МЕТОД ТЕМАТИЧНОЇ КЛАСИФІКАЦІЇ ТЕКСТІВ З ВИКОРИСТАННЯМ МАШИННОГО НАВЧАННЯ

Мазурець О.В., exe.chong@gmail.com, **Віт Р.В.**, vit.roman.vit@gmail.com

Хмельницький національний університет

Процес тематичної класифікації текстів полягає у групуванні текстової інформації за певними категоріями чи темами, що дозволяє виявляти ключові ідеї, тенденції й шаблони у даних. Використання алгоритмів машинного навчання дозволяє автоматизувати процес аналізу текстів, ефективно застосовуючи контекстуальні ознаки, що значно підвищує швидкість і точність класифікації [1].

Метою роботи є розробка та дослідження методу тематичної класифікації текстів з використанням машинного навчання, здатного підвищити точність і релевантність тематичного аналізу, що сприятиме прийняттю обґрунтованих рішень на основі текстових даних.

Метод тематичної класифікації текстів з використанням машинного навчання забезпечує перетворення текстових даних у результати, які включають кількість тем, домінуючу тему кожного документа, а також розширений набір ключових слів для кожної теми (рис. 1). Цей підхід поєднує адаптивність тематичного моделювання з автоматичним розширенням ключових слів, що дозволяє проводити ефективний тематичний аналіз [2]. На вхід методу подається досліджуваний текст разом із попередньо обробленим збалансованим корпусом текстів відповідної предметної області. Перший етап передбачає підготовку текстових даних, яка включає токенізацію, лематизацію та видалення стоп-слів. На другому етапі задаються початкові параметри моделі LDA: якщо кількість тем відома, алгоритм здійснює класифікацію на задану кількість категорій; у разі відсутності цього параметра навчання проводиться для визначення оптимальної кількості тем. Четвертий етап зосереджений на навчанні моделі LDA, де для кожного слова в документі розраховується ймовірність його приналежності до певної теми. На основі цих ймовірностей визначаються розподіли слів і тем, що використовуються для формування ключових слів кожної теми [3].

