УДК 004.9

ДОСЛІДЖЕННЯ ВИКОРИСТАННЯ ТЕХНІК МОДЕЛЮВАННЯ ДЛЯ АНАЛІЗУ ВІДГУКІВ КЛІЄНТІВ

Лаптєв О. А. (ORCID https://orcid.org/0000-0002-4194-402X)[1],
Юзва А. С.[2]
[1] *Київський національний університет імені Тараса Шевченка*
[2] *Волинський національний університет імені Лесі Українки*

RESEARCH ON THE USE OF MODELING TECHNIQUES FOR ANALYZING CUSTOMER FEEDBACK

Oleksandr Laptiev[1], Yuzva Anna[2]
[1] *Taras Shevchenko National University of Kyiv*
[2] *Lesya Ukrainka Volyn National University*

Abstract. The exponential growth of social media platforms and applications on the Internet has resulted in a staggering amount of user-generated textual content, including comments and reviews. Consequently, users often face difficulties in drawing valuable conclusions or relevant information from such content. To solve this problem, machine learning and natural language processing algorithms have been applied to analyze the vast amount of textual data available on the Internet. In recent years, topic modeling techniques have gained considerable popularity in this field. In this study, we comprehensively investigate and compare five commonly used topic modeling techniques that are specifically applied to customer reviews. The methods studied are: Latent Semantic Analysis (LSA), Latent Dirichlet Allocation (LDA), Nonnegative Matrix Factorization (NMF), Pachinko Allocation Model (PAM), Top2Vec, and BERTopic. By practically demonstrating their advantages in identifying important topics, we aim to emphasize their effectiveness in real-world scenarios. To evaluate the effectiveness of these topic modeling techniques, we carefully select two text datasets. The evaluation is based on standard statistical evaluation metrics such as the topic consistency score. Our findings show that BERTopic consistently produces more meaningful topic extractions and achieves favorable results.

Keywords: customer reviews, theme modeling, natural language processing, machine learning, topic coherence, latent Dirichlet allocation, non-negative matrix factorization.

Вступ. У сучасну цифрову епоху відгуки клієнтів стали незамінним джерелом інформації як для споживачів, так і для компаній. Завдяки поширенню платформ електронної комерції, соціальних медіа та вебсайтів з оглядами в Інтернеті клієнти мають можливість ділитися своєю думкою про продукти чи послуги. Ці відгуки клієнтів дають цінну інформацію про сильні та слабкі сторони та загальну якість різних пропозицій, допомагаючи іншим потенційним клієнтам приймати зважені рішення. Однак велика кількість відгуків клієнтів, доступних в Інтернеті, є проблемою. Підприємствам часто важко вручну проаналізувати та витягти значущу інформацію з цієї величезної кількості неструктурованих даних. Ось тут і вступає в гру тематичне моделювання [1].

Моделювання тем – це потужна обчислювальна техніка, яка дозволяє компаніям розкривати основні теми, теми та шаблони в колекції відгуків клієнтів. Застосовуючи алгоритми обробки природної мови (NLP) і машинного навчання, алгоритми моделювання тем можуть автоматично ідентифікувати та витягувати ключові теми з текстових даних. Це дозволяє підприємствам отримати глибше розуміння думок, уподобань і проблем клієнтів.

Моделювання може використовуватися з відгуками клієнтів, щоб отримати уявлення про те, що клієнти говорять про ваші продукти чи послуги. Наприклад, можливо використовувати тематичне моделювання, щоб визначити різні функції, якими клієнти найбільше задоволені, або сфери, де потрібно вдосконалити власні продукти чи послуги.

Моделювання теми також можна використовувати для визначення проблемних точок клієнта. Наприклад, якщо харчова онлайн-платформа бачить, що багато клієнтів говорять про ту саму проблему у своїх відгуках, вони можуть використати цю інформацію, щоб покращити свій продукт чи послугу. Застосування тематичного моделювання до відгуків клієнтів пропонує кілька переваг. По-перше, це систематичний і ефективний спосіб аналізу великої кількості відгуків клієнтів. Замість того, щоб вручну читати та класифікувати кожен відгук, компанії можуть використовувати алгоритми моделювання тем для автоматизації процесу. Це не тільки економить час і ресурси, але й дає можливість підприємствам отримувати інформацію в такому масштабі, який інакше був би неможливий [2].

По-друге, тематичне моделювання дозволяє компаніям визначати нові тенденції та повторювані теми у відгуках клієнтів. Розкриваючи теми, які найчастіше обговорюються, компанії можуть отримати цінну інформацію про те, які аспекти їхніх продуктів чи послуг резонують із клієнтами та які сфери можуть потребувати вдосконалення. Ця інформація може інформувати процеси прийняття рішень, такі як розробка продукту, маркетингові стратегії та покращення обслуговування клієнтів. Крім того, тематичне моделювання може допомогти компаніям відстежувати та оцінювати настрої у відгуках клієнтів. Пов'язуючи показники настрою з різними темами, компанії можуть визначити, які аспекти їхніх пропозицій отримують позитивні чи негативні відгуки. Ці знання можна використати для визначення пріоритетів у сферах покращення або висвітлення позитивних рис у маркетингових кампаніях [3].

Протягом багатьох років у цій галузі проводилися великі дослідження, що призвело до розробки різних методів. Ці методи можна загалом класифікувати на три основні категорії: алгебраїчні, імовірнісні та нейронні моделі [4-5]. Алгебраїчні та ймовірнісні моделі є традиційними статистичними підходами, тоді як нейронні моделі представляють останні досягнення в цій галузі, використовуючи штучні нейронні мережі та виграючи від широкого впровадження глибокого навчання в НЛП. Алгебраїчні моделі охоплюють ряд методів, таких як латентне семантичне індексування (LSI) і невід'ємна матриця факторизації (NMF) [25-26]. Імовірнісні моделі включають ймовірнісне латентне семантичне індексування (pLSI), прикріплене кореляційне пояснення (CorEx), прихований розподіл Діріхле (LDA) і різні розширення та варіанти LDA, такі як ієрархічний процес Діріхле (HDP), корельована тематична модель (CTM) і структурна тема моделі (STM) [6-8]. LDA був найбільш широко використовуваним методом протягом десятиліть, але його застосування в дослідницьких дослідженнях часто не викликало сумнівів, насамперед через його популярність, без надання обґрунтування вибору методу тематичного моделювання [9]. LDA можна розглядати як узагальнену версію pLSI, що включає попередній розподіл Діріхле між розподілами документів за темами та словами за темами. Одним з обмежень LDA є його залежність від представлення пакета слів (BoW), яке не враховує семантичні зв'язки між словами в тексті [9].

