УДК: 004.891.3:004.912

Д. С. Трухін
С. О. Жуков

ПОРІВНЯЛЬНИЙ АНАЛІЗ ІМОВІРНІСНИХ МОДЕЛЕЙ У ЗАДАЧАХ КЛАСИФІКАЦІЇ ТОНАЛЬНОСТІ

Вінницький національний технічний університет

Анотація

Роботу присвячено порівняльному дослідженню двох варіантів наївного Байєсівського класифікатора — мультиноміальної (MultinomialNB) та бернуллієвської (BernoulliNB) моделей — у задачі автоматичної класифікації тональності текстових відгуків покупців. Реалізовано повний ML-пайплайн на датасеті Amazon Reviews: попередня обробка тексту (токенізація, лемматизація, видалення стоп-слів), TF-IDF та бінарна векторизація з біграмами, підбір параметра згладжування Лапласа методом GridSearchCV. Проведено порівняльний аналіз моделей за метриками Accuracy та F1-score, побудовано матриці помилок та визначено найбільш інформативні слова для кожного класу тональності.

Ключові слова: аналіз тональності тексту, наївний Байєсівський класифікатор, MultinomialNB, BernoulliNB, TF-IDF, обробка природної мови, Amazon Reviews, Python, Scikit-learn.

Abstract

The paper is devoted to a comparative study of two variants of the Naive Bayes classifier — Multinomial (MultinomialNB) and Bernoulli (BernoulliNB) models — for automatic sentiment classification of customer reviews. A complete ML pipeline was implemented on the Amazon Reviews dataset: text preprocessing (tokenization, lemmatization, stop-word removal), TF-IDF and binary vectorization with bigrams, and Laplace smoothing parameter tuning via GridSearchCV. A comparative analysis of the models was conducted using Accuracy and F1-score metrics, confusion matrices were constructed, and the most informative features for each sentiment class were identified.

Keywords: sentiment analysis, Naive Bayes classifier, MultinomialNB, BernoulliNB, TF-IDF, natural language processing, Amazon Reviews, Python, Scikit-learn.

Вступ

Стрімке зростання обсягів текстових даних на платформах електронної комерції робить задачу автоматичного аналізу тональності відгуків (Sentiment Analysis) одним із ключових напрямів прикладного машинного навчання. Здатність алгоритму правильно визначити емоційне забарвлення відгуку безпосередньо впливає на якість рекомендаційних систем, маркетингової аналітики та моніторингу репутації брендів.

Наївний Байєсівський класифікатор є одним із найефективніших інструментів для текстової класифікації завдяки простоті навчання, низьким обчислювальним витратам та прозорій інтерпретованості результатів. При цьому існують дві принципово різні реалізації цього алгоритму для роботи з текстом: MultinomialNB, що враховує частоту появи кожного слова, та BernoulliNB, що оперує лише фактом присутності або відсутності слова. Вибір між цими моделями суттєво впливає на якість класифікації залежно від характеристик вхідних даних, однак систематичне порівняння цих підходів на великих реальних датасетах залишається актуальною задачею

Метою роботи є реалізація та порівняльний аналіз MultinomialNB і BernoulliNB на датасеті Amazon Reviews, визначення умов переваги кожної моделі та формулювання практичних рекомендацій щодо їх застосування.

Результати дослідження

Як вхідні дані використано датасет Amazon Reviews з платформи Kaggle [1], що містить текстові відгуки покупців із мітками тональності. Для забезпечення коректного порівняння моделей сформовано збалансовану вибірку: 80 000 прикладів для навчання та 20 000 для тестування — по 50% позитивних та негативних відгуків у кожній підвибірці. Розподіл класів наведено на рис. 1.

