УДК 004.8

Віт Р.В., Мазурець О.В. *(Хмельницький національний університет, м. Хмельницький, Україна)*

ТЕМАТИЧНА КЛАСИФІКАЦІЯ ТЕКСТОВОЇ ІНФОРМАЦІЇ ЗАСОБАМИ ОБРОБКИ ПРИРОДНОЇ МОВИ

Анотація: *Запропоновано підхід до тематичної класифікації текстової інформації засобами обробки природної мови для автоматизованого ідентифікування та групування текстів за основними темами. Цей підхід сприяє ефективній організації збереження і використання великих обсягів текстової інформації, організовуючи структурований доступ до її змісту. Проведена крос-валідація продемонструвала результат точності 0.86, що на 0.15 перевищує точність, отриману при використанні LDA без додаткових модифікацій для класифікації.*

Ключові слова: *тематична класифікація, обробка природної мови, LDA, ідентифікування текстів, групування текстів.*

Abstract: Approach for thematic classification of text information using natural language processing tools for automated identification and grouping of texts by main topics is proposed. This approach contributes to the effective organization of storage and use of large volumes of text information, organizing structured access to its content. The cross-validation demonstrated an accuracy result of 0.86, which is 0.15 higher than the accuracy obtained when using LDA without additional modifications for classification.

Keywords: *thematic classification, natural language processing, LDA, texts identification, texts grouping.*

Постановка проблеми

Тематична класифікація текстів є поширеним підходом до обробки та аналізу неструктурованих і напівструктурованих даних в організаціях [1]. Цей процес полягає у групуванні текстової інформації за певними категоріями чи темами, що дозволяє виявляти ключові ідеї, тенденції та шаблони в даних.

Застосування алгоритмів машинного навчання дає змогу автоматизувати аналіз текстів, використовуючи контекстуальні ознаки, що значно підвищує швидкість і точність класифікації.

Аналіз останніх публікацій

Сучасні огляди часто фокусуються на імовірнісних підходах до тематичного моделювання, але важливо також враховувати методи, які базуються на лінійній алгебрі, оскільки вони здатні ефективно представляти тематичну структуру текстів [2].

У сфері тематичного аналізу текстів, заснованого на машинному навчанні, проведено чимало досліджень, орієнтованих на виявлення ключових слів і фраз, а також на формування n-грам за критерієм релевантності. Наприклад, одне з досліджень використало попередньо навчену модель BERT NLP від SberDevices, адаптовану до російськомовних текстів. Результати свідчать, що цей підхід ефективний для аналізу текстів, якщо тематика добре репрезентована у навчальному наборі даних [3].

Ще одне дослідження вивчало вплив природних криз на функціонування ланцюгів постачання, використовуючи дані соціальних мереж. Для цього була розроблена структура, що дозволяє автоматично оцінювати вплив криз, таких як пандемія COVID-19. Використовуючи аналіз термінів спільного входження та побудову карти знань, дослідники проаналізували 1024 онлайн-звіти. Було визначено п’ять ключових напрямів впливу на ланцюги постачання: роздрібна торгівля продуктами, харчові послуги, виробництво, поведінка споживачів та логістика. Ця модель стала ефективним інструментом для підтримки прийняття рішень у кризових умовах [4].

Мета роботи та постановка завдань

Метою роботи є розробка методу тематичної класифікації текстової інформації засобами обробки природної мови, здатного підвищити точність і релевантність тематичного аналізу, що сприятиме прийняттю обґрунтованих рішень на основі текстових даних.

Виклад основного матеріалу

Метод тематичної класифікації текстової інформації засобами обробки природної мови дає змогу перетворювати вхідні текстові дані у вивід у вигляді кількості тем, домінуючої теми

кожного документа та розширеного списку ключових слів для кожної з тем (рис. 1). Розроблений підхід поєднує гнучкість тематичного моделювання з можливістю автоматичного розширення ключових слів, забезпечуючи ефективний тематичний аналіз текстів.

