РОЗДІЛ ІІІ. КОМП'ЮТЕРНІ НАУКИ

УДК 004.82
DOI https://doi.org/10.26661/2786-6254-2023-1-05

ПОРІВНЯННЯ КЛАСИФІКАТОРІВ ДЛЯ ЗАДАЧІ АНАЛІЗУ ТОНАЛЬНОСТІ ТЕКСТУ

Бойко Н. І.
кандидат економічних наук, доцент,
доцент кафедри систем штучного інтелекту
Національний університет «Львівська політехніка»
вул. Князя Романа, 5, Львів, Україна
orcid.org/0000-0002-6962-9363
Nataliya.i.boyko@lpnu.ua

Кулінченко А-М. Р.
студентка IV курсу кафедри систем штучного інтелекту
Національний університет «Львівська політехніка»
вул. Князя Романа, 5, Львів, Україна
orcid.org/0000-0002-8987-1851
anna-mariia.kulinchenko.knm.2019@lpnu.ua

Газдюк К. П.
доктор філософії за спеціальністю 121 (інженерія програмного забезпечення),
асистент кафедри програмного забезпечення комп'ютерних систем
Чернівецький національний університет імені Юрія Федьковича
вул. Коцюбинського, 2, Чернівці, Україна
orcid.org/0000-0002-7568-4422
kateryna.gazdyik@gmail.com

Ключові слова:
сентимент аналіз, наївний Баєс, логістична регресі, дерева рішень, випадковий ліс, машинне навчання, класифікатор, гіперпараметри.

Метою дослідження є визначення найбільш ефективного класифікатора для задачі аналізу тональності тексту. Серед вибраних у роботі для порівняння наводиться наївний Баєс, логістична регресія, дерево рішень, випадковий ліс. Для задачі аналізу тональності тексту вибрано сентимент аналіз. Основою для проведення дослідження вибрано набір відгуків на фільми, що надані критиками IMDB. Об'єктами дослідження є безпосередньо вибрані класифікатори, а предметом відповідно є визначення їх ефективності у разі застосування до вищезгаданої задачі. Задачею цього розділу є ознайомлення та оцінка методів класифікації у контексті задачі сентимент аналізу. Для порівняння було вибрано такі класифікатори, як наївний Баєс, логістична регресія, дерево рішень та випадковий ліс. Перейдемо до детальнішого опису кожного з них. Це дослідження дозволить визначити найефективніший алгоритм класифікації для аналізу тональності тексту. Це, своєю чергою, дає можливість програмам, які виконують такий аналіз, покращити якість розподілення тексту на різні групи. У цій роботі було визначено класифікатор, який серед вибраних є найефективнішим, – це метод логістичної регресії. Під час виконання такої роботи були проведені аналізи актуальності задачі та наукових джерел, серед яких було досліджено точність класифікаторів наївного Баєса, логістичної

регресії, дерев рішень та випадкового лісу. Перед безпосередньою класифікацією наведений розподіл набору даних на навчальну та тестову вибірки. Проводиться Тренування кожного класифікатора з певними гіперпараметрами. Також був виконаний детальний аналіз та підготовка даних для задачі бінарної класифікації. Паралельно виконувалось безпосереднє тренування класифікаторів та проведення експериментів з кожним. Було обговорено результати дослідження за допомогою повної статистики всіх метрик та всіх вибраних класифікаторів. Для покращення точності класифікаторів необхідно підбирати відповідні гіперпараметри на кожен тип. Проводиться аналіз самих слів. Проведено статистичні обчислення слів, вживаних у позитивних та негативних відгуках, та побудовані, відповідно, «хмари слів» з найбільш вживаними словами. Для детальнішого аналізу побудовано також матриці невідповідностей по кожному методу.

COMPARISON OF CLASSIFIERS FOR THE TASK OF TEXT TONALITY ANALYSIS

Boyko N. I.
PhD, Associate Professor,
Associate Professor at the Department of Artificial Intelligent Systems
Lviv Polytechnic National University
Prince Roman str., 5, Lviv, Ukraine
orcid.org/0000-0002-6962-9363
Nataliya.i.boyko@lpnu.ua

Kulinchenko A-M. R.
4th year Student at the Department of Artificial Intelligent Systems
Lviv Polytechnic National University
Prince Roman str., 5, Lviv, Ukraine
orcid.org/0000-0002-8987-1851
anna-mariia.kulinchenko.knm.2019@lpnu.ua

Hazdiuk K. P.
Doctor of Philosophy in Speciality 121 (Software Engineering),
Assistant at the Department of the Software of Computer Systems
Yuriy Fedkovych Chernivtsi National University
Kotsyubynsky str., 2, Chernivtsi, Ukraine
orcid.org/0000-0002-7568-4422
kateryna.gazdyik@gmail.com

Key words: *sentiment analysis, naive Bayes, logistic regression, decision trees, random forest, machine learning, classifier, hyperparameters.*

The study aims to determine the most effective classifier for the task of analyzing the tonality of the text. Naive Bayes, logistic regression, decision trees, and random forests are among the ones specified in work for comparison. Sentiment analysis was chosen for the task of analyzing the tonality of the text. A set of movie reviews provided by IMDB critics was selected as the basis for the research. The objects of the study are chosen directly as classifiers, and the subject, accordingly, is the determination of their effectiveness when applied to the problem mentioned above. This chapter aims to introduce and evaluate classification methods in the context of sentiment analysis. Classifiers such as naive Bayes, logistic regression, decision tree and random forest were compared. Let's move on to a more detailed description of each of them. This study determines the most effective classification algorithm for the analysis of the text's tonality. This, in turn, enables programs that perform such analysis

to improve the quality of text distribution into different groups. In this work, the classifier was determined which among the selected ones is the most effective – it is the method of logistic regression. During the implementation of this work, analyses of the relevance of the problem and scientific sources were conducted, among which the accuracy of the naive Bayes, logistic regression, decision trees, and random forest classifiers was investigated. Before direct classification, the distribution of the data set into training and test samples is given. Each classifier is trained with specific hyperparameters. Detailed analysis and data preparation for the binary classification task was also performed. In parallel, training classifiers and conducting experiments with each were carried out. The study results were discussed using complete statistics of all metrics and all selected classifiers. To improve the accuracy of classifiers, choosing appropriate hyperparameters for each type is necessary. Analysis of the words themselves is carried out. Statistical calculations of the terms used in positive and negative reviews were carried out, and accordingly, "word clouds" with the most used words were constructed. For a more detailed analysis, inconsistency matrices were also created for each method.