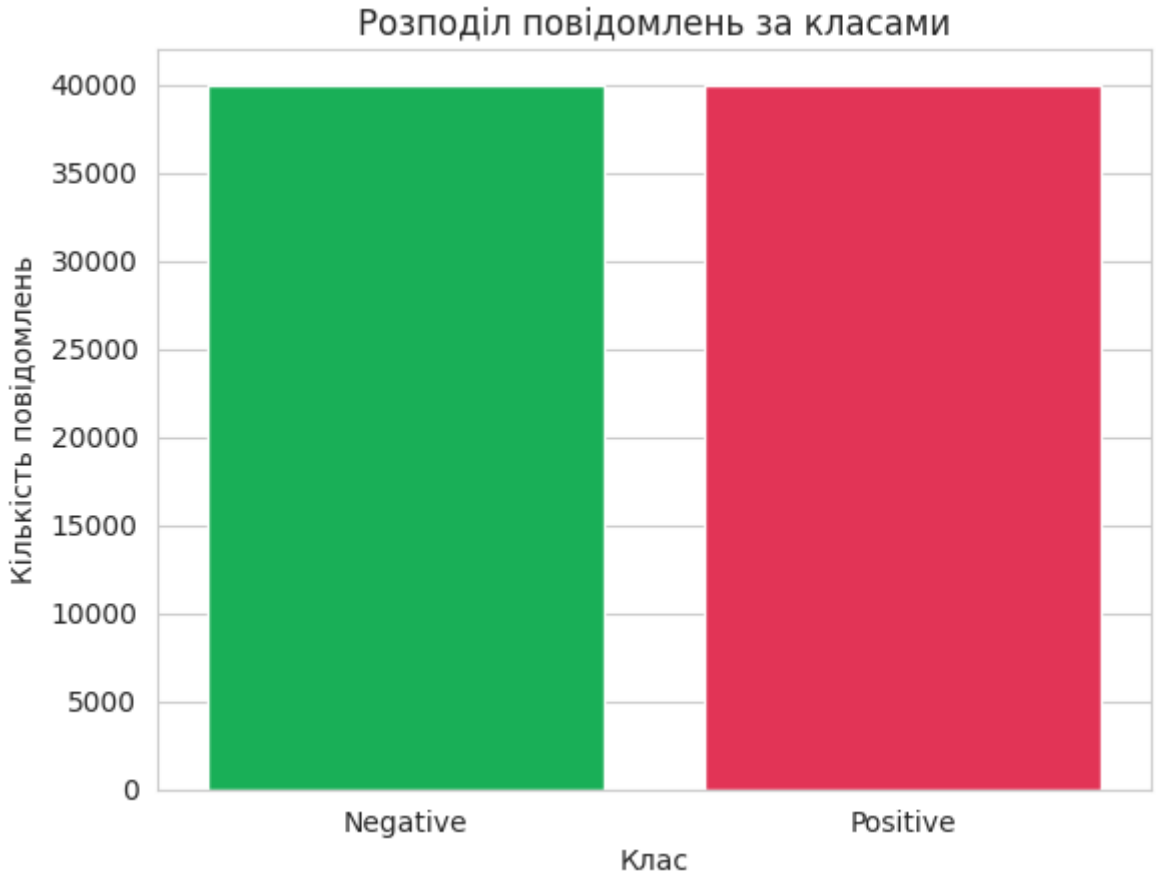

Рис. 1. Розподіл класів у тренувальній вибірці

Попередня обробка тексту виконувалась за допомогою бібліотеки NLTK [2] та включала: приведення до нижнього регістру, токенізацію (*word_tokenize*), видалення стоп-слів англійської мови та лемматизацію (*WordNetLemmatizer*). Застосування лемматизації дозволило скоротити розмір словника та зменшити розмірність простору ознак без втрати семантичної інформації.

Для MultinomialNB застосовано TF-IDF векторизацію (*TfidfVectorizer*) з параметрами *ngram_range=(1,2)*, *max_features=30000*, *sublinear_tf=True*. Використання біграм дозволяє моделі враховувати словосполучення на кшталт «not good» або «highly recommend», що підвищує точність класифікації порівняно з уніграмним представленням. Для BernoulliNB застосовано бінарну CountVectorizer-векторизацію (*binary=True*) з аналогічними параметрами біграм та обмеження словника.

Для обох моделей виконано автоматичний підбір параметра згладжування Лапласа α методом GridSearchCV із 5-кратною крос-валідацією по сітці значень {0.01, 0.05, 0.1, 0.5, 1.0, 2.0, 5.0, 10.0} [3]. Матриці помилок фінальних моделей наведено на рис. 2.

