УДК 004.932

**Ю. І. Маршук** [1]
**О. В. Бісікало** [1]

# РОЗРОБКА МОДУЛЯ КЛАСТЕРИЗАЦІЇ НОВИН ДЛЯ СИСТЕМИ JETIQ

[1]Вінницький національний технічний університет

***Анотація***

*У статті представлено розробку програмного модуля для автоматизованої кластеризації новин в інформаційно-аналітичній системі JetIQ. Запропоноване рішення базується на використанні методів обробки природної мови, векторного представлення текстів (embeddings), алгоритму K-Means та локальної великої мовної моделі Llama 3.1. Розроблений модуль забезпечує автоматичне групування новин за тематикою, фільтрацію шумового контенту та генерацію назв сформованих категорій. Програмну реалізацію виконано мовою Python із використанням FastAPI та REST API архітектури*.

***Ключові слова:*** *кластеризація текстів, NLP, K-Means, Llama 3.1, JetIQ, FastAPI, embeddings, REST API*.

***Abstract***

*The paper presents the development of a news clustering module for the JetIQ information system. The proposed solution is based on natural language processing methods, text embeddings, the K-Means clustering algorithm, and the local large language model Llama 3.1. The developed module provides automatic thematic grouping of news articles, noise filtering, and generation of cluster labels. The software implementation is developed in Python using FastAPI and REST API architecture..*

***Keywords:*** *text clustering, NLP, K-Means, Llama 3.1, JetIQ, FastAPI, embeddings, REST API.*

## Вступ

У сучасних інформаційних системах обсяги текстових даних постійно зростають, що ускладнює процес їх опрацювання та структуризації. Особливо актуальною ця проблема є для освітніх порталів, де щоденно публікуються новини, оголошення та повідомлення різної тематики [1].

Інформаційно-аналітична система JetIQ використовується для організації освітнього процесу та інформування користувачів про актуальні події. Однак ручна категоризація новин є трудомісткою та потребує значних часових витрат. Тому актуальним є впровадження автоматизованих методів тематичного групування новин на основі сучасних технологій штучного інтелекту та машинного навчання.

Метою роботи є розробка програмного модуля семантичної кластеризації новин для системи JetIQ на основі алгоритму K-Means [2, 3] та локальної мовної моделі Llama 3.1 [6].

## Результати дослідження

У результаті аналізу предметної області було визначено основні вимоги до системи автоматизованої кластеризації новин та обґрунтовано вибір технологій реалізації:

**1. Попередню обробку текстових даних:**
- очищення HTML-розмітки;
- видалення службових символів та шуму;
- нормалізацію текстової інформації.

**2. Семантичну векторизацію новин:**
- формування embeddings для кожного повідомлення;
- використання трансформерних моделей [4] для отриманняконтекстних векторів [5];
- кешування результатів для зменшення часу обробки.

**3. Кластеризацію новин:**
- застосування алгоритму K-Means [2, 3];
- формування тематичних груп повідомлень;
- визначення належності нових новин до існуючих кластерів.

**4. Автоматичне іменування категорій:**
- використання локальної LLM Llama 3.1 [6];
- генерація коротких та зрозумілих назв кластерів;
- забезпечення автономної роботи без використання зовнішніх API.

**5. REST API інтеграцію:**
- обмін даними у форматі JSON;

- асинхронну обробку запитів;
- можливість інтеграції з існуючою інфраструктурою JetIQ.

Розробка програмного модуля здійснювалася з використанням сучасного стеку технологій:

1. **Python + FastAPI —** для реалізації серверної логіки, обробки запитів та побудови високопродуктивного REST API.
2. **Ollama / LM Studio + Meta Llama 3.1 [6] —** для локального запуску великих мовних моделей, генерації векторних представлень текстів (embeddings), семантичного аналізу новин та автоматичного формування назв кластерів.
3. **Scikit-learn —** для реалізації алгоритму кластеризації K-Means [2, 3] та математичної обробки векторних даних.
4. **SQLite** — для збереження новин, кешування ембедингів та результатів кластеризації.

У порівнянні з існуючими аналогами (Google News Engine, статистичними TF-IDF системами та хмарними AI API), розроблений модуль має такі переваги:

1. Повністю локальна обробка даних без передачі інформації стороннім сервісам.
2. Висока точність семантичного групування завдяки використанню сучасних embedding-моделей [5] та трансформерної архітектури [4].
3. Відсутність витрат на використання хмарних API та незалежність від стабільності інтернет-з'єднання.
4. Автоматичне формування людиночитабельних назв тематичних категорій за допомогою локальної великої мовної моделі.
5. Можливість інтеграції з інформаційною системою JetIQ через REST API.