1. Вступ

Завдання аналізу емоційного забарвлення тексту поступово набуває популярності у зв'язку зі збільшенням активної аудиторії та їх середнім часом проведення у соціальних мережах та Інтернеті загалом [1; 4].

Безліч виробничих компаній, використовуючи соціальні мережі та вебсторінки поширення своєї продукції, а саме коментарі з цих джерел, здійснюють аналіз тексту з метою збору та обробки рейтингової інформації про власні продукти, а відповідно, і задля покращення якості продукту в майбутньому [6; 2].

Такий аналіз також використовується державними органами безпеки багатьох країн, а саме на предмет виявлення інформації, що несе незаконний характер. Наприклад: торгівля незаконними товарами, погрози, попередження про можливі терористичні дії, інформація про діяльність небезпечних організацій тощо.

Метою дослідження є визначення найбільш ефективного класифікатора для задачі аналізу тональності тексту. Серед вибраних у роботі для порівняння наводиться наївний Баєс, логістична регресія, дерево рішень, випадковий ліс. Для задачі аналізу тональності тексту вибрано сентимент аналіз.

Основою для проведення дослідження вибрано набір відгуків на фільми, що надані критиками IMDB. Об'єктами дослідження є безпосередньо вибрані класифікатори, а предметом відповідно є визначення їх ефективності у разі застосування до вищезгаданої задачі.

Це дослідження дозволить визначити найефективніший алгоритм класифікації для аналізу тональності тексту. Це, своєю чергою, дає можливість програмам, які виконують такий аналіз, покращити якість розподілення тексту на різні групи.

2. Аналіз літературних джерел

У 2020 році в науковому журналі «Комп'ютерно-інтегровані технології: освіта, наука, виробництво» було опубліковано статтю «Порівняльний аналіз методів для вирішення задачі сентимент аналізу тексту» за авторством аспіранта С.С. Мироненка та студентки Є.А. Онищенко Національного технічного університету України «Київський політехнічний інститут імені Ігоря Сікорського» [7; 3]. У своєму дослідженні вони провели порівняння таких класифікаторів, як:

1) наївний Баєсівський класифікатор, який у разі класифікації на вибірці для тренування дав точність 98%, а на вибірці для тестування – 88%;

2) класифікатор на основі рекурентної нейронної мережі з довгою короткочасною пам'яттю, який у разі класифікації на вибірці для тренування дав точність 95%, а на вибірці для тестування – 75%;

3) класифікатор на основі одновимірної згорткової нейронної мережі, який у разі класифікації на вибірці для тренування дав точність 99%, а на вибірці для тестування – 86%;

4) класифікатор на основі трьохвимірної загорткової нейронної мережі з попередньою обробкою тексту за допомогою токенізатора BERT, який у разі класифікації на вибірці для тренування дав точність 99%, а на вибірці для тестування – 89%.

Найпопулярнішими методами класифікації для сентимент аналізу у сфері туризму були визначені метод опорних векторів та наївний Баєс. Китайські дослідники Йе, Жанг і Ло у своїй праці (Qiang Ye, Ziqiong Zhang, Rob Law, 2009, "Sentiment classification of online reviews to travel destinations by supervised machine learning approaches") [3] порівняли ці два методи і зазначили, що добилися кращої точності у разі класифікації методом опорних векторів, а саме 80%.

Через півтора року науковці Ші та Лі здійснили додаткові до попереднього дослідження (Han-Xiao Shi, Xiao-Jun Li, 2011, "A Sentiment Analysis Model for Hotel Reviews Based on Supervised Learning" In Machine Learning and Cybernetics (ICMLC)) [4], порівнявши ті самі методи, в результаті вони отримали майже на 5% вищу точність, використовуючи той же метод опорних векторів.

У тому ж році дослідники Йе, Жанги і Ло повторили своє дослідження, проте за основу взяли інший набір даних. У своїй роботі (Ziqiong Zhang, Qiang Ye, Zili Zhang, Yijun Li, 2011, "Sentiment Classification of Internet Restaurant Reviews Written in Cantonese") вони порівняли такі класифікатори на основі ресторанних відгуків та отримали вищу точність на класифікаторі наївного Баєса, а саме 91% [5].

У 2018 році магістрант Харківського національного університету радіоелектроніки М.Г. Литвинов у своїй роботі («Дослідження моделей оцінювання тонального забарвлення тексту») [6] порівняв класифікацію наївним Баєсом та методом k найближчих сусідів. За допомогою наївного класифікатора Баєса йому вдалось досягнути точності 84%, а за допомогою метода k найближчих сусідів – 62%.

3. Матеріали та методи

Завданням цього розділу є ознайомлення та оцінка методів класифікації у контексті задачі сентимент аналізу. Для порівняння було вибрано такі класифікатори, як наївний Баєс, логістична регресія, дерево рішень та випадковий ліс. Перейдемо до детальнішого опису кожного з них.

