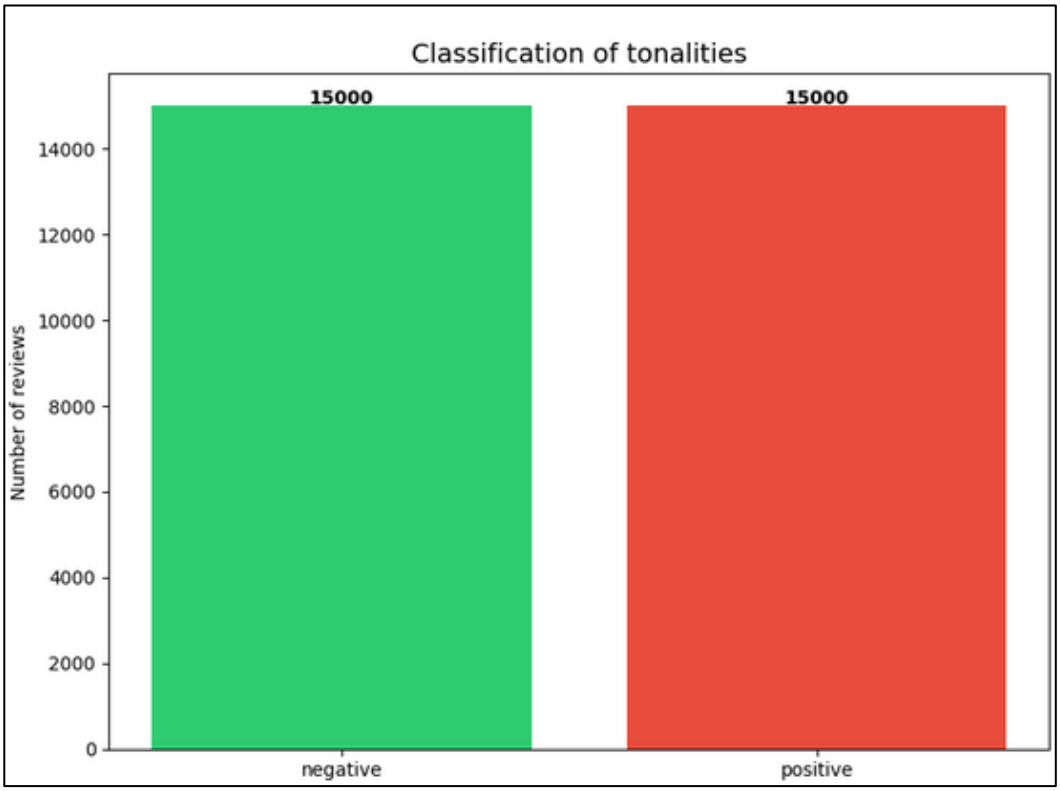

Рис. 1. Розподіл класів тональності після балансування даних

На рисунку 1 відображено симетричний розподіл цільової змінної. Така конфігурація є критично важливою для того, щоб модель об'єктивно вивчала характерні ознаки обох категорій, а не максимізувала загальну точність за рахунок ігнорування міноритарного класу.

Процес підготовки текстової інформації у даному дослідженні розглядається як багатоступеневий інтелектуальний конвеєр, спрямований на максимальну деструктуризацію шуму при повному збереженні семантичного ядра повідомлення. На відміну від класичних задач класифікації, аналіз літературних відгуків вимагає прецизійного підходу до морфології та синтаксису, оскільки читачі часто використовують складні мовленнєві звороти для опису сюжету, які не завжди прямо корелюють із їхнім власним ставленням до твору.

Програмна реалізація лінгвістичного конвеєра базується на використанні бібліотек NLTK та re [3-5]. Основні етапи включали:

- Очищення тексту від HTML-тегів, URL-адрес та неалфавітних символів;
- Приведення до нижнього регістру та токенізацію;
- Лематизацію: Використання *WordNetLemmatizer* для зведення слів до словникової форми, що дозволяє зберегти семантичний зв'язок між різними формами одного слова.