Останнім часом, Нейронні моделі набули значної популярності з 2016 року, що збіглося зі швидким прогресом глибокого навчання. Приклади нейронних моделей включають lda2vec, SBM, deep LDA, Top2Vec і BERTopic [10-14, 26-28]. Розробка цих

моделей узгоджується з експоненціальним зростанням методів глибокого навчання. Наприклад, глибока LDA – це гібридна модель, яка поєднує LDA з базовою нейронною мережею багатошарового персептрона (MLP). На відміну від цього, останній BERTopic використовує двонаправлене представлення кодера від трансформаторів (BERT) і TF-IDF на основі класу (c-TF-IDF) для досягнення найсучаснішої продуктивності. BERTopic став домінуючим підходом у сфері тематичного моделювання, демонструючи вражаючі результати на різних наборах даних із мінімальними вимогами до попередньої обробки даних [14].

У цьому дослідженні нашою головною метою є оцінка та порівняння продуктивності п'яти методів тематичного моделювання: LDA, NMF, LSA, PAM і BERTopic. Ці методи забезпечують унікальні методології для вилучення тем із текстових даних. Прихований розподіл Діріхле (LDA) – це генеративна статистична модель, яка ідентифікує теми на основі розподілу ймовірностей слів [8]. Факторизація невід'ємної матриці (NMF) використовує підхід лінійної алгебри для розкриття прихованих тем шляхом факторизації матриці термін-документ [9]. Латентний семантичний аналіз (LSA) використовує техніку зменшення розмірності для виявлення семантичних зв'язків між словами та документами [10]. Модель розподілу Pachinko (PAM) об'єднує тематичне моделювання з визначенням авторства, щоб розкрити як тему, так і інформацію про автора [11-12]. Top2Vec поєднує в собі сильні сторони word2vec, кластеризації та вбудовування документів для ідентифікації узгоджених тем у текстових даних [13]. Нарешті, BERTopic використовує потужність вбудовування BERT (Bidirectional Encoder Representations from Transformers) для визначення зв'язних тем у тексті [14]. Вивчаючи та порівнюючи ефективність цих методів, ми прагнемо отримати уявлення про їхні відповідні сильні та слабкі сторони в програмах тематичного моделювання. Щоб оцінити продуктивність п'яти різних моделей, було розраховано показник когерентності.

Аналіз останніх досліджень і публікацій. Останніми роками спостерігається помітне збільшення досліджень, присвячених тематичному моделюванню для оглядів клієнтів. Науковці детально дослідили різні техніки, моделі та застосування, проливаючи світло на переваги та перешкоди, пов'язані з використанням тематичного моделювання для аналізу відгуків клієнтів. Ця зростаюча увага зумовлена необхідністю ефективного аналізу величезних обсягів текстових даних і виявлення основних тем і шаблонів у них. Дослідники зважилися на різноманітні методології, поєднуючи можливості машинного навчання з техніками тематичного моделювання, щоб покращити свій аналіз [1]. За останні десять років тематичне моделювання (ТМ) широко застосовувалося в різних областях, включаючи охорону здоров'я, гостинність, освіту, соціальні мережі та фінанси [15]. Його впровадження виявилося корисним як для академічних, так і для промислових цілей, особливо в міждисциплінарних дослідженнях. Однак нещодавня література вказує на тенденцію, коли більшість досліджень зосереджуються на застосуванні конкретної техніки ТМ, часто обмеженої однією технікою, у певній галузі.

У дослідженні, проведеному дослідниками, корпус вебсторінок, пов'язаних з питаннями безпеки харчових продуктів, був проаналізований за допомогою латентного розподілу Діріхле (LDA) [16]. Дослідження продемонструвало потенціал LDA як цінного інструменту для дослідників комунікацій у визначенні вебсайтів, пов'язаних з безпекою харчових продуктів. Крім того, автори провели інтелектуальний огляд літератури,

застосувавши LDA до наукових статей [17]. Створюючи теми за допомогою LDA, автори змогли вибрати відповідні теми для огляду літератури, сприяючи ефективному та цілеспрямованому процесу огляду. Інше дослідження було зосереджено на оглядах готелів у Нью-Йорку [18]. Автори застосували структурне тематичне моделювання (STM), щоб проаналізувати відгуки та підкреслили покращення висновків щодо незадоволеності споживачів завдяки використанню STM.

Дослідницька стаття використовувала STM для визначення тем дослідження з назви, ключових слів і анотації статей, опублікованих у журналі Computers & Education протягом сорока двох років [19]. Цей підхід дозволив виявити повторювані теми та тенденції в галузі. Дослідники досліджували новини, пов'язані з COVID, у Великобританії, Індії, Японії та Південній Кореї [20]. Вони також використовували алгоритм Top2Vec для виявлення широко розповсюджених тем і згодом провів аналіз настроїв за допомогою моделі RoBERTa. Дослідження було зосереджено на аналізі дописів, пов'язаних з біткойнами, у Twitter, Reddit і Bitcoin Talk. Вони досліджували LDA, щоб отримати теми з дописів, які потім використовувалися нейронною системою на основі LSTM для прогнозування курсу акцій [21].

Інше дослідження використало та порівняло чотири методи моделювання тем, включаючи LDA, NMF, Top2Vec і BERTopic, на даних соціальних мереж для досліджень соціальних наук. NMF і BERTopic продемонстрували кращу продуктивність порівняно з двома іншими методами в цьому сценарії [2]. У дослідницькій статті проаналізовано твіти, пов'язані з вакциною проти COVID-19. Для виділення тем із твітів було використано LDA, а також проведено аналіз полярності настроїв за допомогою методу на основі словника [22].

У іншому дослідженні було використано три методи моделювання тем (LDA, CorEx і NMF), щоб ідентифікувати досвід мандрівників з публікацій в Instagram за допомогою певного хештегу [23]. Нарешті, деякі автори застосували LDA для аналізу фінансових новин. Дослідження мало на меті висвітлити прогнози та спекулятивні заяви в новинних статтях за допомогою графічного інтерфейсу [24]. LDA стала переважною технікою тематичного моделювання (ТМ) в останній літературі, як свідчать дослідження [17-23]. Ці дослідження в сукупності демонструють потенціал ТМ у розкритті тем у різних сферах. Однак існує дефіцит досліджень із використанням ТМ у завданнях, пов'язаних із онлайн-рецензуванням. Недавнє дослідження, наприклад, використовувало LDA для визначення тем із коротких твітів і створення прогнозів спеціально для фінансових інструментів [24]. Ці дослідження демонструють різноманітні застосування тематичного моделювання в різних сферах і ефективність різних алгоритмів у вилученні цінної інформації з різних типів текстових даних.