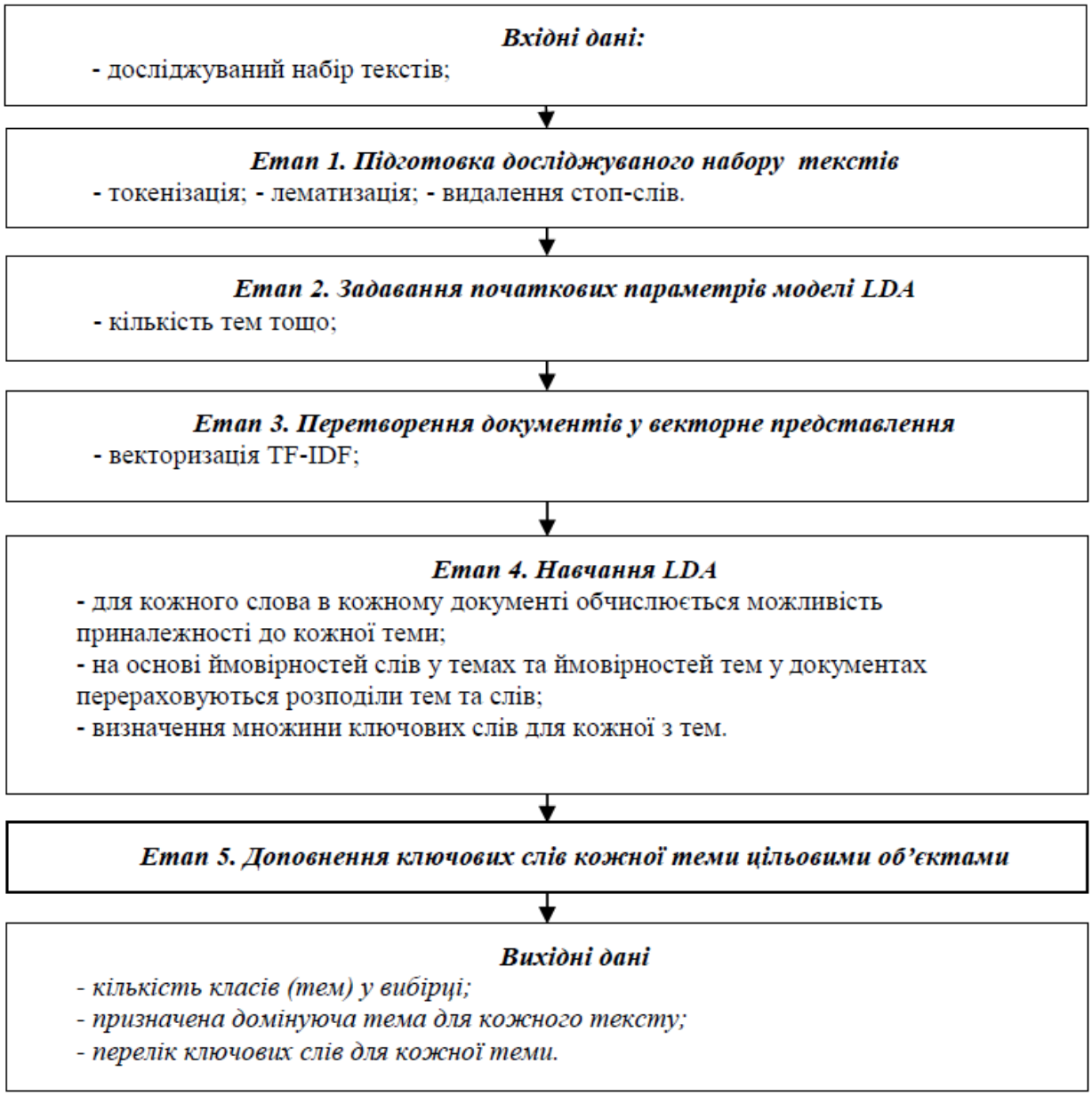

Рис. 1. Етапи методу тематичної класифікації текстів

П'ятий етап полягає в доповненні набору ключових слів кожної теми за допомогою цільових об'єктів, враховуючи терміни та іменникові сутності, релевантні до предметної області. Це дозволяє підвищити точність ідентифікації цільових об'єктів шляхом врахування згрупованих через лематизацію іменникових сутностей. Цільові об'єкти формуються шляхом об'єднання унікальних ключових слів, отриманих різними методами, та сутностей NER.

Результатом застосування методу є інформація про кількість тем у вибірці, визначення основної теми для кожного тексту та розширений список ключових слів для кожної теми.

Для проведення дослідження методу тематичної класифікації текстів з використанням машинного навчання використано англомовний датасет «fake-and-real-news-dataset», що включає 23,502 недостовірні статті 21,417 достовірних новин [4]. Програмна реалізація методу була виконана за допомогою середовища Google Colab із використанням Jupyter Notebook. У процесі тематичного моделювання, без попереднього визначення кількості тем, оптимальна їх кількість була визначена на основі значення когерентності моделі і становила 14 тем.

Оптимальне число тем встановлюється на основі максимальної когерентності моделі [5]. Якщо когерентність зростає, це свідчить про можливість додаткового виділення тем, тоді як зниження чи стабілізація цього показника вказує на досягнення найкращого розподілу. У цьому дослідженні тематичне моделювання виконувалося з класифікацією текстів на 14 тем. Представлення ключових слів для кількох категорій наведено на рис. 2.