Ключовим науковим внеском у препроцесинг є впровадження функції handle_negation. На відміну від стандартних підходів, де частка "not" видаляється як стоп-слово, розроблений алгоритм ідентифікує заперечні маркери (*not, no, never, n't*) і зливає їх із наступними значущими словами в єдині вектори (наприклад, *"not_good", "not_recommended"*). Це дозволяє моделі враховувати інверсію сенсу, яка є поширеною у розлогих книжкових рецензіях.

Ефективність описаного конвеєра наочно підтверджується порівняльним аналізом найбільш частотних термінів (Рис. 2).

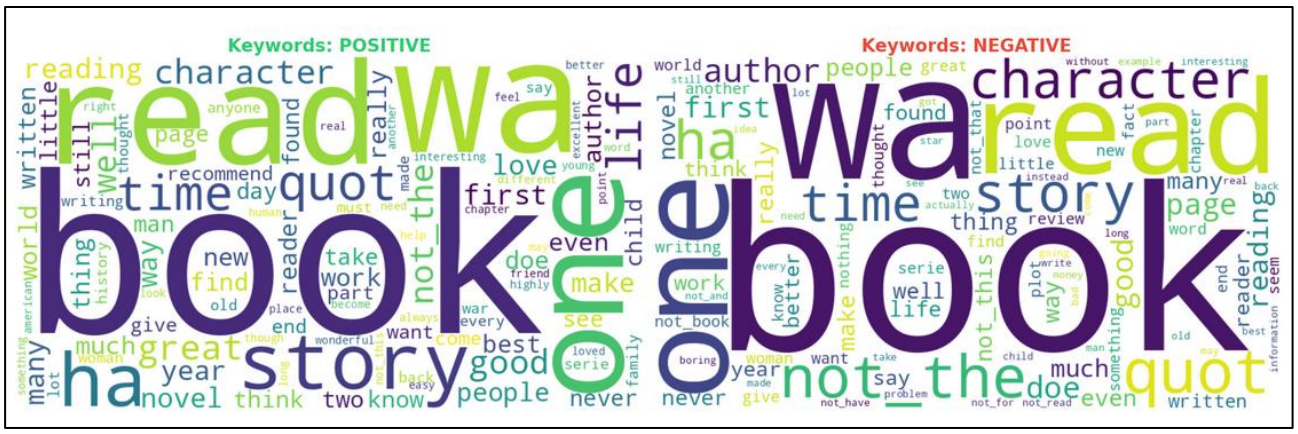

Рис. 2. Графік хмари слів позитивних та негативних

Рисунок 2 демонструє різницю в полярності термінів. У позитивному сегменті домінують лексеми *great, love, masterpiece*, тоді як у негативному – завдяки конвеєру обробки – чітко виділяються предиктори *not_good, not_read, bad* та *boring*. Це підтверджує, що попередня обробка успішно виділила ключові емоційні маркери.

Для виявлення найбільш способу представлення інформації проведено порівняльний аналіз статистичного методу Bag-of-Words (BoW) та лінійного методу TF-IDF. Векторизація виконувалася з розширенням простору ознак до триграм (n=1,2,3), що дозволило вловити стійкі ідіоми та літературні звороти. Розмірність простору була обмежена 5000 найбільш значущих термінів для запобігання «прокляттю розмірності».

На основі сформованих матриць було навчено моделі MultinomialNB (як статистичний базис) та Logistic Regression (як інтерпретований еталон лінійної класифікації).