Разом із перевагами існують певні обмеження, що потребують подальшого вдосконалення:

1. Швидкість обробки великих наборів новин залежить від обчислювальних ресурсів локального сервера.
2. Для досягнення максимальної точності необхідне періодичне оновлення мовних моделей та векторних представлень.
3. Якість кластеризації може знижуватися при наявності великої кількості коротких або недостатньо інформативних повідомлень.

Розроблений модуль демонструє високу ефективність у задачах автоматичного тематичного групування новин та має значний потенціал для впровадження в освітні інформаційні системи.

На рисунках 1–3 представлено діаграму варіантів використання, діаграму послідовності та діаграму класів програмного модуля кластеризації новин для системи JetIQ.

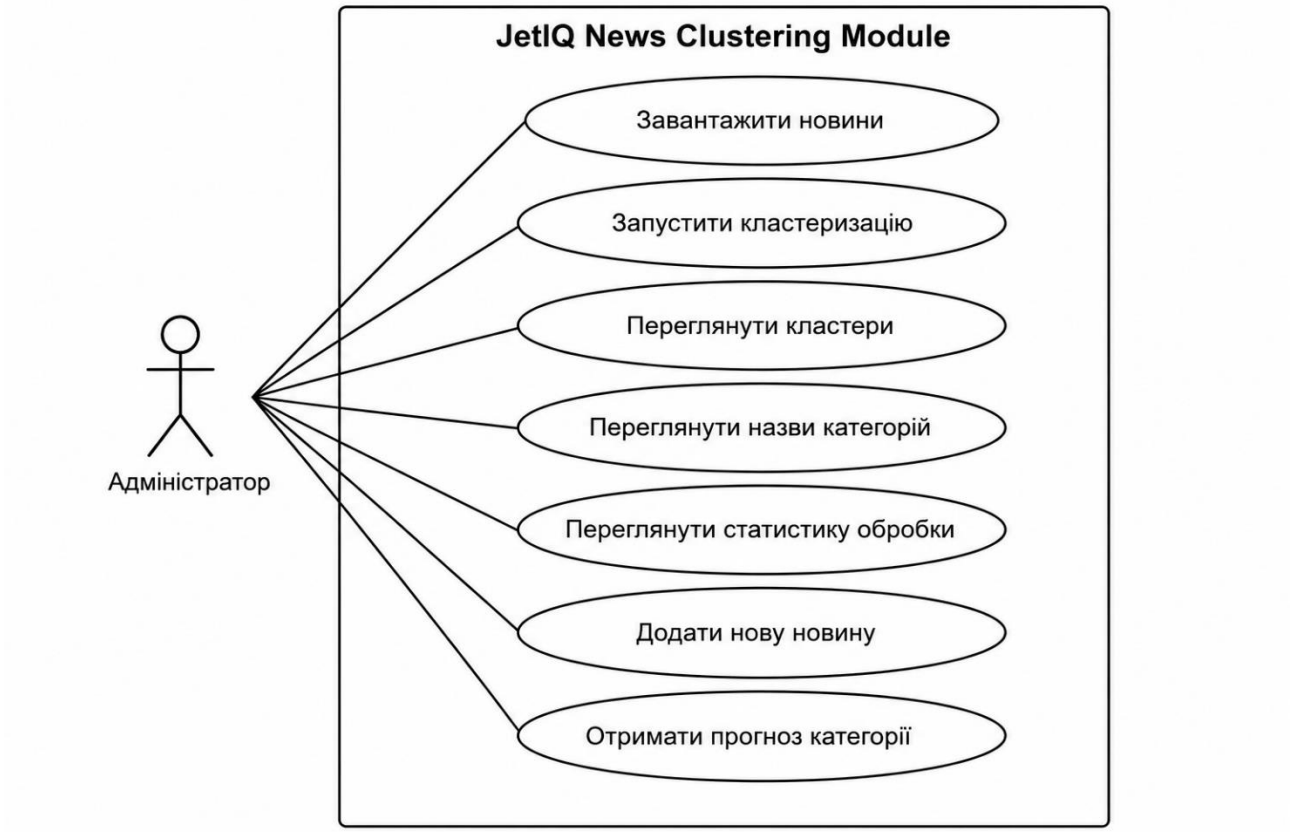


Рисунок 1 – Діаграма варіантів використання (Use Case Diagram)