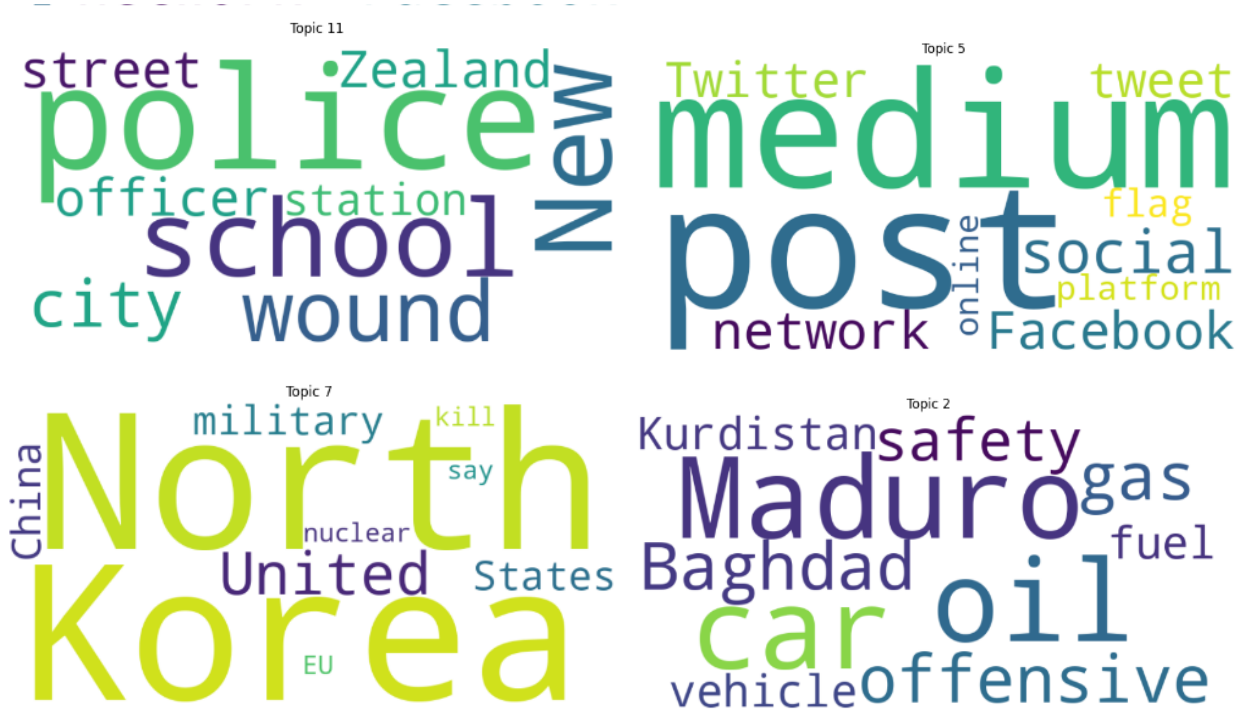

Рис. 2. Ключові слова за тематичної класифікації на 14 тем

Як видно з рис. 2, виділені теми мають чіткий розподіл і майже не перетинаються, що підтверджує оптимальність вибору 14 тем для цього датасету.

Отже, метод тематичної класифікації текстів з використанням машинного навчання розроблено для автоматичного визначення та групування текстів за їх основними темами. Такий метод дозволяє ефективно впорядковувати великі обсяги текстової інформації та забезпечувати доступ до її змісту в структурованій формі.

Соціальні мережі створюють значний масив текстових даних, які містять різноманітні думки, коментарі та обговорення. Використання створеного методу для аналізу таких даних допомагає виявляти ключові теми, які цікавлять користувачів, а також визначати загальні настрої в спільноті.

Список використаних джерел

1. Sarin G., Kumar P., Mukund M. Text classification using deep learning techniques: a bibliometric analysis and future research directions. *Benchmarking: An International Journal.* 2024. 31(8). P. 2743-2766.
2. 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.
3. 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».* Одеса. 2024. С. 205-208.
4. Датасет «fake- and- real- news- dataset». URL: https://www.kaggle.com/datasets/clmentbisaillon/fake-and-real-news-dataset
5. Віт Р.В., Мазурець О.В. Метод виявлення множин цільових об’єктів предметної області у текстовому контенті. *Актуальні проблеми комп’ютерних наук АПКН-2024.* Хмельницький, 2024. С. 78-82.