SUMMARY TABLE OF RESULTS

Модель	Accuracy	Precision	Recall	ROC-AUC	F1 negative	F1 positive	Time (sec)
TF-IDF + LogisticRegression	0.8822	0.8822	0.8822	0.9504	0.8824	0.8819	21.29
TF-IDF + MultinomialNB	0.8680	0.8680	0.8680	0.9387	0.8686	0.8674	20.04
BoW + LogisticRegression	0.8637	0.8637	0.8637	0.9266	0.8637	0.8637	53.63
BoW + MultinomialNB	0.8565	0.8566	0.8565	0.9203	0.8578	0.8552	19.60

Рис. 3. Порівняння ефективності моделей за комплексом метрик

На основі отриманих експериментальних даних (рис. 3) встановлено, що лінійний метод векторизації TF-IDF демонструє системну перевагу над статистичним підходом Bag-of-Words (BoW) за всіма ключовими метриками якості. Така різниця в ефективності класифікації текстових відгуків зумовлена фундаментальними математичними особливостями представлення ознак.

Метод Bag-of-Words базується на простому частотному підрахунку токенів. Це призводить до того, що загальновживані слова (наприклад, *"book", "read", "author"*), які часто зустрічаються в обох класах відгуків, отримують найбільшу вагу, створюючи значний інформаційний шум. Натомість метод TF-IDF використовує механізм зворотньої частоти документа (IDF), який дозволяє «штрафувати» такі неінформативні лексеми, одночасно підсилюючи вагу рідкісних емоційних маркерів (*"excellent", "disappointing", "boring"*). Саме це дозволило моделі Logistic Regression на базі TF-IDF досягти найвищої точності (88,22%) та найкращої роздільної здатності за метрикою ROC-AUC (0,9504).

Окремим чинником успіху є використання L_2-нормалізації векторів, яка інтегрована в алгоритм TF-IDF. Завдяки цьому лінійні класифікатори швидше знаходять глобальний оптимум при навчанні. Як видно з результатів експерименту, час обробки даних для логістичної регресії при переході від BoW до TF-IDF скоротився з 53,63 с до 21,29 с (у 2,5 раза), що підтверджує не лише вищу точність, а й кращу

обчислювальну масштабованість лінійних методів векторизації. Таким чином, TF-IDF визначено як найбільш оптимальний метод для задач аналізу тональності розріджених текстових даних.

Для оцінки здатності системи диференціювати класи та виявлення «зон слабкості» використано Confusion Matrix. На рисунку 4 видно, що модель TF-IDF + Logistic Regression успішно ідентифікувала 2653 негативних та 2640 позитивних відгуків на тестовій вибірці. Помилки першого та другого роду (347 та 360 випадків відповідно) є симетричними, що вказує на відсутність алгоритмічного зміщення.