3.1. Аналіз класифікатора наївного Баєса

Для визначення належності певного об'єкта (відгука) до того чи іншого класу (негативного чи позитивного) наївний класифікатор Баєса [7] використовує ймовірності, визначені відповідно за допомогою теореми Баєса (Формула 1) з умовою незалежності змінних (змінними в такому випадку є слова, які містяться в відгуках).

$$P(c|d) = \frac{P(d|c)P(c)}{P(d)}, \qquad (1)$$

де:

– $P(c|d)$ – ймовірність того, що об'єкт d належить класу c ;

– $P(d|c)$ – ймовірність зустріти об'єкт d серед усіх об'єктів класу c ;

– $P(c)$ – апріорна ймовірність зустріти об'єкт класу c серед усіх об'єктів;

– $P(d)$ – апріорна ймовірність об'єкта d серед усіх об'єктів.

Тобто для його побудови не потрібно вивчати взаємодію всіх можливих комбінацій змінних (кількість яких експоненційно зростає зі збільшенням числа змінних), а можна обмежитись впливом кожної змінної окремо на належність певного об'єкта до одного з класів. Тобто ми виключаємо ймовірності потраплянь у відгук різних комбінацій слів і розглядаємо лише окремі слова. Таким чином, розмір необхідної вибірки для побудови скорочується, проте модель такого класифікатора буде точною лише за виконання умови незалежності змінних, що є недоліком такого методу. Однак практика показує, що малі відхилення у сторону залежності змінних лише зовсім трохи знижують точність.

Оскільки ціль задачі класифікації полягає в тому, щоб визначити, до якого класу належить, нам потрібно отримати не самі ймовірності, а найбільш імовірний клас, якому може належати об'єкт. Це і є фінальним кроком математичної моделі цього методу [2; 7]. У такому класифікаторі для цього використовується оцінка апостеріорного максимуму (Формула 2).

$$c_{map} = \arg\max \frac{P(d \mid c)P(c)}{P(d)} \ . \qquad (2)$$

Тобто спершу потрібно обчислити ймовірності належності об'єкта до кожного класу, а потім вибрати той, в якого максимальна ймовірність. Оскільки в такій задачі ймовірність зустріти об'єкт серед усіх інших – це константа (бо вибірка є сталою), то вона не буде впливати на розподіл імовірностей класів, отже, ми можемо її ігнорувати, кінцева формула оцінки апостеріорного максимуму буде мати такий вигляд (Формула 3):

$$c_{map} = \arg\max\left[P(d|c)P(c)\right]. \qquad (3)$$

Перевагами цього класифікатора є висока швидкість, простота і масштабованість. Проаналізувавши, можна зробити висновок, що переваги однозначно переважують недоліки, саме тому цей класифікатор є одним з найпопулярніших, особливо для задачі сентимент аналізу, оскільки він дає можливість без значних обчислювальних чи часових затрат надати досить хороший результат навіть без регулювання гіпер-параметрів.

3.2. Аналіз класифікації методом логістичної регресії

Логістична регресія, також відома як логіт-регресія, є часто використовуваним статистичним регресійним методом, який застосовується у випадку, коли залежна змінна може набувати лише бінарних значень, як і є в такому випадку, оскільки відгуки на фільми будуть класифікуватись на позитивні (label = 1) та негативні (label = 0) [8].

Логістична функція – це сигмоїдальна функція, яка приймає будь-яке дійсне значення і повертає значення з проміжку [1]. Це відбувається за допомогою Формули 4 (логістична функція):

$$p = \frac{1}{1 + e^{-y}} . \quad (4)$$

де:

– p – ймовірність того, що подія відбудеться;

– e – число Ейлера;

– y – рівняння регресії.

Наступне перетворення (Формула 5) використовується, щоб лінеаризувати таку функцію і називається логітом або логістичним:

$$p' = \ln \frac{p}{p - 1} . \quad (5)$$

Для визначення коефіцієнтів логістичної регресії, за якими можна відновити ймовірності, використовується метод максимальної правдоподібності.

Логістична регресія є популярним методом, оскільки багато задач можна звести до задачі бінарної класифікації. Її основною перевагою є те, що вона легка в інтерпретації і дає одні з найкращих результатів для задач бінарної класифікації. Також вона відзначається швидкістю та правильністю визначення на нових даних.

3.3. Аналіз класифікації за допомогою дерева рішень

Дерево рішень складається з двох елементів: вузлів та гілок. Рішення вважається прийнятим, коли об'єкт пройшов шлях від кореня дерева до якогось з листків (кінцевих вузлів). Усі вузли, окрім листків, є умовами для певної ознаки поточного об'єкта, які розділяють декілька можливих значень цієї ознаки і від того, під яку умову підпаде така ознака, залежить, в який вузол перейде об'єкт на наступному кроці. В контексті такої задачі умовами є порівняння частот певних слів у відгуку. Тобто якщо частоти позитивних слів будуть більшими за частоти негативних, то ітеративно прогноз буде схилятися в сторону позитивного відгуку, і навпаки. Порівняння продовжуватимуться до того часу як об'єкт не закінчить свій шлях по дереву на якомусь листку, який і буде визначати, до якої категорії він потрапив.

Перевагами методу є:

– простота в розумінні та інтерпретації;

– можливість виконувати класифікацію як категорійну, так і числову;

– дія за принципом «білого ящика», тобто ми можемо точно пояснити будь-який вибір моделі, оскільки всі умови вказані.

Основним недоліком цього методу є те, що в процесі навчання можуть створюватися занадто складні шляхи, які неповною мірою описують дані. Це виникає внаслідок перенавчання моделі. Для того щоб уникнути такої проблеми, потрібно визначити оптимальну глибину дерева та обмежитись нею.

3.4. Аналіз класифікації за допомогою випадкового лісу

Метод випадкового лісу передбачає утворення багатьох окремих дерев рішень, кожне з яких виконує свій незалежний прогноз. Після того як усі дерева завершать прогнозування йде підрахунок усіх результатів, внаслідок якого визначається, яка категорія чи значення набрали найбільше голосів, що і визначає кінцевий прогноз усієї моделі.

Основним чинником ефективності такої моделі є низька кореляція між моделями окремих дерев, які тим самим захищають один одного від поодиноких помилок, за умови, що вони всі не повторюють одні й ті ж помилки, що трапляється досить рідко. Таким чином, збірний прогноз великого числа моделей є більш точним, ніж прогнози окремих моделей.