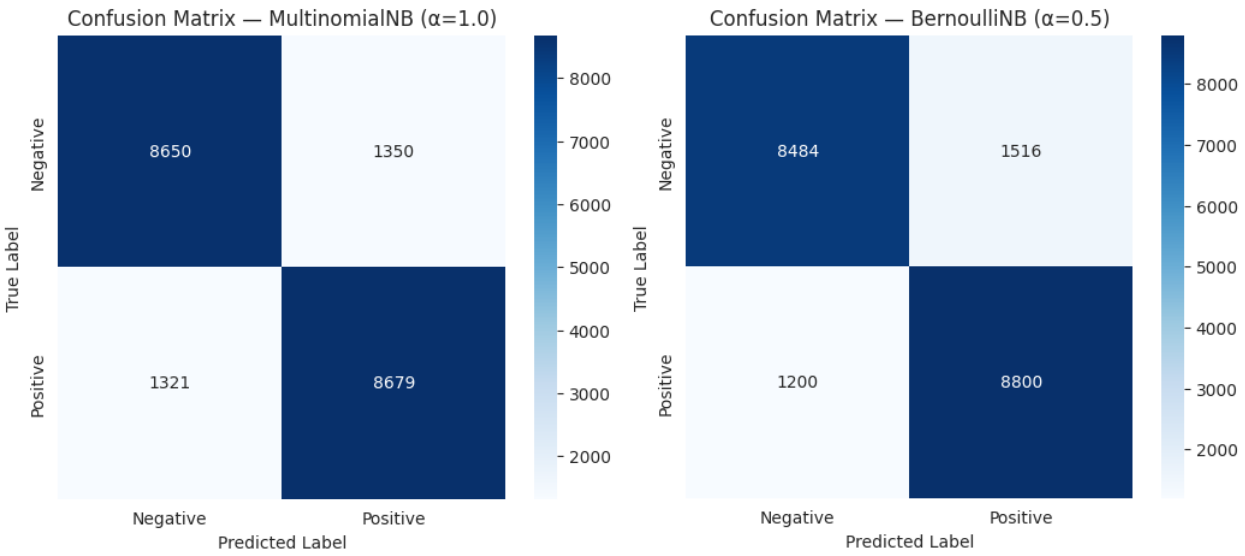

Рис. 2. Матриці помилок: MultinomialNB та BernoulliNB

Порівняльний аналіз моделей виконано за метриками Accuracy та F1-score (macro-усереднення). Вибір макро-усереднення обумовлений збалансованістю вибірки та необхідністю рівноцінного врахування якості класифікації обох класів. Результати порівняння наведено у табл. 1 та на рис. 3.

Метрика	MultinomialNB	BernoulliNB	Різниця
Accuracy	0.8665 (86.65%)	0.8642 (86.42%)	+0.0023
F1-score (macro)	0.8664	0.8642	+0.0023
F1-score Positive	0.8666	0.8663	+0.0003
F1-score Negative	0.8663	0.8620	+0.0042
Training Time (c)	16.5200	15.0025	–1.5175
Prediction Time (c)	0.0044	0.0129	+0.0085