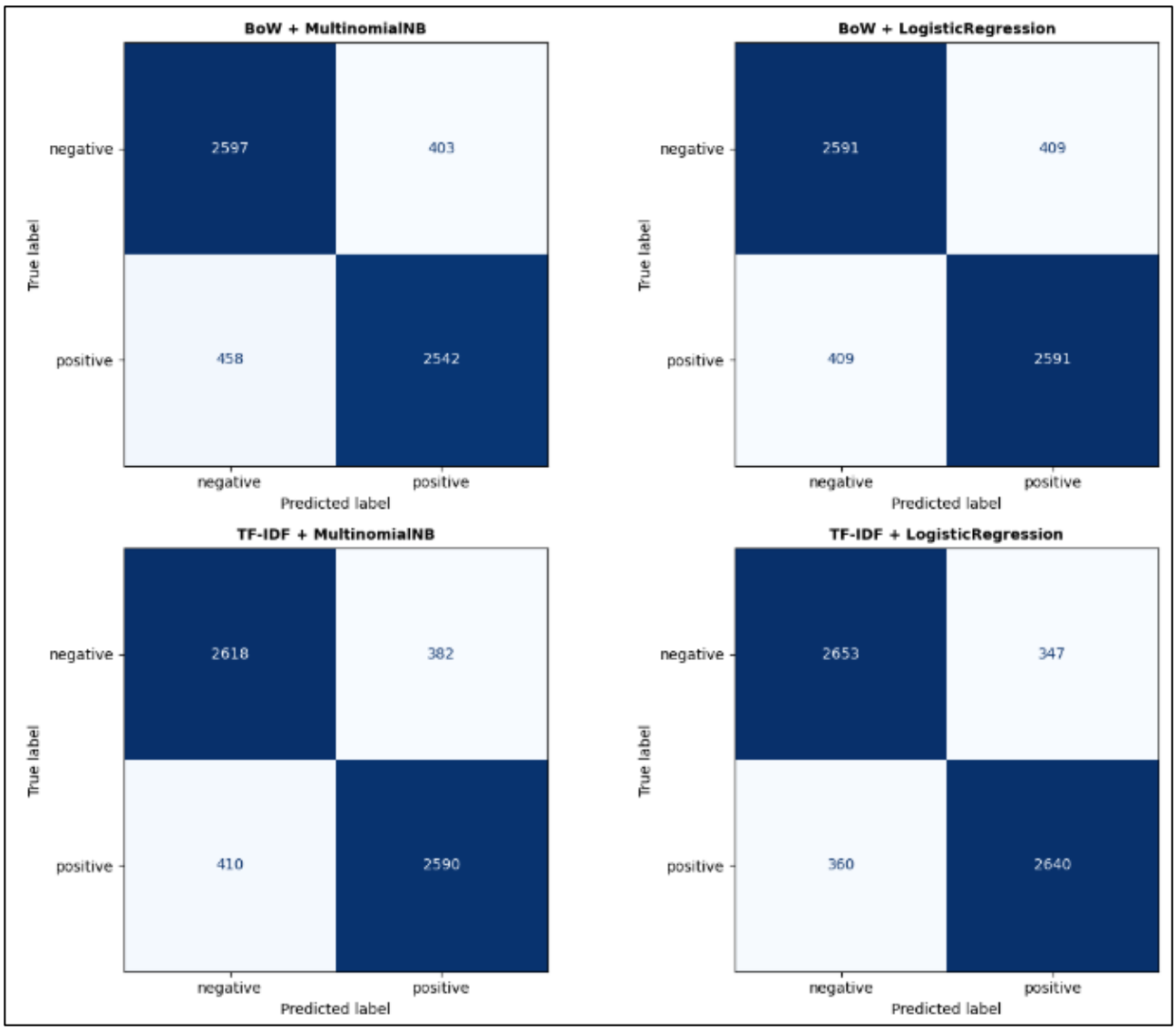

Рис. 4. Матриця помилок моделей

Детальний аналіз хибних передбачень показав, що основні труднощі викликані:

1. Сарказмом: вислови типу *"настільки 'чудова' книга, що не зміг дочитати"* сприймаються як позитивні через вагу слова "чудова".
2. Змішаними контекстами: коли читач хвалить стиль автора, але критикує сюжет.
3. Описами сюжету: використання негативної лексики для опису трагічних подій у книзі, яку користувач оцінив позитивно.

Незважаючи на ці складні випадки, показник ROC-AUC 0,9504 свідчить про високу надійність системи та її готовність до практичного впровадження в аналітичні платформи електронної комерції.

Загалом, результати дослідження продемонстрували високу ефективність комбінації логістичної регресії та методу векторизації TF-IDF при класифікації емоційної тональності неструктурованих текстів. Отримані показники підтверджують, що лінійний підхід до зважування ознак, посилений алгоритмом Negation Handling, дозволяє успішно диференціювати складні лінгвістичні конструкції та

заперечення. Модель успішно використовує статистичну значущість n-грам для створення потужного класифікатора, який забезпечує точність на рівні 88,22%. Розроблений програмний конвеєр може стати надійною основою для інтелектуальних систем моніторингу споживчих настроїв у реальному часі на платформах електронної комерції.