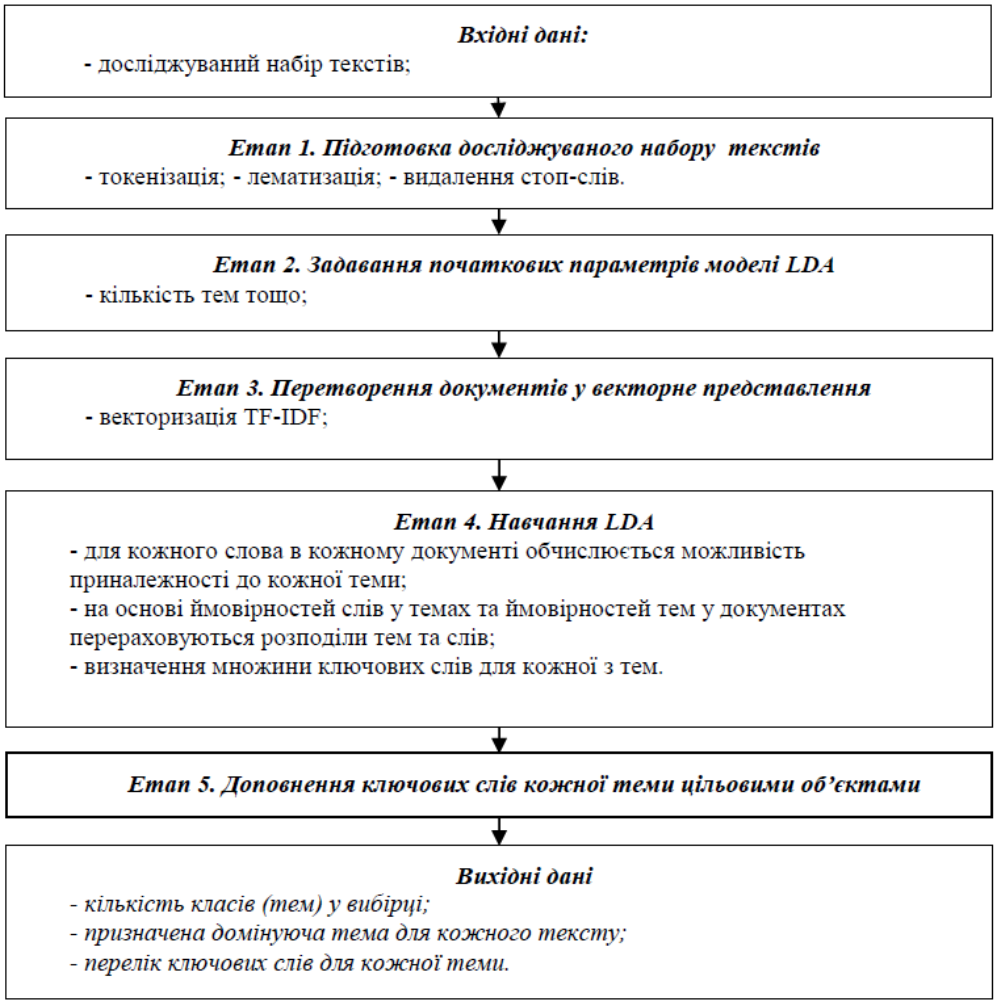

Рисунок 1. Етапи методу тематичної класифікації текстової інформації засобами обробки природної мови

Етап 1 (Підготовка текстових даних) включає токенізацію, лематизацію та видалення стоп-слів, що забезпечує чистоту та стандартизацію вхідного тексту для подальшої обробки. Етап 2 (Налаштування параметрів моделі Latent Dirichlet Allocation) LDA налаштовується на визначення кількості тем у текстах. Якщо параметр кількості тем не вказаний, модель автоматично обирає оптимальну кількість тем. Етап 3 (Навчання моделі LDA), на цьому етапі обчислюються ймовірності того, що слова належать до певних тем, а документи – до окремих категорій. Це дозволяє за розробленим методом тематичної класифікації текстової інформації засобами обробки природної мови визначити розподіл тем у текстах і сформувати ключові слова для кожної теми.

Етап 5 доповнює множини ключових слів кожної теми цільовими об'єктами із врахуванням ключових слів й іменникових сутностей предметної області, що досягти підвищити точність виявлення цільових об'єктів предметної області внаслідок врахування іменникових сутностей. Цільові об'єкти виступають об'єднаною множиною ключових слів, знайденими методами пошуку ключових слів без повторів, та NER-множиною, яка згруповані шляхом лематизації.

Вихідними даними методу є кількість тем вибірки, визначена домінуюча тема для кожного тексту, перелік ключових слів до кожної теми.