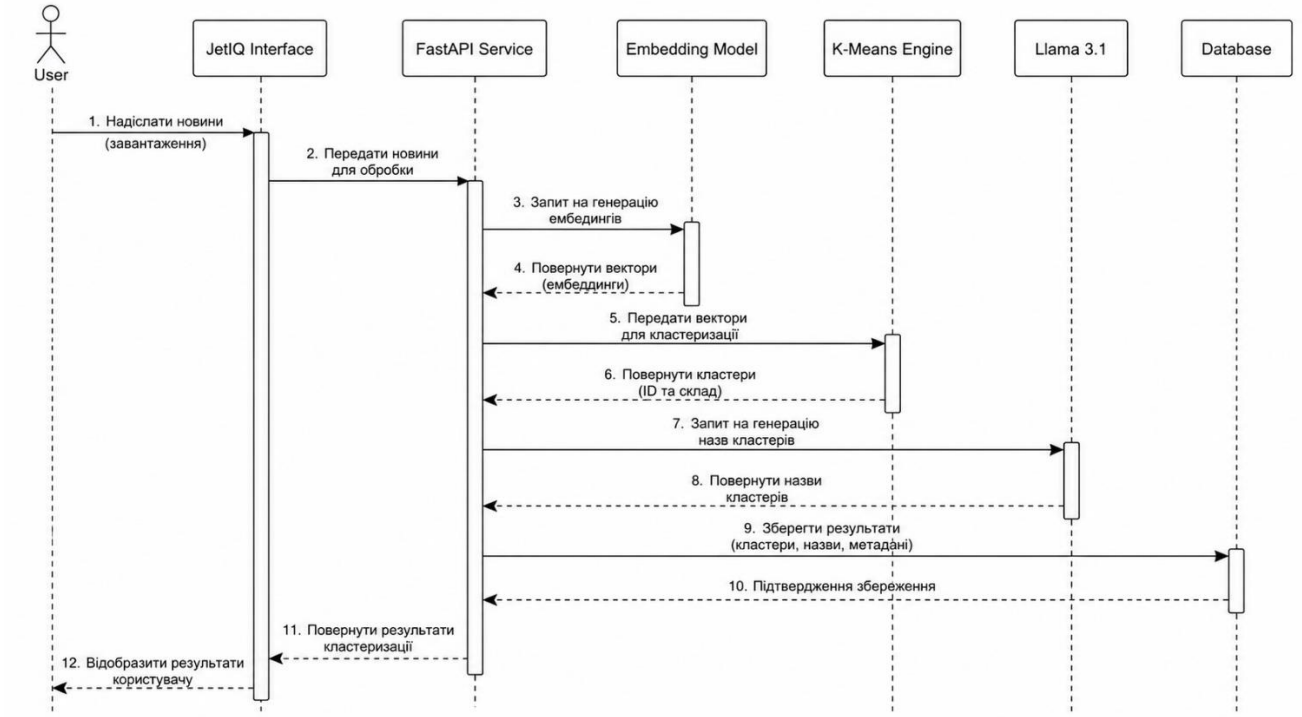


Рисунок 2 – Діаграма послідовності (Sequence Diagram)

Діаграма послідовності ілюструє порядок взаємодії між користувачем і компонентами системи під час процесу кластеризації новин. Послідовність цього процесу така:

1. Надіслати новини: користувач завантажує новинні повідомлення для подальшої обробки через інтерфейс системи JetIQ.
2. Передати новини для обробки: інтерфейс JetIQ передає отримані новини до сервісу FastAPI для запуску процесу аналізу.
3. Запит на генерацію ембедингів: сервіс FastAPI надсилає текстові дані до моделі векторизації для формування семантичних представлень новин.
4. Повернути вектори (ембединги): модель векторизації обробляє тексти та повертає їхні векторні представлення сервісу FastAPI.
5. Передати вектори для кластеризації: FastAPI передає отримані ембединги до модуля K-Means для виконання групування новин.
6. Повернути кластери: модуль K-Means формує тематичні групи новин та повертає інформацію про створені кластери.
7. Запит на генерацію назв кластерів: після завершення кластеризації FastAPI надсилає інформацію про сформовані групи до великої мовної моделі Llama 3.1.
8. Повернути назви кластерів: модель Llama 3.1 аналізує зміст кожного кластера та автоматично генерує зрозумілі назви категорій.
9. Зберегти результати: сервіс FastAPI передає сформовані кластери, їх назви та службові метадані до бази даних.
10. Підтвердження збереження: база даних виконує запис інформації та повертає підтвердження успішного збереження результатів.
11. Повернути результати кластеризації: сервіс FastAPI надсилає сформовані результати до інтерфейсу JetIQ.
12. Відобразити результати користувачу: інтерфейс системи оновлюється та відображає користувачу сформовані тематичні кластери новин і назви відповідних категорій.