У комплексному оглядовому дослідженні досліджуються різні моделі та рамки аналізу думок і настроїв, включаючи тематичне моделювання. У ньому обговорюється роль тематичного моделювання у визначенні та виділенні ключових тем із відгуків клієнтів і підкреслюється його корисність у розумінні настроїв і вподобань клієнтів. Дослідники зосереджуються на застосуванні тематичного моделювання в аналізі настроїв відгуків клієнтів. Автори пропонують новий підхід, який поєднує тематичне моделювання з аналізом настроїв, щоб виявити приховані теми та настрої у відгуках клієнтів [28]. Дослідження демонструє ефективність цього підходу у визначенні аспектів продуктів або послуг, які сприяють позитивним чи негативним настроям [29]. Інше дослідження досліджує використання тематичного режиму.

Головне дослідження. У цьому дослідженні досліджується вибір методів тематичного моделювання (ТМ) через всебічний огляд. Зокрема, ми зосереджуємось на шести широко використовуваних методах ТМ, які використовують різноманітні форми представлення та статистичні моделі. Рисунок 1 ілюструє стандартний процес створення тем, який служить основою для нашого аналізу. Наша оцінка охоплює як якість теми, так і показники ефективності. Важливо відзначити, що принципові відмінності між цими методами полягають у їхніх підходах до захоплення структур і специфічних аспектах цих структур, які вони використовують. Хоча текстові дані соціальних медіа охоплюють чисельні методи ТМ, ми вибрали найпопулярніші з метою порівняння, оскільки недоцільно згадувати кожен метод окремо.

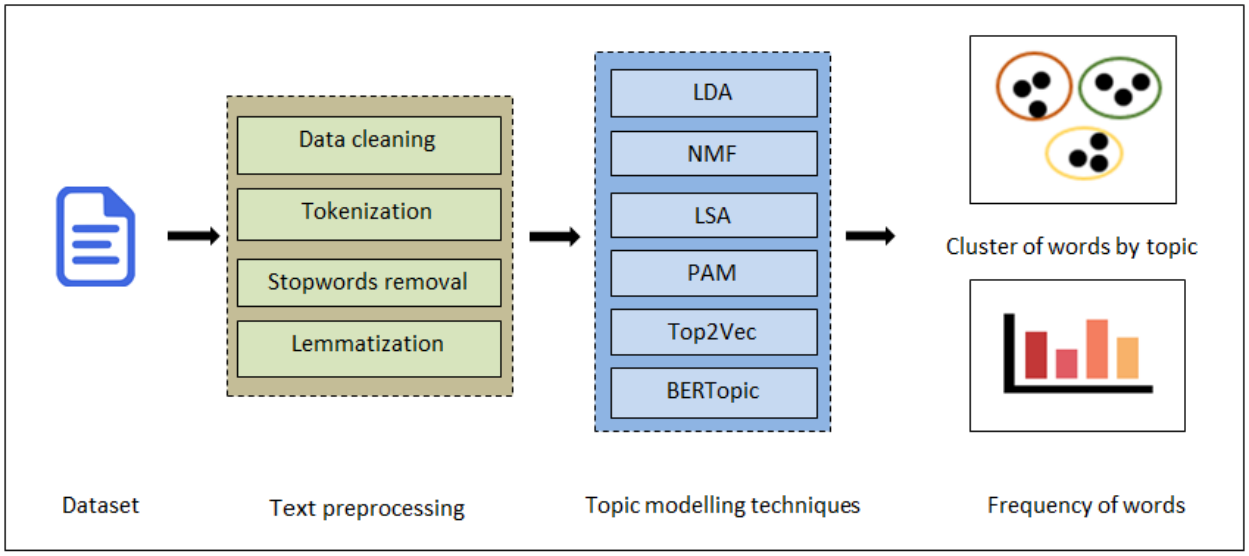

Рис. 1. Огляд схеми тематичного моделювання

Збір даних. У цьому дослідженні було використано два набори даних: набір даних про задоволеність клієнтів з урядового вебсайту Міністерства економіки ОАЕ (МОЕ) і набір даних OpSpam. Перший набір даних містить 29 200 відгуків, отриманих від Міністерства економіки ОАЕ. Ці відгуки є достовірними та відображають рівень задоволеності клієнтів додатком МОЕ. Вони дають уявлення про досвід і думки реальних користувачів щодо послуг, які пропонує уряд. Другий набір даних — це набір даних OpSpam, який є загальнодоступним у kaggle. Він містить 1600 відгуків клієнтів [32]. Ці відгуки були зібрані з таких популярних платформ, як TripAdvisor і Amazon Mechanical Turk. Набір даних служить цінним ресурсом для аналізу настроїв і завдань з аналізу думок, надаючи різноманітні огляди для аналізу та досліджень. Дослідники та практики можуть використовувати цей набір даних для вивчення думок клієнтів, тенденцій настроїв і пов'язаних тем у різних областях

Попередня обробка даних. Модуль попередньої обробки даних відіграє вирішальну роль у вдосконаленні та очищенні необроблених даних, гарантуючи, що вони містять лише необхідні функції для завдання моделювання теми. Цей модуль використовує різні методи для підвищення якості даних. Він ефективно видаляє нерелевантні фрази та символи, усуваючи будь-які шуми та відволікання від тексту. Крім того, для видалення часто

використовуваних слів у мові, які не сприяють суттєвому вирішенню завдань інтелектуального аналізу тексту, застосовується видалення стоп-слова. Ці слова, такі як прийменники, числа та інші нерелевантні терміни, не мають відповідної інформації для дослідження.

Для полегшення глибокого аналізу використовується токенізація, яка розділяє введений текст на значущі одиниці, такі як фрази, слова або лексеми. Результатом токенізації є послідовність цих токенів, які служать основними одиницями для подальшого аналізу та обробки. Крім того, для оптимізації даних застосовано лематизацію з тегами частини мови (POS). Цей процес зменшує простір функцій, зіставляючи слова з їх базовими або словниковими формами, зменшуючи надмірність і покращуючи ефективність наступного аналізу.