Зважаючи на цей факт, часто використовуються ансамблі моделей, що поєднують прогнози не лише однотипних підмоделей, але й зовсім різні, що також забезпечує вищу ефективність та точність кінцевої моделі.

Основним недоліком випадкового лісу є те, що велика кількість дерев може спричинити до тривалішого часу прогнозування порівняно з іншими моделями. Внаслідок цього класифікація великих наборів може бути не завжди ефективною в часі. Проте якщо час виконання не є ключовим, то він цілком компенсується вищою точністю класифікації, що є основною перевагою такого методу.

4. Експерименти

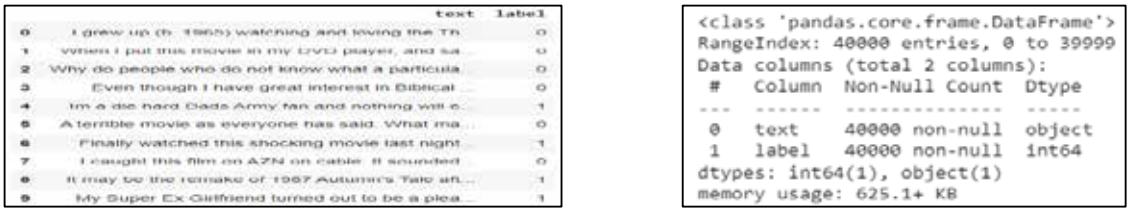
4.1.Аналіз та попередня обробка набору даних

Перед безпосереднім проведенням експериментів необхідно проаналізувати та підготувати набір даних. Для цього завантажимо його і переглянемо декілька перших записів:

Як видно з рис. 1a, у наборі даних містяться відгуки на фільми у текстовій формі (колонка text), а також на кожен відгук є визначена позначка (колонка label), яка свідчить про те, чи він позитивний (значення 1) чи негативний (значення 0). Також кожне значення є автоматично проіндексоване.

Отож, колонка text має тип object, що свідчить про вміст даних стрічкового типу, а колонка label містить значення числового типу, тобто в такому наборі всі типи даних є визначеними коректно та не вимагають приведення (рис. 1 b). Також видно, що набір даних містить у загальному 40000 записів, серед яких немає пропущених даних.

Тепер перевіримо набір даних на вміст повторювальних даних. Для цього підрахуємо кількість унікальних значень у колонці відгуків та віднімемо її від загальної кількості записів.



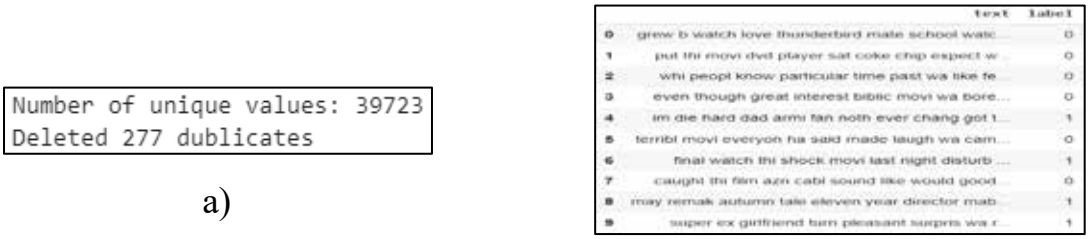
	text	label
0	I grew up (b. 1965) watching and loving the Th...	0
1	When I put this movie in my DVD player, and sa...	0
2	Why do people who do not know what a particula...	0
3	Even though I have great interest in Biblical ...	0
4	Im a die hard Dads Army fan and nothing will e...	1
5	A terrible movie as everyone has said. What ma...	0
6	Finally watched this shocking movie last night	1
7	I caught this film on AZN on cable. It sounded ...	0
8	It may be the remake of 1987 Autumn's Tale wh...	1
9	My Super Ex-Girlfriend turned out to be a plea...	1

a)

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 40000 entries, 0 to 39999
Data columns (total 2 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   text    40000 non-null  object
 1   label   40000 non-null  int64
dtypes: int64(1), object(1)
memory usage: 625.1+ KB
```

b)

Рис. 1. a) приклад запису перших 10 записів набору даних; b) інформація про вміст набору даних