Висновки

У результаті проведеного дослідження розроблено та апробовано систему класифікації емоційної тональності розріджених текстових даних. У результаті проведеного системного аналізу доведено, що вибір методу векторизації є визначальним фактором при моделюванні емоційної тональності текстів. Дослідження підтвердило, що лінійний метод TF-IDF, завдяки штрафуванню часто вживаних стоп-слів, формує більш інформативний простір ознак порівняно з базовим Bag-of-Words. Це дозволяє логістичній регресії та наївному класифікатору Баєса працювати з меншим рівнем шуму, що безпосередньо впливає на підвищення точності класифікації відгуків Amazon Books.

Попередня обробка тексту, зокрема впровадження алгоритму Negation Handling для обробки заперечень, значно підвищує якість моделювання шляхом збереження семантичного контексту. Типові маркери позитивної тональності включають лексеми *"excellent"*, *"amazing"*, *"wonderful"*, тоді як негативної – *"boring"*, *"disappointing"* та специфічні заперечні біграми на кшталт *"not_read"*.

Застосування L_2-регуляризації та логарифмічного масштабування термів (sublinear_tf) дозволило уникнути перенавчання моделі та утричі прискорити процес обчислень порівняно зі статистичними методами. Розроблений програмний конвеєр демонструє здатність ефективно класифікувати відгуки навіть за наявності складних літературних зворотів, що робить його цінним інструментом для систем автоматизованого моніторингу ринку.

Унікальність реалізованого підходу полягає в інтеграції спеціалізованого алгоритму обробки лінгвістичних заперечень (Negation Handling), який дозволяє зберігати семантичний контекст фрази шляхом трансформації заперечних конструкцій у стійкі семантичні біграми. Незважаючи на складність літературної мови та наявність розлогих описів сюжету, поєднання методів векторизації TF-IDF із лінійними класифікаторами демонструє високу точність у розпізнаванні емоційного забарвлення.

Отримані результати мають практичне значення для розробки систем підтримки прийняття рішень у видавничій галузі та на платформах електронної комерції. Подальші дослідження можуть включати впровадження методів глибокого навчання (Deep Learning) та семантичних векторів для розпізнавання прихованого сарказму та іронії.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Amazon Books Reviews. 2022 [Електронний ресурс]. – Режим доступу: https://www.kaggle.com/datasets/mohamedbakhet/amazon-books-reviews
2. Sentiment Analysis. 2025 [Електронний ресурс]. – Режим доступу: https://www.kaggle.com/code/mariana65/sentiment-analysis/
3. В. Б. Мокін, М. В. Дратований. Наука про дані: машинне навчання та інтелектуальний аналіз даних — Вінниця, ВНТУ, 2024. – 258 с.
4. Matplotlib Pyplot Documentation. 2025 [Електронний ресурс]. – Режим доступу: https://matplotlib.org/3.5.3/api/_as_gen/matplotlib.pyplot.html
5. Bag-of-words vs TF-IDF. 2025 [Електронний ресурс]. – Режим доступу: https://www.geeksforgeeks.org/nlp/bag-of-words-vs-tf-idf/

Білецька Мар'яна Володимирівна – студентка групи СА-22б, факультет інтелектуальних інформаційних технологій та автоматизації, Вінницький національний технічний університет, м. Вінниця. e-mail: marjnabilecka@gmail.com

Жуков Сергій Олександрович – к.т.н., доцент кафедри системного аналізу та інформаційних технологій, Вінницький національний технічний університет, e-mail: sazhukov@gmail.com

Biletska Mariana – student of Faculty of Intelligent Information Technologies and Automation, SA-22b, Vinnytsia National Technical University, Vinnytsia, e-mail marjnabilecka@gmail.com

Zhukov Serhii O. – Ph.D., Assistant Professor of the Department of Systems Analysis and Information Technology, Vinnytsia National Technical University, Vinnytsia, e-mail: sazhukov@gmail.com