Прийоми моделювання теми. Моделювання тем – це широко використовувана техніка в аналізі тексту, яка спрямована на ідентифікацію кластерів слів, відомих як теми, які часто з'являються разом у корпусі документів. Ці теми виявляються за допомогою імовірнісних моделей. По суті, тема являє собою розподіл ймовірностей за всіма словами в лексиці корпусу, де найбільш імовірні слова вказують на його зміст. Для обробки вхідних документів тематичне моделювання покладається на числові вектори, наприклад вектори документ × термін. Потім ці вектори перетворюються на вектори тема × термін і документ × тема. Віднесення слів до певних тем здійснюється за допомогою матриці документ × термін. У цьому дослідженні використовуються різні методології тематичного моделювання, кожна з яких має власний підхід. Ці методології будуть коротко описані в наступних розділах.

LDA. Прихований розподіл Діріхле (LDA) є широко визнаною генеративною імовірнісною моделлю, яка використовується для виявлення прихованої тематичної інформації у великій колекції чи корпусі документів[8]. Ця модель використовує підхід пакета слів (BoW), розглядаючи кожен документ як вектор частоти слів. Роблячи це, він перетворює текстову інформацію в числові дані, які можна легко проаналізувати. LDA ефективно зменшує розмірність моделі BoW, представляючи документ як комбінацію тем [12]. Як правило, використовується кілька сотень тем, що призводить до векторного представлення документа з кількома сотнями вимірів. Таке зменшення розмірів значно прискорює навчання та мінімізує ризик переобладнання. Математично модель LDA можна підсумувати таким чином:

$$\Phi n \sim \text{Діріхле}(\beta) \quad (1)$$

$$\Theta p \sim \text{Діріхле}(\alpha) \quad (2)$$

У наведеному вище формулюванні рівняння (1) представляє розподіл слів за темою та рівняння Eq. (2) представляє розподіл тем документа. Φn позначає розподіл слів для теми n, а Θp позначає розподіл тем для документа p. Гіперпараметри моделі α і β представляють щільність тем-документів і щільність слів-тем відповідно.

LSA. Метод латентного семантичного аналізу (LSA) використовує двоетапний підхід для вивчення теми [10]. На першому кроці він створює звичайну матрицю термінів за документами, поширену техніку пошуку інформації. Щоб підвищити значущість інформативних слів, підрахунки матриці потім згладжуються. Перетворення логарифмічної ентропії застосовано для покращення оригінальної моделі LSA. На другому кроці LSA розкладає згладжену матрицю, виявляючи загальні закономірності та зв'язки між словами та документами. LSA ефективно обробляє великі обсяги необробленого тексту, розбиваючи

його на окремі слова та організовуючи у зв'язні речення чи абзаци. Він розглядає як подібність між термінами в тексті, так і їхні зв'язки з іншими термінами, що веде до глибшого розуміння основної теми. Крім того, модель LSA генерує векторні представлення для текстів, полегшуючи групування пов'язаних слів. Метод LSA вивчає приховані теми через матричну декомпозицію матриці термін-документ [10]. Припустімо, що ми маємо подокументну матрицю X, яку можна розкласти на три інші матриці: a, b і c. Перемножуючи ці матриці разом, ми отримуємо реконструйовану матрицю y, виражену у формулі. (3) як:

$$\{y\} = \{a\}\{b\}\{c\} \quad (3)$$

У цій декомпозиції стовпці представляють абзаци, тоді як рядки представляють унікальні слова.

NMF. Факторизація невід'ємної матриці (NMF) – це ще один підхід до факторизації матриці M шляхом мінімізації помилки реконструкції, але з додатковим обмеженням, що розкладені матриці містять лише невід'ємні значення [9]. У цьому сенсі NMF можна розглядати як вивчення ненормалізованих розподілів ймовірностей за темами. Формально NMF визначається як:

$$V = FB \quad (4)$$

У рівнянні (4), матриця термін-документ V розміром (m × n) представляє зв'язок між m термінами та n документами. Матриця F з розмірами k×n представляє розподіл тем за документами, де k — кількість тем. Кожен стовпець F відповідає темі та пов'язаним з нею вагам термінів. Крім того, матриця B розміром (k×n) фіксує ваги документів для кожної теми. Кожен стовпець B відповідає вазі документа для відповідної теми. Забезпечуючи невід'ємність, NMF забезпечує корисну структуру для вивчення та інтерпретації тем у різних програмах.

LDA. Модель розподілу Пачінко (PAM) – це передовий метод моделювання тем, який усуває деякі обмеження моделі латентного розподілу Діріхле (LDA). PAM призначений для охоплення ієрархічної структури тем у колекції документів. PAM вводить ієрархічну структуру в процес моделювання теми. Він представляє теми як орієнтований ациклічний граф (DAG) або структуру дерева. Ця ієрархія дозволяє моделювати зв'язки між темами на різних рівнях.

У PAM кожен документ пов'язаний із набором тем, які вибираються з різних рівнів ієрархії. Модель розглядає як теми документа, так і теми вищого рівня, які забезпечують контекст і узгодженість теми документа. Ця ієрархічна організація дозволяє PAM фіксувати складніші тематичні зв'язки та надавати точніші представлення документів. Імовірність створення цілого корпусу P(E|α) у контексті тематичного моделювання можна розрахувати як добуток імовірності для кожного окремого документа P(n|α), де α представляє гіперпараметри моделі. Математично PAM можна виразити у рівнянні. (5) як:

$$P(E|\alpha) = \Pi\, P(n|\alpha) \quad (5)$$

Тут P(n|α) представляє ймовірність створення окремого документа «n» із заданими гіперпараметрами α. Добуток Π позначає множення цих ймовірностей для всіх документів у корпусі.

Top2Vec. Top2Vec – це інноваційний підхід до неконтрольованого машинного навчання, який пропонує масштабовані та ефективні рішення для тематичного моделювання та кластеризації документів. Його основною метою є ідентифікація відповідних тем у великомасштабних текстових корпусах [13]. Top2Vec відображає як документи, так і слова в спільному семантичному векторному просторі за допомогою методу Doc2Vec. Вектори

документа згодом кластеризуються, у результаті чого утворюється кілька кластерів, кожен з яких представляє окрему тему. Представлення теми в кластері визначається шляхом усереднення векторів документів у цьому кластері та вилучення найближчих N слів до вектора теми. Однією з примітних особливостей Top2Vec є його здатність працювати з багатослівними фразами та рідко вживаними термінами, що відрізняє його від традиційних технік тематичного моделювання.