```
Number of unique values: 39723
Deleted 277 dublicates
```

a)

	text	label
0	grew b watch love thunderbird mate school watc...	0
1	put thi movi dvd player sat coke chip expect w...	0
2	whi peopl know particular time past wa like fe...	0
3	even though great interest biblic movi wa bore...	0
4	im die hard dad armi fan noth ever chang got t...	1
5	terribl movi everyon ha said made laugh wa cam...	0
6	final watch thi shock movi last night disturb ...	1
7	caught thi film azn cabl sound like would good...	0
8	may remak autumn tale eleven year director mab...	1
9	super ex girlfriend turn pleasant surpris wa r...	1

b)

Рис. 2. a) виявлення та видалення дублікатів; b) набір даних після очистки від стоп-слів

Було виявлено 277 дубльованих даних, задля коректної класифікації залишимо лише унікальні значення, видаливши усі дублікати (рис. 2a).

Після цього потрібно почистити відгуки від так званих «стоп-слів» [9]. Це слова, які є найбільш вживаними під час побудови речення, проте не несуть важливої інформації для задачі класифікації через те, що вони не містять емоційного забарвлення. Такими словами переважно є займенники, сполучники, прийменники та частки. Для такої фільтрації скористаємося наперед визначеним набором таких слів з бібліотеки NLTK (Natural Language Toolkit).

На рис. 2b бачимо, що з набору даних було видалено всі стоп-слова, а також розділові знаки, оскільки вони також не дають ніякого впливу на емоційне забарвлення тексту.

Після цього проведена векторизація тексту за допомогою векторизатора з бібліотеки sklearn [10], що використовує TF-IDF метрику. Така метрика дозволяє визначити важливості слів у контексті відгуків. А саме за допомогою обрахування добутку нормалізованої частоти входження слова у відгуки та логарифму зворотної частоти відгуків, в які входить таке слово. Таким чином, більша вага буде у менш вживаних слів, а загальні слова, що частіше трапляються, будуть мати меншу вагу.

Тепер перевіримо збалансованість цільового поля (колонка label), оскільки задля коректного тренування класифікатора необхідно, щоб у кожного класу була приблизно однакова кількість записів, тоді класифікатор показує найбільш правдиву та ефективну оцінку на тестових даних. Для цього візуалізуємо розподіл цільової змінної (рис. 3).

З рис. 3a видно, що цільова змінна є практично ідеально збалансованою.

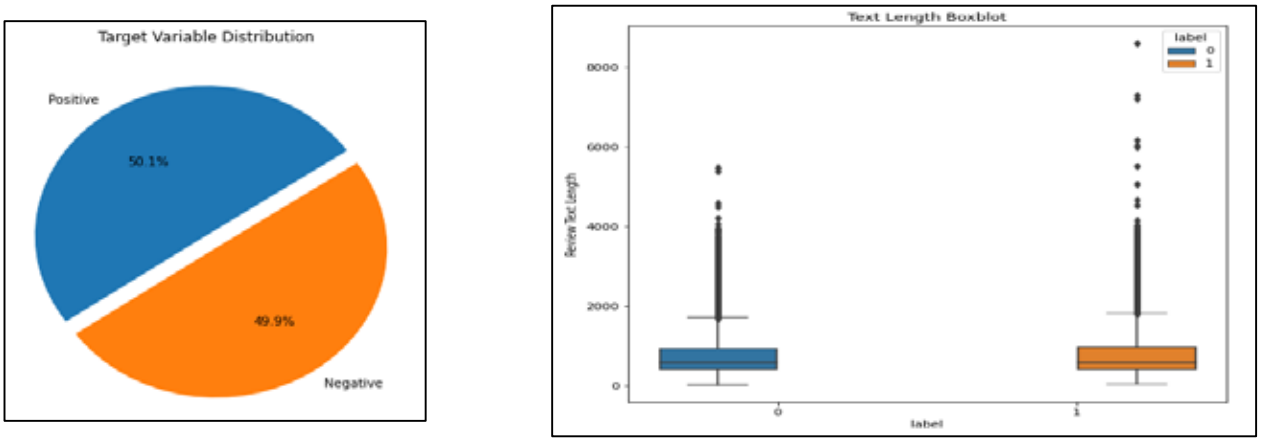
Проаналізуємо також довжини відгуків за допомогою коробчастої діаграми (рис. 3b). Бачимо, що довжини повідомлень також є рівномірно збалансованими, за винятком декількох викидів, проте вони не чинитимуть вагомої різниці саме у задачі сентимент аналізу, тому немає сенсу очищувати від них набір даних.

Після цього ми можемо перейти до аналізу самих слів. Проведемо статистичні обчислення слів, вживаних у позитивних та негативних відгуках, та побудуємо відповідні «хмари слів» з найбільш вживаними словами в кожному з класів за допомогою бібліотеки WordCloud. Хмара будується таким чином, що чим більша частота слова в певній вибірці, тим більшим шрифтом воно зображене на хмарі (рис. 4a, 4b).

4.2.Проведення експериментів

Перед безпосередньою класифікацією потрібно розділити набір даних на навчальну та тестову вибірки, і звісно ж, натренувати кожен класифікатор з певними гіперпараметрами.