Табл. 1. Порівняння результатів MultinomialNB та BernoulliNB

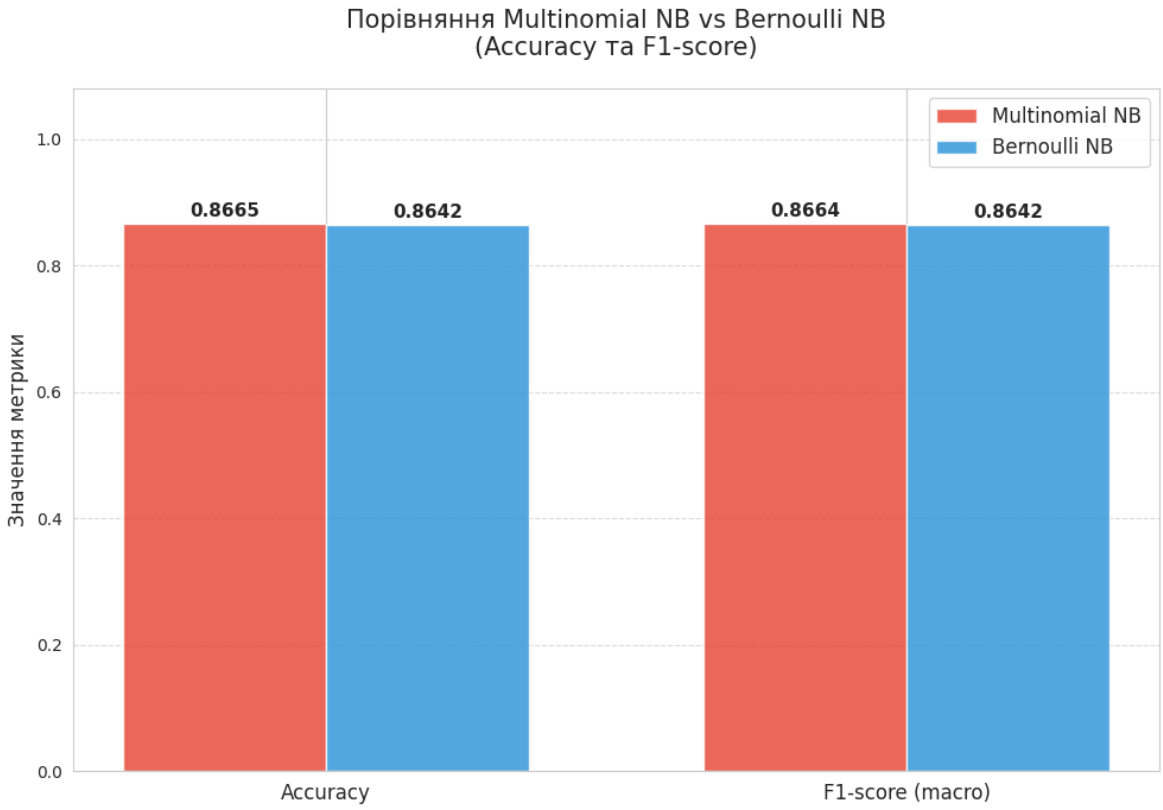

Рис. 3. Порівняльна гістограма метрик MultinomialNB та BernoulliNB

Для інтерпретації результатів класифікації визначено топ-15 найбільш інформативних слів для кожного класу на основі лог-ймовірностей *feature_log_prob_* обох моделей (рис. 4-5). Аналіз показує, що для позитивного класу обидві моделі виявляють слова на кшталт «great», «book», «good», тоді як для негативного — «much», «would», «money». Відмінність між моделями полягає у тому, що MultinomialNB додатково виявляє інформативні біграми, тоді як BernoulliNB більш чутлива до рідкісних, але специфічних термінів.