Запропонований метод тематичної класифікації текстових даних створено для автоматизованого ідентифікування та групування текстів за основними темами. Цей підхід сприяє ефективній організації великих обсягів текстової інформації, забезпечуючи структурований доступ до її змісту.

Соціальні медіа генерують значні обсяги текстових даних, що містять думки, коментарі та обговорення. Використання розробленого методу для аналізу таких даних дозволяє зрозуміти основні теми, які цікавлять користувачів, а також виявити загальні настрої у спільноті.

Для проведення дослідження обрано англомовний датасет "fake-and-real-news-dataset", який поділено на два файли: "Fake.csv" (містить 23,502 фейкові статті) та "True.csv" (містить 21,417 достовірних новин) [5]. Програмну реалізацію методу виконано у середовищі Google Colab із використанням Jupyter Notebook. У процесі тематичного моделювання без попереднього визначення кількості тем оптимальна їх кількість була встановлена на основі когерентності моделі – 14 тем. Графік когерентності, який демонструє максимальне значення, наведено на рис. 2.

Як показано на рис. 2, оптимальна кількість тем визначається точкою максимального значення когерентності. Якщо когерентність продовжує зростати, це свідчить про можливість витягування додаткових тем. Зворотна тенденція чи стабілізація вказує на досягнення оптимального розподілу. Відповідно, тематичне моделювання було виконано з класифікацією на 14 тем.

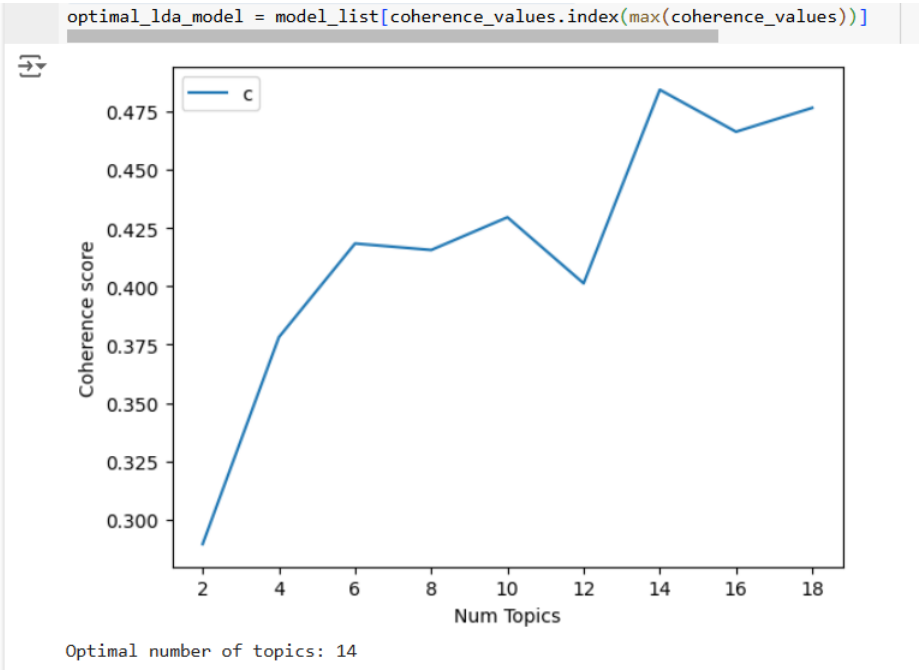

Рисунок 1. Визначення оптимальної кількості тем

Для перевірки якості моделі використовувалась крос-валідація з п’ятьма фолдами [6, 7]. У рамках цієї процедури дані поділяли на п’ять частин, де чотири використовувались для навчання, а одна — для тестування [8]. Процес повторювався п’ять разів, що забезпечило рівномірне використання всіх частин у ролі тестової.

Результати показали, що оцінка точності становила 0.71 для базового підходу без доповнення ключових слів для тем та 0.86 при використанні розширеного набору ключових слів [8, 9]. Як класифікатор застосовувався алгоритм SVC, де навчання проводилось на ключових словах [10]. Оцінка 0.71 отримана шляхом класичної класифікації моделі LDA. Високі значення точності свідчать про ефективність методу, незважаючи на нерівномірний розподіл даних між класами та значну кількість тем [11].

Подальші дослідження будуть спрямовані на покращення точності класифікації за умов нерівномірного розподілу даних між класами, а також на вивчення альтернативних алгоритмів для тематичного моделювання.