Задля того щоб зробити порівняння класифікаторів, найбільш об'єктивним для їх навчання буде використаний однаковий розмір вибірок, а саме 70% даних на тренування, а 30% на проведення тестів. А також до всіх класифікаторів будуть застосовані гіпер-параметри, які були встановлені



a) b)

Рис. 3. a) розподіл цільової змінної; b) коробчаста діаграма довжин відгуків

a) b)

Рис. 4. Хмара найбільш вживаних слів: a) у позитивних відгуках; b) у негативних відгуках

за замовчуванням у бібліотеці sklearn [10]. Визначено буде лише глибину дерева рішень задля уникнення перенавчання.

Загальний алгоритм виконання класифікації виглядатиме таким чином (рис. 5):

Результати класифікації методом наївного Баєса представлені на рис. 6a, 6b.

З рис. 6a, 6b видно, що точність класифікації за допомогою метода наївного Баєса становила 84,8%, а F1-score рівний 84,5%. Також на рис. 6a, 6b зображені ROC-криві для обох класів, з яких можна побачити, що площа під кривою дорівнює 0,92, що є досить хорошим показником.

Результати класифікації методом логістичної регресії представлені на рис. 7a, 7b.

З рис. 7a, 7b видно, що точність класифікації за допомогою метода логістичної регресії дорівнює 88,4%, а F1-score рівний 88,6%. З ROC-кривих можна побачити, що площа під кривою дорівнює 0,95.

Результати класифікації методом дерева рішень представлені на рис. 8a, 8b.

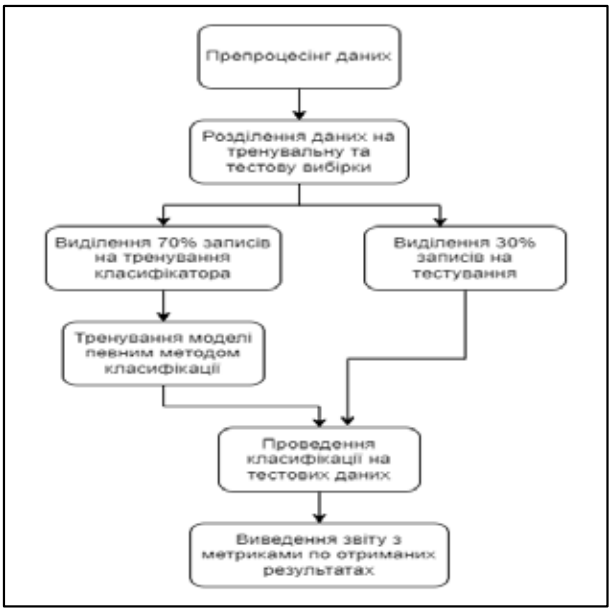


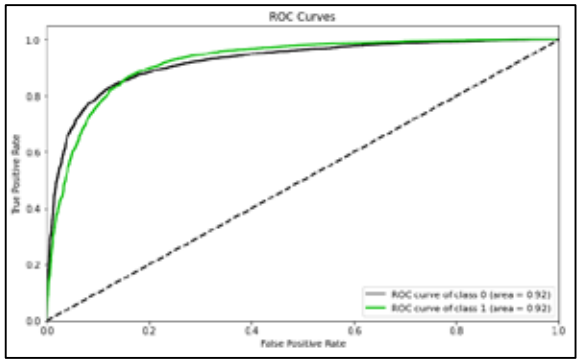
Рис. 5. Алгоритм проведення класифікації

Accuracy = 84.8%
F1 Score = 84.5%

Classification Report:

	precision	recall	f1-score	support
0	0.83	0.87	0.85	5934
1	0.86	0.83	0.85	5983
accuracy			0.85	11917
macro avg	0.85	0.85	0.85	11917
weighted avg	0.85	0.85	0.85	11917

a)



b)

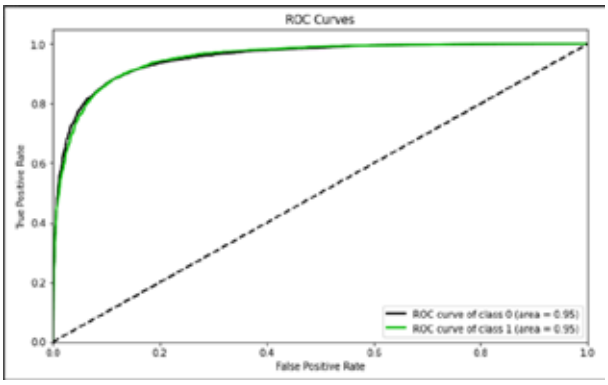
Рис. 6. Результати класифікації: a) звіт по класифікації наївним Баєсом; b) ROC-криві для класифікації наївним Баєсом

Accuracy = 88.4%
F1 Score = 88.6%

Classification Report:

	precision	recall	f1-score	support
0	0.89	0.87	0.88	5934
1	0.87	0.90	0.89	5983
accuracy			0.88	11917
macro avg	0.88	0.88	0.88	11917
weighted avg	0.88	0.88	0.88	11917

a)



b)

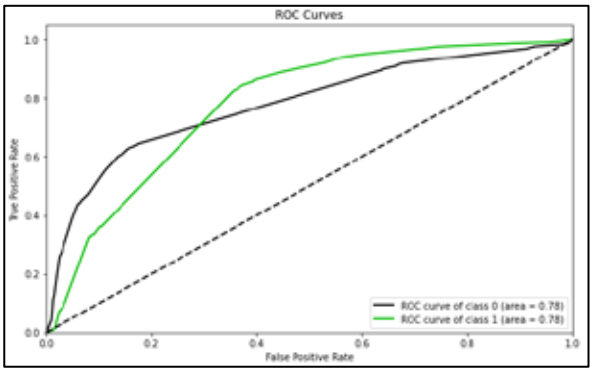
Рис. 7. Результати класифікації: a) звіт по класифікації методом логістичної регресії; b) ROC-криві для класифікації методом логістичної регресії

Accuracy = 73.6%
F1 Score = 76.1%

Classification Report:

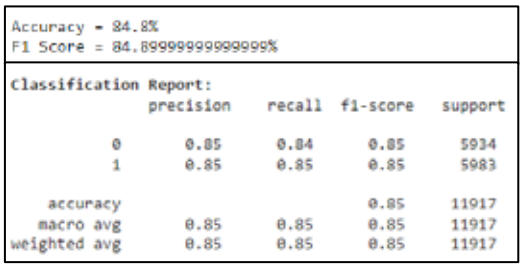
	precision	recall	f1-score	support
0	0.80	0.63	0.70	5934
1	0.70	0.84	0.76	5983
accuracy			0.74	11917
macro avg	0.75	0.74	0.73	11917
weighted avg	0.75	0.74	0.73	11917

a)



b)

Рис. 8. Результати класифікації: a) звіт по класифікації деревом рішень; b) ROC-криві для класифікації деревом рішень

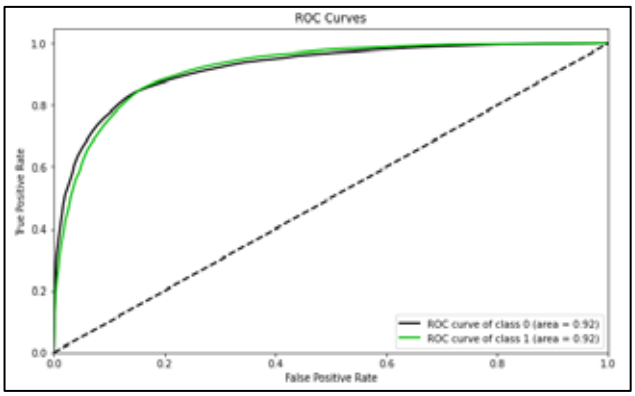


Accuracy = 84.8%
F1 Score = 84.89999999999999%

Classification Report:

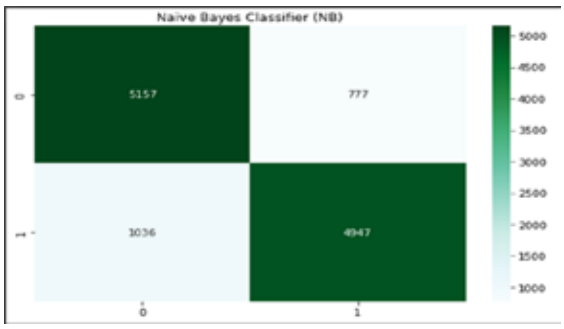
	precision	recall	f1-score	support
0	0.85	0.84	0.85	5934
1	0.85	0.85	0.85	5983
accuracy			0.85	11917
macro avg	0.85	0.85	0.85	11917
weighted avg	0.85	0.85	0.85	11917

a)

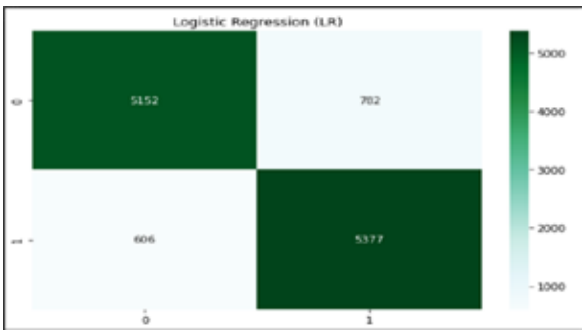


b)

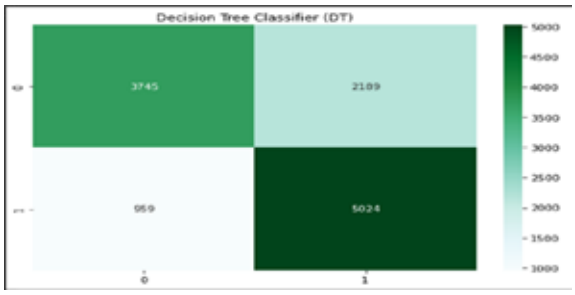
Рис. 9. Результати класифікації: a) звіт по класифікації випадковим лісом; b) ROC-криві для класифікації випадковим лісом



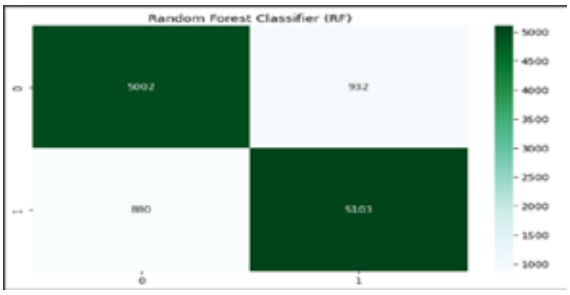
a)



b)



c)



d)

Рис. 10. Матриця невідповідностей: a) для наївного Баєса; b) для логістичної регресії; c) для дерева рішень; d) для випадкового лісу