Процес Top2Vec складається з кількох кроків. По-перше, вбудовані вектори та слова генеруються за допомогою таких методів, як Doc2Vec. Далі зменшується розмірність векторів вбудовування, часто за допомогою таких методів, як UMAP. Згодом виконується кластеризація скорочених векторів, як правило, за допомогою HDBSCAN. Потім обчислюються центроїди отриманих кластерів, що представляють різні теми. Вектор кожної теми виходить шляхом усереднення векторів документа в межах відповідного кластера. Нарешті, слова, тісно пов'язані з вектором кожного кластера, призначаються відповідній темі.

Загалом Top2Vec пропонує новий підхід до масштабованого тематичного моделювання та кластеризації документів. Використовуючи семантичну подібність вбудовування слів та ієрархічну кластеризацію, це забезпечує ефективне рішення для вилучення значущих тем із великих текстових корпусів. Його здатність працювати з багатослівними фразами та рідкісними термінами ще більше розширює його можливості, відрізняючи його від традиційних методів тематичного моделювання.

BERTopic. BERTopic – це розширена попередньо навчена техніка моделювання тем, яка використовує BERT і c-TF-IDF для створення щільних кластерів, що дозволяє інтерпретувати теми, зберігаючи важливі слова в описах тем [14]. На відміну від традиційних підходів до тематичного моделювання, BERTopic використовує контекстуалізовані вбудовування слів, надані BERT, що дозволяє фіксувати семантичне значення та контекст слів у корпусі. Крім того, BERTopic забезпечує зручний інтерфейс, який дозволяє дослідникам спостерігати та аналізувати результати процесу моделювання теми.

Подібно до Top2Vec, BERTopic включає кілька кроків. Він починається з вбудовування документів, а потім зменшення розмірності за допомогою таких методів, як UMAP. Далі виконується кластеризація за допомогою таких методів, як HDBSCAN, в результаті чого утворюються кластери. З цих кластерів генеруються представлення тем. Однак те, що відрізняє BERTopic, це включення c-TF-IDF на останньому етапі. Ця техніка використовується для виділення тематичних слів, зменшення кількості тем і підвищення зв'язності та різноманітності слів за допомогою максимальної граничної релевантності (MMR).

Метрика оцінки. Оцінка тематичних моделей не має універсального підходу. Як зазначено в дослідженні [7], показники когерентності вважаються найбільш відповідним показником, коли результати тематичного моделювання використовуються користувачами. Для оцінки доступні різні показники когерентності, наприклад c_v і u_mass. Обидві метрики когерентності (c_v і u_mass) оцінюють когерентність теми шляхом обчислення суми балів схожості попарного розподілу серед слів у наборі тем, TS[34]. Загалом це можна виразити у формулі. (6) як:

$$\text{Когерентність (TS)} = \Sigma(wi, wj, \varepsilon) \quad (6)$$

Тут TS представляє набір слів, що описують тему (wi, wj), а ε означає коефіцієнт згладжування, який гарантує, що оцінка узгодженості повертає дійсні числа. Оцінка c_v створює вектори вмісту на основі спільного входження слів і обчислює оцінку за допомогою косинусної подібності та поточкової взаємної інформації (PMI). З іншого боку, метрика u_mass фокусується на аналізі розподілу слів у темі, щоб визначити його зв'язність. Як правило, вищі показники узгодженості вказують на кращі теми. Однак важливо зазначити, що сама по собі узгодженість може не забезпечити ідеальний вимір, і для доповнення показників узгодженості часто необхідна людська перевірка результатів для інтерпретації. Врахувати ефективність, особливо під час роботи з великі набори документів, ми також оцінили час обчислення або час навчання кожної моделі як показник ефективності в цьому експерименті.

Результати дослідження. У цьому розділі ми представляємо експериментальну установку та результати нашого дослідження, зосереджуючись на порівнянні результатів різних підходів до тематичного моделювання (ТМ). Мета цих експериментів полягає в тому, щоб забезпечити комплексний аналіз результатів, отриманих від кожного методу ТМ у заданому контексті. Ми заглиблюємося в деталі, виділяючи ключові висновки та ідеї, отримані в результаті наших порівнянь. Вивчаючи результати цих експериментів, ми отримуємо глибше розуміння продуктивності та ефективності різних методів ТМ.

Експериментальне дослідження. У цьому розділі ми очистили необроблений набір даних у структурований формат. Для підготовки даних було реалізовано процес очищення, який передбачав видалення зайвих символів, цифр, стоп-слів і символів. Крім того, було виконано лематизацію за допомогою тегів POS, щоб забезпечити виділення осмислених речень. На наступному кроці ми застосували техніку втручання в тему до цих очищених даних.

На етапі вилучення даних нашою метою було витягти теми з вхідних даних. Ми провели кілька оцінок, змінюючи кількість тем (t = 5, 10). Початкові результати ефективності та узгодженості теми, проаналізовані за допомогою загальних стандартних показників, застосовних до методів тематичного моделювання (ТМ), представлені в табл. 1.

Таблиця 1. Порівняння продуктивності різних методів тематичного моделювання в наборі даних 1

Topic modeling techniques	**Coherence Score**	
	K=5	**K = 10**
LDA	0.45	0.40
NMF	0.49	0.50
LSA	0.50	0.53
PAM	0.49	0.44
Top2Vec	0.56	0.54
BERTopic	0.62	0.56

BERTopic продемонструвала найкращу ефективність щодо когерентності та інтерпретації при застосуванні до набору даних 1. Вона перевершила всі інші тематичні моделі, досягнувши найвищих балів когерентності (c_v = 0,62, u_mass = -1,156). LSA мав дещо нижчі показники когерентності (c_v = 0,56) порівняно з Top2vec (c_v = 0,56). Проте LDA дав найнижчий бал когерентності (c_v = 0,40). PAM і NMF показали помірні

результати та досягли помірних показників когерентності. Табл. 2 демонструє 5 найпопулярніших ключових слів, згенерованих BERTopic. Примітно, що BERTopic відзначився у створенні описових тем порівняно з іншими методами.