ВИСНОВКИ

Запропоновано тематичної класифікації текстової інформації засобами обробки природної мови для автоматизованого ідентифікування та групування текстів за основними темами. Цей підхід сприяє ефективній організації збереження і використання великих обсягів текстової інформації, організовуючи структурований доступ до її змісту.

Запропонований метод відрізняється від аналогів можливістю динамічного визначення тем завдяки використанню тематичного моделювання, а також розширеним набором ключових слів. У цьому методі поєднуються ключові слова, отримані через LDA, з додатковими цільовими термінами, релевантними до предметної області.

Метод був реалізований програмно та протестований на англомовному наборі даних. За результатами тематичного моделювання було встановлено оптимальну кількість тем у датасеті 14. Проведена крос-валідація продемонструвала результат точності 0.86, що на 0.15 перевищує точність, отриману при використанні LDA без додаткових модифікацій для класифікації.

Список посилань:

1. *Sarin, G., Kumar, P., & Mukund, M. (2024). Text classification using deep learning techniques: a bibliometric analysis and future research directions. Benchmarking: An International Journal, 31(8), 2743-2766.*
2. *Osuntoki S., Odumuyiwa V., & Sennaike O. (2022). Understanding document thematic structure: A systematic review of topic modeling algorithms. Journal of Information and Organizational Sciences, 46(2), 305-322.*
3. *Mindubaev, A., Prokopyev, N., & Burnashev, R. (2022, November). Implementation of the Thematic Text Analysis Algorithm Using Machine Learning. In World Conference Intelligent System for Industrial Automation (pp. 1-11). Cham: Springer Nature Switzerland.*
4. *Sheikhattar, M.R., Nezafati, N. & Shokouhyar, S. A thematic analysis–based model for identifying the impacts of natural crises on a supply chain for service integrity: a text analysis approach. Environ Sci Pollut Res 29, 79413–79433 (2022). https://doi.org/10.1007/s11356-022-21380-x*
5. *Датасет «fake-and-real-news-dataset». URL: https://www.kaggle.com/datasets/clmentbisaillon/fake-and-real-news-dataset*
6. *Залуцька О.О., Молчанова М.О., Віт Р.В., Мазурець О.В. Конфігурування нейронної мережі для класифікації емоційної тональності текстової інформації за показниками семантичної зв'язності. Збірник наукових праць за матеріалами XV Всеукраїнської науково-практичної конференції «Актуальні проблеми комп'ютерних наук АПКН-2023». Хмельницький, 2023. с. 102-107.*
7. *Молчанова М.О., Мазурець О.В., Собко О.В., Віт Р.В., Назаров В.В. Алгоритм виявлення аб'юзивного вмісту в україномовному аудіоконтенті для імплементації в об'єктно-орієтовану інформаційну систему. Науковий журнал «Вісник Хмельницького національного університету» серія: Технічні науки. Хмельницький, 2024. №1 (331). С. 101-106.*
8. *Мазурець О.В., Віт Р.В. Інтелектуальний метод виявлення цільових об'єктів предметної області за показниками семантичної зв'язності для класифікації текстової інформації. Розвитки інформаційно-керуючих систем та технологій.: монографія. Львів-Торунь : Lina-Pres, 2024. – С.223-244.*
9. *Мазурець О., Віт Р. Інтелектуальний метод виявлення цільових об'єктів предметної області для класифікації текстової інформації. Матеріали XII Міжнародної науково-практичної конференції «Інформаційні управляючі системи та технології ІУСТ-ОДЕСА-2024». 23-25.09.2024. Одеса. 2024. С.205-208.*
10. *Мазурець О.В., Віт Р.В. Дослідження ефективності методу виявлення цільових об'єктів предметної області. Інформаційні технології і автоматизація. Матеріали XVII міжнародної науково-практичної конференції. 31 жовтня – 1 листопада 2024 р. Одеса, ОНТУ. 2024. С.650-653.*
11. *Віт Р.В., Мазурець О.В. Метод виявлення множин цільових об'єктів предметної області у текстовому контенті. Збірник наукових праць за матеріалами XVI Всеукраїнської науково-практичної конференції «Актуальні проблеми комп'ютерних наук АПКН-2024». 15-16 листопада 2024. Хмельницький, 2024. с. 78-82.*