З рис. 8a, 8b видно, що точність класифікації за допомогою дерева рішень становить 73,6%, а F1-score рівний 76,1%. З ROC-кривих можна побачити, що площа під кривою дорівнює 0,78.

Результати класифікації методом випадкового лісу представлені на рис. 9a, 9b.

З рис. 9a, 9b видно, що точність класифікації за допомогою випадкового лісу дорівнює 84,8%, а F1-score рівний 84,9%. З ROC-кривих можна побачити, що площа під кривою дорівнює 0,92.

4.3.Обговорення результатів дослідження

Для детальнішого аналізу побудуємо також матриці невідповідностей для кожного з результатів класифікації (рис. 10a, 10b, 10c, 10d):

А також виведемо загальний графік для вказаних метрик усіх класифікаторів (рис. 11):

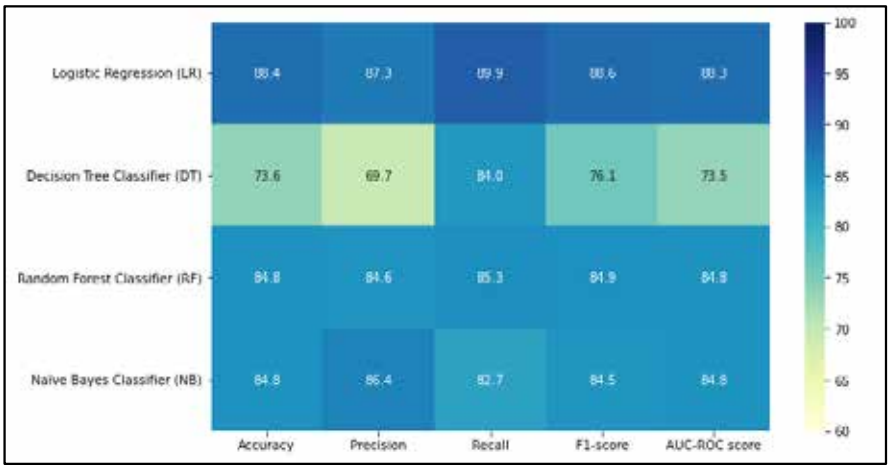


Рис. 11 Представлення результатів для описаних метрик усіх класифікаторів

Як видно з попередніх рис. 6–9, найкращий результат дав класифікатор, побудований на логістичній регресії з результуючою точністю 88,4% і 88,6% f1-score, що є дійсно дуже хорошим результатом порівняно зі всіма іншими і навіть з томностями, які згадувались для такої задачі в опублікованій літературі.

Далі по рейтингу майже на одному рівні перебувають класифікатори наївного Баєса та випадкового лісу з точністю 84,8%. З огляду на те, що час тренування та виконання класифікації випадковим лісом є в рази більшим, ніж наївним Баєсом, роблю висновок, що наївний Байєс є оптимальнішим. З огляду на це саме він і посяде друге місце в рейтингу, а третє – класифікатор випадковим лісом.

Останнє місце по точності займає одиночне дерево рішень, маючи точність 73,6%. Решта метрик цього класифікатора відрізняється від решти в околі на 15% менше. Особливо гіршу статистику цей класифікатор показав на визначенні негативних відгуків, оскільки він класифікував 2189 негативних відгуків як позитивні, що є мінімум на 1000 більше від усіх інших відхилень. Саме цей відгук зіпсував методу загальну статистику.

Висновки

У цій роботі було визначено класифікатор, який серед вибраних є найефективнішим, – це метод логістичної регресії.

Під час виконання цієї роботи були проведені аналізи актуальності задачі та наукових джерел, серед яких було досліджено точність класифікаторів наївного Баєса, логістичної регресії, дерев рішень та випадкового лісу.

Також був виконаний детальний аналіз та підготовка даних для задачі бінарної класифікації. Паралельно виконувалось безпосереднє тренування класифікаторів та проведення експериментів з кожним.

Було обговорено результати дослідження за допомогою повної статистики усіх метрик та всіх вибраних класифікаторів.

Для покращення точності класифікаторів необхідно підбирати відповідні гіперпараметри на кожен з них. Можливе проведення такого дослідження у подальших наукових роботах.

ЛІТЕРАТУРА

1. Мироненко С.С., Онищенко Є.А. Порівняльний аналіз методів для вирішення задачі сентимент аналізу тексту. *Науковий журнал «Комп'ютерно-інтегровані технології: освіта, наука, виробництво»*, 2020. URL: https://cit-journal.com.ua/index.php/cit/article/view/170/243 (дата звернення: 30.05.2022).
2. Мороз Б., Кабак Л., Ширін А., Овчаренко С. Використання Data Mining в інформаційних бібліотечних системах. *Computer-integrated technologies: education, science, production*, 42, 2021, с. 177–184. URL: http://dx.doi.org/10.36910/6775-2524-0560-2021-42-26.