Таблиця 2. Приклади тем, створені BERTopic (найкращі 5 тем) для набору даних 1

Topics	Top 5 keywords
0	Thanks, great, fast, approval, easy
1	Excellent, immediate, support, thing, pay
2	Good, bad, thank, application, time
3	Happy, thanks, happiness, user, satisfied
4	Excellent, service, job, keep, comment

Рис. 2. Хмара слів для тем, згенерованих BERTopic на основі набору даних 1

На рис. 2 показано хмарні графіки слів для тем, створених BERTopic. Хмара слів служить інструментом візуального представлення тем, виділяючи найбільш відповідні слова за допомогою використання яскравих кольорів і більшого тексту. Слова «дякую» і «добре» повторюються і мають найбільший вплив на виявлення відгуку клієнта. Щоб отримати всебічне розуміння значущості кожного терміну, генерація хмари слів може запропонувати візуально покращене представлення та показано на малюнку 2. Ця техніка візуалізації допомагає надати огляд важливості та частоти різних термінів у певному контексті.

Обговорення. Після оцінки наших шести моделей стає очевидним, що BERTopic перевершує решту за показниками когерентності. Розглядаючи набір даних 1, який складається з коротких текстових коментарів, LSA, NMF і Top2Vec працюють чудово, даючи оцінку когерентності вище 0,50. Однак при роботі з набором даних 2, що містить довші коментарі до відгуків про готелі, показник узгодженості падає нижче 0,35. Все ж таки, BERTopic вдалося зберегти свою перевагу в обох наборах даних, досягнувши вражаючого показника узгодженості вище 0,56.

Згідно з нашими спостереженнями в табл. 1, кожен застосований метод ТМ демонстрував свої сильні та слабкі сторони. Під час оцінювання результати всіх методів продемонстрували однакову ефективність. Загалом BERTopic дав найвищу ймовірність термін-тема, тоді як моделі Top2Vec, NMF і LSA продемонстрували порівнянну ефективність. З іншого боку, показники когерентності LDA були порівняно нижчими, ніж інші методи в наборі даних 1. Однак у наборі даних 2 LDA показали хороші результати. NMF і PAM були нижчими, ніж інші моделі в наборі даних 2. В обох наборах даних LSA працював порівняно добре. Імовірності термінів-тем варіюються від 0 до 1 для всіх оцінюваних методів ТМ. Однак слід зазначити, що певні звичайні методи генерували незначущі слова, включно з предметно-специфічними стоп-словами, які були непридатними для подальшої обробки. Загалом BERTopic підтримував послідовну узгодженість для обох наборів даних у різних налаштуваннях параметрів, що свідчить про те, що такі параметри, як кількість тем і максимальна кількість ітерацій, мали обмежений вплив на узгодженість.

З точки зору ефективності обчислень, BERTopic перевершив Top2Vec у цьому сценарії. Традиційні методи, такі як LDA, LSA, NMF і PAM, мали час обчислення, який залежав від кількості тем, що робило їх повільніше порівняно з BERTopic і Top2Vec для аналогічної кількості тем (500). Однак LDA міг би працювати швидше з меншою кількістю тем. При виборі тематичної моделі важливо враховувати компроміс між узгодженістю та використанням ресурсів. Що стосується підготовки даних, звичайні тематичні методи (LDA, LSA, NMF, PAM) вимагали складніших етапів попередньої обробки, включаючи видалення знаків пунктуації, непотрібних символів, стоп-слів і нормалізації тексту. Навпаки, Top2Vec і BERTopic вимагали мінімальної попередньої обробки даних, оскільки їхні базові моделі могли зрозуміти контекст із вихідною текстовою структурою. Однак для невеликих вибірок даних Top2Vec і BERTopic можуть створити теми з деякими стоп-словами. Цю проблему було пом'якшено в сценарії аналізу впливу новин, де документи новин були достатньо довгими.

Що стосується тонкого налаштування параметрів, і Top2Vec, і BERTopic використовують HDBSCAN для кластеризації, що не дозволяє безпосередньо вказувати кількість тем за замовчуванням. Однак BERTopic надає більше можливостей налаштування, показуючи налаштування параметрів основних компонентів, таких як UMAP і HDBSCAN. Навпаки, звичайні методи моделювання тем (LDA, LSA, NMF, PAM) дозволяють користувачам гнучко визначати певні параметри, включаючи кількість тем. Беручи до уваги всі ці фактори, BERTopic став найкращим вибором для аналізу впливу як довгих, так і коротких текстових даних на аналіз відгуків клієнтів у цьому дослідженні. Його чудова когерентність, можливість інтерпретації та обчислювальна ефективність, а також переваги мінімальної попередньої обробки даних і налаштування параметрів роблять його надійним підходом до тематичного моделювання.

Висновки. Інтернет відіграє вирішальну роль у стимулюванні попиту на бізнес-додатки та послуги, які покращують досвід покупок і комерційну діяльність для клієнтів у всьому світі. Однак величезна кількість інформації та знань, доступних в Інтернеті, іноді може перевантажити користувачів, що призведе до додаткового часу та зусиль, витрачених на пошук актуальної інформації. Зростання онлайн-платформ, таких як Twitter, Facebook, Instagram, підкреслило потребу в аналізі відгуків клієнтів, які створюють проблеми через їх обмеженість і галасливість, що часто призводить до неточного висновку щодо теми.

У цьому дослідницькому дослідженні представлено порівняльний аналіз шести основних методів тематичного моделювання, а саме LDA, LSA, PAM, NMF і двох сучасних нейронних моделей Top2Vec і BERTopic, у контексті аналізу відгуків клієнтів на онлайн-платформах соціальних мереж. Мета полягає в тому, щоб оцінити продуктивність цих тематичних моделей на коротких текстових даних і вивчити їх інтеграцію в сценарій аналізу відгуків клієнтів. Експериментальні результати демонструють, що BERTopic є найефективнішою моделлю в цілому, вимагаючи мінімальної попередньої обробки даних, досягаючи найвищої оцінки узгодженості та демонструючи прийнятний обчислювальний час.

Однак це дослідження визнає кілька обмежень, які відкривають потенційні шляхи для майбутніх досліджень. У цьому дослідженні шість моделей (LDA, LSA, PAM, NMF, Top2Vec і BERTopic) порівнювалися з точки зору когерентності, інтерпретації та часу обчислення. Майбутні дослідження можуть передбачати порівняння додаткових тематичних моделей. Крім того, оцінці тематичних моделей бракує консенсусу щодо відповідних заходів, які слід застосовувати. Узгодженість і інтерпретація не завжди можуть бути найбільш релевантними показниками оцінки для різних програм. У певних контекстах перевагу можуть мати інші заходи, такі як різноманітність тем. Необхідні додаткові дослідження для вивчення та аналізу сильних і слабких сторін різних показників оцінки. Крім того, дослідження потенціалу великих мовних моделей (LLM) як альтернативи звичайним тематичним моделям є багатообіцяючим напрямком, що потребує комплексної оцінки на основі встановлених методів тематичного моделювання.