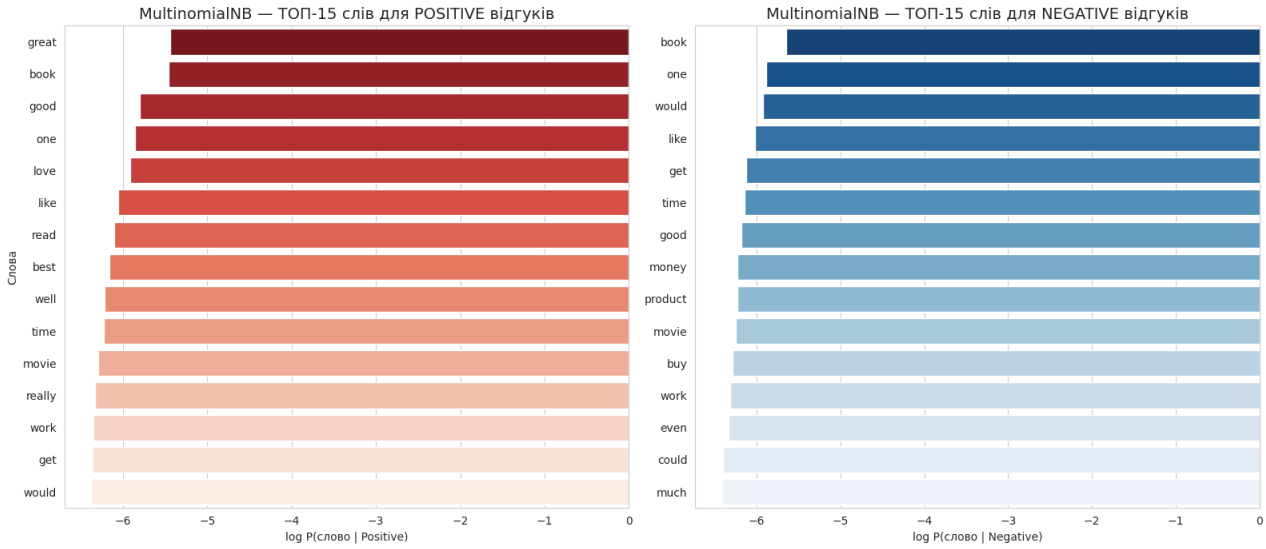

Рис. 4. Графік топ-15 слів для моделі MultinomialNB

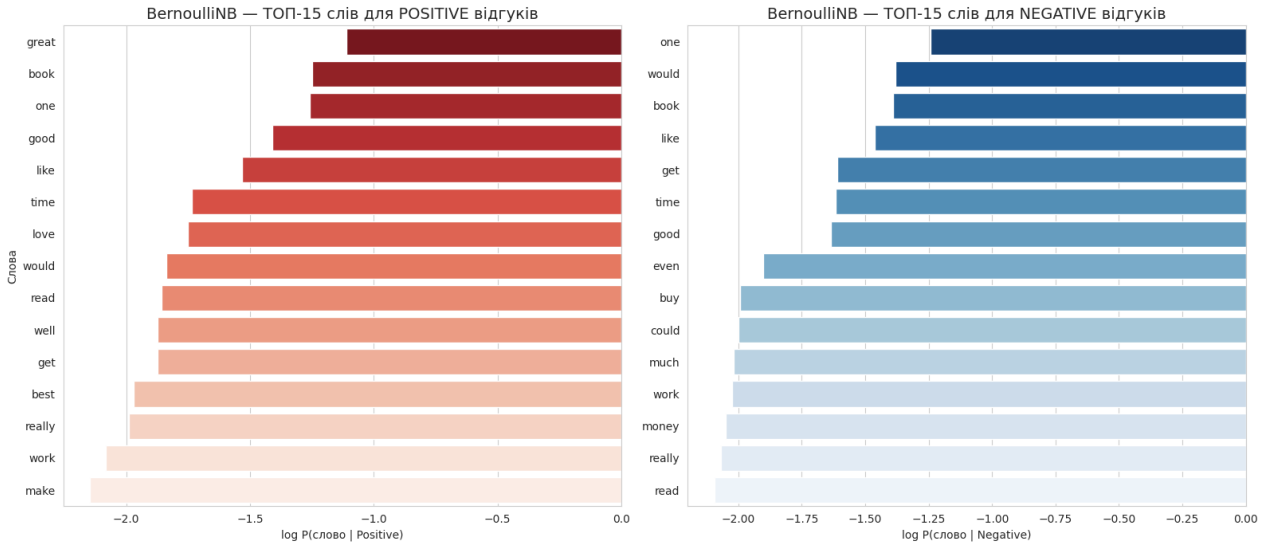

Рис. 5. Графік топ-15 слів для моделі BernoulliNB

Аналіз показує, що для позитивного класу обидві моделі виявляють слова на кшталт «great», «book», «good», тоді як для негативного — «much», «would», «money». Відмінність між моделями полягає у тому, що MultinomialNB додатково виявляє інформативні біграми, тоді як BernoulliNB більш чутлива до рідкісних, але специфічних термінів.

Висновки

Реалізовано та порівняно два підходи до байєсівської класифікації тональності відгуків Amazon Reviews. Встановлено, що MultinomialNB із TF-IDF векторизацією демонструє вищу якість класифікації завдяки врахуванню частоти термінів та ефективному використанню біграм. BernoulliNB із бінарною векторизацією поступається за точністю, проте є швидшою у навчанні та більш доцільною для коротких текстів з обмеженим словником.

Підбір параметра згладжування α методом GridSearchCV дозволив підвищити якість обох моделей порівняно з використанням значення за замовчуванням (α=1.0). Аналіз матриць помилок підтвердив симетричний розподіл помилок між класами, що є очікуваним результатом для збалансованого датасету

Отримані результати демонструють ефективність MultinomialNB для задачі Sentiment Analysis з результатом 86.65%, та можуть бути використані як базовий орієнтир при виборі між частотним та бінарним представленням тексту у практичних системах класифікації відгуків.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Amazon Reviews Dataset. Kaggle, 2021. URL: https://www.kaggle.com/datasets/kritanjalijain/amazon-reviews
2. Bird S., Klein E., Loper E. Natural Language Processing with Python. O'Reilly Media, 2009. URL: https://www.nltk.org/book/
3. Pedregosa F. та ін. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research. 2011. Vol. 12. P. 2825–2830
4. McCallum A., Nigam K. A Comparison of Event Models for Naive Bayes Text Classification. AAAI-98 Workshop on Learning for Text Categorization. 1998. P. 41–48
5. Концевой А., Бісікало О. Моделі глибокого навчання для вирішення задачі класифікації текстової інформації. Інформаційні технології та комп'ютерна інженерія. 2022. Т. 55, № 3. С. 13–20

Трухін Дмитро Сергійович – студент групи СА-23б, факультет інтелектуальних інформаційних технологій та автоматизації, Вінницький національний технічний університет, м. Вінниця, e-mail: dimatr2016@gmail.com.

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

Trukhin Dmytro S. – student of group SA-23b, Faculty of Intelligent Information Technologies and Automation, Vinnytsia National Technical University, Vinnytsia, e-mail: dimatr2016@gmail.com.

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