3. Qiang Ye., Ziqiong Z., Law R. Sentiment classification of online reviews to travel destinations by supervised machine learning approaches. *Expert Systems with Applications*, Vol. 36. Issue 3. Part 2, 2009, pp. 6527–6535. 2009. URL: https://doi.org/10.1016/j.eswa.2008.07.035.
4. Shi H.-X., Li X.-J. A Sentiment Analysis Model for Hotel Reviews Based on Supervised Learning. *Machine Learning and Cybernetics (ICMLC),* 2011. DOI: 10.1109/ICMLC.2011.6016866.
5. Zhang Z., Ye Q., Zhang Z., Li, Y. Sentiment Classification of Internet Restaurant Reviews Written in Cantonese. *Expert Systems with Applications,* 2011. URL: https://doi.org/10.1016/j.eswa.2010.12.147.

6. Литвинов М.Г. Дослідження моделей оцінювання тонального забарвлення тексту, 2018. URL: https://openarchive.nure.ua/bitstream/document/20018/1/2018_Vesnyana_shkola_42-49_PI.pdf (дата звернення: 30.05.2022).
7. Naive Bayes classifier. Wikipedia. URL: https://en.wikipedia.org/wiki/Naive_Bayes_classifier (дата звернення: 01.02.2023).
8. What is logistic regression? IBM. URL: https://www.ibm.com/topics/logistic-regression. Дата звернення: 01.02.2023.
9. Kavita Ganesan. What are Stop Words? 2020. URL: https://kavita-ganesan.com/what-are-stop-words/#.YpxIp6hByUk (дата звернення: 15.12.2022).
10. Scikit-learn documentation. Scikit-learn. URL: https://scikit-learn.org/stable/ (дата звернення: 02.06.2022).

REFERENCES

1. Myronenko, S.S., Onyshchenko, Ye.A. (2020). Porivnialnyi analiz metodiv dlia vyrishennia zadachi sentyment analizu tekstu. Naukovyi zhurnal «Kompiuterno-intehrovani tekhnolohii: osvita, nauka, vyrobnytstvo», 2020. Retrieved from: https://cit-journal.com.ua/index.php/cit/article/view/170/243 [in Ukrainian].
2. Moroz, B., Kabak, L., Shyrin, A., Ovcharenko, S. (2021). Vykorystannia Data Mining v informatsiinykh bibliotechnykh systemakh. Computer-integrated technologies: education, science, production, 42, s. 177–184. Retrieved from: http://dx.doi.org/10.36910/6775-2524-0560-2021-42-26 [in Ukrainian].
3. Qiang Ye., Ziqiong Z., Law R. (2009). Sentiment classification of online reviews to travel destinations by supervised machine learning approaches. Expert Systems with Applications, Vol. 36. Issue 3. Part 2, s. 6527–6535. Retrieved from: https://doi.org/10.1016/j.eswa.2008.07.035 [in English].
4. Shi, H.-X., Li, X.-J. (2011). A Sentiment Analysis Model for Hotel Reviews Based on Supervised Learning. Machine Learning and Cybernetics (ICMLC). DOI: 10.1109/ICMLC.2011.6016866 [in English].
5. Zhang, Z., Ye, Q., Zhang, Z., Li, Y. (2011). Sentiment Classification of Internet Restaurant Reviews Written in Cantonese. Expert Systems with Applications. Retrieved from: https://doi.org/10.1016/j.eswa.2010.12.147 [in English].
6. Lytvynov, M.H. (2018). Doslidzhennia modelei otsiniuvannia tonalnoho zabarvlennia tekstu. Retrieved from: https://openarchive.nure.ua/bitstream/document/20018/1/2018_Vesnyana_shkola_42-49_PI.pdf [in Ukrainian].
7. Naive Bayes classifier. Wikipedia. Retrieved from: https://en.wikipedia.org/wiki/Naive_Bayes_classifier [in English].
8. What is logistic regression? IBM. Retrieved from: https://www.ibm.com/topics/logistic-regression [in English].
9. Kavita Ganesan. What are Stop Words? 2020. Retrieved from: https://kavita-ganesan.com/what-are-stop-words/#.YpxIp6hByUk [in English].
10. Scikit-learn documentation. Scikit-learn. Retrieved from: https://scikit-learn.org/stable/ [in English].