Бібліографія

1. R. Albalawi, T. H. Yeap, and M. Benyoucef, "Using Topic Modeling Methods for Short-Text Data: A Comparative Analysis," Frontiers in Artificial Intelligence, vol. 3, Jul. 2020, doi: https://doi.org/10.3389/frai.2020.00042.
2. R. Egger and J. Yu, "A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts," Frontiers in Sociology, vol. 7, May 2022, doi: https://doi.org/10.3389/fsoc.2022.886498.
3. G. Papadia, M. Pacella, M. Perrone, and V. Giliberti, "A Comparison of Different Topic Modeling Methods through a Real Case Study of Italian Customer Care," Algorithms, vol. 16, no.2, pp.94, 2023.
4. A. Abdelrazek, Y. Eid, E. Gawish, W. Medhat and A. Hassan, "Topic modeling algorithms and applications: A survey," Information Systems, p.102131, 2022.
5. R.J. Gallagher, K. Reing, D. Kale and G. Ver Steeg, "Anchored correlation explanation: Topic modeling with minimal domain knowledge," Transactions of the Association for Computational Linguistics, 5, pp.529-542, 2017.
6. D.M. Blei, A.Y. Ng and M.I. Jordan, "Latent dirichlet allocation, Journal of machine Learning research," 3(Jan):993-1022, 2003.
7. W. Li and A. McCallum, "Pachinko allocation: Scalable mixture models of topic correlation," J. of Machine Learning Research, Submitted, 2008. [12] A. Afaq, G. Loveleen and S. Gurmeet, "A Latent Dirichlet Allocation Technique for Opinion Mining of Online Reviews of Global Chain Hotels", In 2022 3rd International Conference on Intelligent Engineering and Management (ICIEM), pp. 201-206, IEEE, 2022.
8. D. Angelov, "Top2vec: Distributed representations of topics," arXiv preprint arXiv:2008.09470, 2020.
9. M. Grootendorst, "BERTopic: Neural topic modeling with a class-based TF-IDF procedure," arXiv preprint arXiv:2203.05794, 2022.
10. G. Bonifazi, E. Corradini, D. Ursino and L. Virgili, "Defining user spectra to classify Ethereum users based on their behavior." Journal of Big Data, 9(1), pp.1-39, 2022.
11. Лукова-Чуйко Н.В., Лаптєв О.А., Барабаш О.В., Мусієнко А.П., Ахрамович В.М. Метод розрахунку захисту персональних даних з урахуванням комплексу специфічних параметрів соціальних мереж.

Збірник наукових праць Військового інституту Київського національного університету імені Тараса Шевченка. Київ: ВІКНУ, 2022. № 76. С. 54 – 68. https://doi.org/10.17721/2519-481X/2022/76-05.

12. Беркман Л.Н., Барабаш О.В., Ткаченко О.М., Мусієнко А.П., Лаптєв О.А., Свинчук О.В. Інтелектуальна система управління для інфокомунікаційних мереж. Системи управління навігації і зв'язку. Том 3. №69. 2022. С 54–59. https://doi.org/10.26906/SUNZ.2022.3.
13. Лаптєв О.А., Бучик С.С., Савченко В.А., Наконечний В.С. , Михальчук І.І, Шестак Я.В., Виявлення та блокування повільних ddos-атак за допомогою прогнозування поведінки користувача. Наукоємні технології. Інформаційні технології, кібербезпека. Том 55 № 3 (2022) стр.184-192. DOI: https://doi.org/10.18372/2310-5461.55.16908.
14. Serhii Yevseiev, Khazail Rzayev, Oleksandr Laptiev, Ruslan Hasanov, Oleksandr Milov, Bahar Asgarova, Jale Camalova, Serhii Pohasii. Development of a hardware cryptosystem based on a random number generator with two types of entropy sources. Eastern-European journal of enterprise technologies. Vol.5№9 (119), 2022 pp. 6–16. ISSN (print) 1729 - 3774. ISSN (on-line) 1729-4061. DOI: https://doi.org/10.15587/1729-4061.2022.265774.
15. Олександр Лаптєв, Віталій Савченко, Віталій Пономаренко, Сергій Копитко, Іван Пархоменко. Удосконалення методу підвищення завадостійкості систем виявлення сигналів засобів негласного здобуття інформації. Захист інформації. Том 24 № 3 (2022): Захист інформації. С.128-136. https://jrnl.nau.edu.ua/index.php/ZI/issue/view/906.
16. Лаптєв, О. і Гришанович, Т. Комплексна методика оцінювання ефективності функціонування системи дистанційного навчання. Прикладні проблеми комп'ютерних наук, безпеки та математики. Волинський національний університет імені Лесі Українки, Луцьк. №1 (2023). 2023.С.63–75.
17. Laptiev, O.,Sobchuk, V.,Subach, I.,Barabash, A.,Salanda, I. The Method of Detecting Radio Signals Using the Approximation of Spectral Function. CEUR Workshop Proceedings, 2022, 3384, pp. 52–61.
18. Valentyn Sobchuk , Iryna Zelenska and Oleksandr Laptiev. Algorithm for solution of systems of singularly perturbed differential equations with a differential turning point. Bulletin of the Polish Academy of Sciences Technical Sciences, Vol.71, No 3, 2023, Article number: e145682. DOI: https://doi.org/10.24425/bpasts.2023.145682.
19. Barabash, O., Sobchuk, V., Musienko, A., Laptiev, O., Bohomia, V., Kopytko, S. (2023). System Analysis and Method of Ensuring Functional Sustainability of the Information System of a Critical Infrastructure Object. In: Zgurovsky, M., Pankratova, N. (eds) System Analysis and Artificial Intelligence . Studies in Computational Intelligence, vol 1107. P. 177-192, Springer, Cham. https://doi.org/10.1007/978-3-031-37450-0_11.
20. Savchenko V., Sobchuk A., Korzh A., Laptiev O., Barabash A. Assessment of the efficiency of protection of the information system of enterprises. Telecommunications and Information Technologies. 2023. No. 2 (79). P. 63-75. DOI: https://doi.org/10.31673/2412-4338.2022.026375.
21. Ю. В. Щербина, Н. Ф. Казакова, О. О. Фразе-Фразенко, О. А. Лаптєв, А. В. Собчук.Вибір джерела випадковості для комп'ютерного моделювання. Наукоємні технології. Інформаційні технології, кібербезпека. Том 59 № 3 (2023) стр.233-238. DOI: https://doi.org/10.18372/2310-5461.59.17944.

References

1. R. Albalawi, T. H. Yeap, and M. Benyoucef, "Using Topic Modeling Methods for Short-Text Data: A Comparative Analysis," Frontiers in Artificial Intelligence, vol. 3, Jul. 2020, doi: https://doi.org/10.3389/frai.2020.00042.
2. R. Egger and J. Yu, "A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts," Frontiers in Sociology, vol. 7, May 2022, doi: https://doi.org/10.3389/fsoc.2022.886498.
3. G. Papadia, M. Pacella, M. Perrone, and V. Giliberti, "A Comparison of Different Topic Modeling Methods through a Real Case Study of Italian Customer Care," Algorithms, vol. 16, no.2, pp.94, 2023.
4. Abdelrazek, Y. Eid, E. Gawish, W. Medhat and A. Hassan, "Topic modeling algorithms and applications: A survey," Information Systems, p.102131, 2022.
5. R.J. Gallagher, K. Reing, D. Kale and G. Ver Steeg, "Anchored correlation explanation: Topic modeling with minimal domain knowledge," Transactions of the Association for Computational Linguistics, 5, pp.529-542, 2017.

6. D.M. Blei, A.Y. Ng and M.I. Jordan, "Latent dirichlet allocation, Journal of machine Learning research," 3(Jan):993-1022, 2003.
7. W. Li and A. McCallum, "Pachinko allocation: Scalable mixture models of topic correlation," J. of Machine Learning Research, Submitted, 2008. [12] A. Afaq, G. Loveleen and S. Gurmeet, "A Latent Dirichlet Allocation Technique for Opinion Mining of Online Reviews of Global Chain Hotels", In 2022 3rd International Conference on Intelligent Engineering and Management (ICIEM), pp. 201-206, IEEE, 2022.
8. D. Angelov, "Top2vec: Distributed representations of topics," arXiv preprint arXiv:2008.09470, 2020.
9. M. Grootendorst, "BERTopic: Neural topic modeling with a class-based TF-IDF procedure," arXiv preprint arXiv:2203.05794, 2022.
10. G. Bonifazi, E. Corradini, D. Ursino and L. Virgili, "Defining user spectra to classify Ethereum users based on their behavior." Journal of Big Data, 9(1), pp.1-39, 2022.
11. Lukova-Chuiko H., Laptev O., Barabash O., Musienko A., & Akhramovich B. The method of calculation of personal data protection on the basis of a set of specific parameters of social networks. *Collection of Scientific Works of the Military Institute of Kyiv National Taras Shevchenko University*, (76), 54–68. https://doi.org/10.17721/2519-481X/2022/76-05.
12. Berkman Lubov Intelligent control system for infocomunication networks / Lubov Berkman, Oleh Barabash, Andrii Musienko, Olga Tkachenko, Oleksandr Laptiev, Olha Svynchuk // Control, Navigation and Communication Systems. Academic Journal. – Poltava: PNTU, 2022. – VOL. 3 (69). – PP. 54-59. – doi: https://doi.org/10.26906/SUNZ.2022.3.054.
13. Laptiev O., Buchyk S., Savchenko V., Nakonechnyy V., Mikhalchuk I., Shestak Y. Detection and blocking slow ddos attacks based on predicting user behavior. Science-based technologies. Information technologies, cybersecurity Vol 55 № 3 (2022) p.184-192. DOI: https://doi.org/10.18372/2310-5461.55.16908.
14. Serhii Yevseiev, Khazail Rzayev, Oleksandr Laptiev, Ruslan Hasanov, Oleksandr Milov, Bahar Asgarova, Jale Camalova, Serhii Pohasii. Development of a hardware cryptosystem based on a random number generator with two types of entropy sources. Eastern-European journal of enterprise technologies. Vol.5№9 (119), 2022 pp. 6–16. ISSN (print) 1729 - 3774. ISSN (on-line) 1729-4061. DOI: https://doi.org/10.15587/1729-4061.2022.265774.
15. Oleksandr Laptiev, Vitalii Savchenko, Serhii Kopytko, Vitaliy Ponomarenko, Ivan Parkhomenko. Improvement of the method of increase the interruption resistance of the signals detection systems of the means of covert information collection. Ukrainian Information Security Research Journal. Vol 24 № 3 (2022). P.128-136. https://jrnl.nau.edu.ua/index.php/ZI/issue/view/906.
16. Laptev, O. and Hryshanovych. Comprehensive methodology for assessing the effectiveness of the distance learning system. *Applied Problems of Computer Science, Security and Mathematics*. №1, 2023, 63–75.
17. Laptiev, O.,Sobchuk, V.,Subach, I.,Barabash, A.,Salanda, I. The Method of Detecting Radio Signals Using the Approximation of Spectral Function. CEUR Workshop Proceedings, 2022, 3384, pp. 52–61.
18. Valentyn Sobchuk , Iryna Zelenska and Oleksandr Laptiev. Algorithm for solution of systems of singularly perturbed differential equations with a differential turning point. Bulletin of the Polish Academy of Sciences Technical Sciences, Vol.71, No 3, 2023, Article number: e145682. DOI: https://doi.org/10.24425/bpasts.2023.145682.
19. Barabash, O., Sobchuk, V., Musienko, A., Laptiev, O., Bohomia, V., Kopytko, S. (2023). System Analysis and Method of Ensuring Functional Sustainability of the Information System of a Critical Infrastructure Object. In: Zgurovsky, M., Pankratova, N. (eds) System Analysis and Artificial Intelligence . Studies in Computational Intelligence, vol 1107. P. 177-192, Springer, Cham. https://doi.org/10.1007/978-3-031-37450-0_11.
20. Savchenko V., Sobchuk A., Korzh A., Laptiev O., Barabash A. Assessment of the efficiency of protection of the information system of enterprises. Telecommunications and Information Technologies. 2023. No. 2 (79). P. 63-75. DOI: https://doi.org/10.31673/2412-4338.2022.026375.
21. Yu. V. Shcherbyna, N. F. Kazakova, O. O. Frazhe-Frazhenko, O. A. Laptiev, A. V. Sobchuk. Detection and blocking slow ddos attacks based on predicting user behavior. Science-based technologies. Information technologies, cybersecurity Vol 59 № 3 (2023) p.233-238. DOI: https://doi.org/10.18372/2310-5461.55.16908.