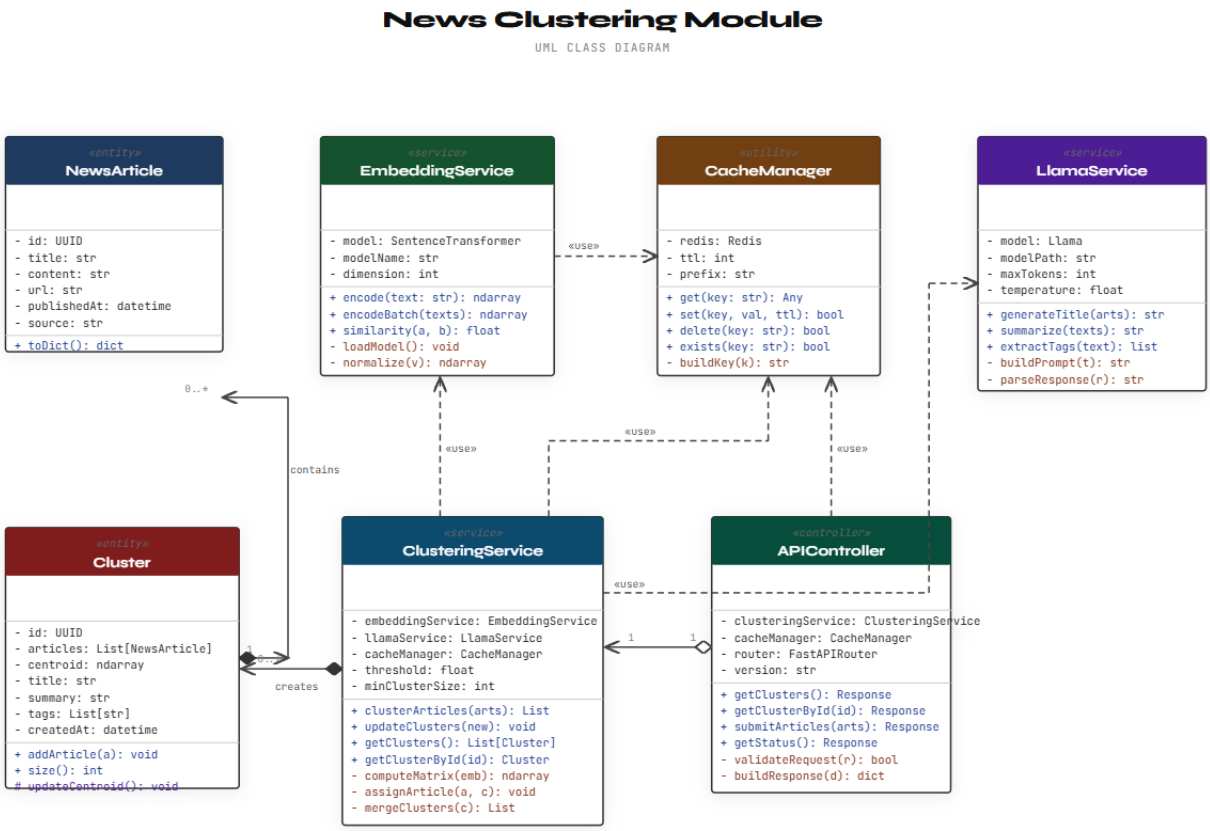


Рисунок 3 – Діаграма класів (Class Diagram)

Діаграма класів відображає структуру модуля кластеризації новин з підтримкою семантичного аналізу та штучного інтелекту. Основні класи та їх взаємозв'язки модуля такі:

1. **NewsArticle** – клас, що представляє окрему новинну статтю з її метаданими (ідентифікатор, заголовок, вміст, URL, дата публікації, джерело). Містить метод для серіалізації статті у словник.
2. **EmbeddingService** – сервіс для векторного представлення текстів на основі моделі SentenceTransformer [5]. Відповідає за кодування окремих текстів та пакетне кодування, обчислення схожості між векторами, завантаження моделі та нормалізацію векторів.
3. **CacheManager** – допоміжний клас для управління кешем через Redis. Забезпечує збереження, отримання, видалення та перевірку наявності даних у кеші, а також формування ключів із префіксом.
4. **LlamaService** – сервіс для взаємодії з мовною моделлю Llama. Відповідає за генерацію заголовків кластерів, створення резюме групи статей, вилучення тегів тощо.
5. **Cluster** – клас, що представляє кластер новин — групу тематично споріднених статей. Зберігає список статей, центроїд кластера, згенерований заголовок, резюме, теги та дату створення.
6. **ClusteringService** – головний сервіс кластеризації, що координує роботу EmbeddingService, LlamaService та CacheManager. Відповідає за кластеризацію статей, оновлення існуючих кластерів, отримання кластерів за ідентифікатором, обчислення матриці схожості, призначення статей до кластерів та їх злиття.
7. **APIController** – контролер, що надає REST API для взаємодії із системою. Обробляє HTTP-запити на отримання списку кластерів, пошук кластера за ідентифікатором, приймання нових статей та перевірку стану системи. Делегує бізнес-логіку до ClusteringService та використовує CacheManager для оптимізації відповідей.

Результати експерименту з запропонованим модулем, представлені в таблиці 1, відповідають експертним оцінкам змісту датасету новин команди розробників системи JetIQ.

Таблиця 1 – Результати експериментів з модулем кластеризації новин.

| № | Тематична група | Кількість повідомлень | Частка, % |
|---|---|---|---|
| 1 | Освітній процес та розклад | 145 | 14.5 |
| 2 | Наукові конференції та гранти | 112 | 11.2 |
| 3 | Спортивні змагання | 67 | 6.7 |
| 4 | Благодійність та волонтерство | 85 | 8.5 |
| 5 | Адміністративні накази | 98 | 9.8 |
| 6 | Студентське самоврядування | 74 | 7.4 |
| 7 | Міжнародні партнерства | 55 | 5.5 |
| 8 | Культурне та мистецьке життя | 62 | 6.2 |
| 9 | IT та інновації | 89 | 8.9 |
| 10 | Соціальні питання | 71 | 7.1 |
| 11 | Інше / Різне | 142 | 14.2 |
| **Разом** | | **1000** | **100** |

## Висновки

У результаті виконання роботи розроблено програмний модуль кластеризації новин для системи JetIQ, який забезпечує автоматизоване тематичне групування текстових повідомлень на основі сучасних методів обробки природної мови. Запропоноване рішення використовує семантичні embeddings, алгоритм K-Means та локальну мовну модель Llama 3.1 для формування тематичних категорій та автоматичної генерації їх назв. Розроблений модуль характеризується автономністю, конфіденційністю обробки даних та можливістю інтеграції у наявну інфраструктуру JetIQ через REST API. Проведені експериментальні дослідження підтвердили ефективність підходу та перспективність його використання для автоматизації інформаційних процесів в освітніх системах.

## СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ


1. О. В. Бісікало, Д. П. Урлапова, Д. І. Тєлєга. «Класифікація жартів за категоріями гумору з використанням методів машинного навчання, » в Матеріали конференції «Молодь в науці: дослідження, проблеми, перспективи (МН-2025)», Вінниця, 2025. [Електронний ресурс]. Режим доступу: https://conferences.vntu.edu.ua/index.php/mn/mn2025/paper/view/25576. Дата звернення: Черв. 2025. – 4 с.
2. MacQueen J. Some methods for classification and analysis of multivariate observations // Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability. 1967. Vol. 1. P. 281–297.
3. Arthur D., Vassilvitskii S. k-means++: The advantages of careful seeding // Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms. 2007. P. 1027–1035.
4. Vaswani A., Shazeer N., Parmar N. et al. Attention Is All You Need // Advances in Neural Information Processing Systems. 2017. Vol. 30. P. 5998–6008.
5. Reimers N., Gurevych I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks // Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing. 2019. P. 3982–3992.
6. Meta Llama 3: The most capable openly available LLM to date. Meta AI. URL: https://ai.meta.com/blog/meta-llama-3/ (дата звернення: 10.04.2026).



***Маршук Юрій Ігорович*** – студент групи 1ІСТ-22б, факультет інтелектуальних інформаційних технологій та автоматизації, Вінницький національний технічний університет, Вінниця.

***Бісікало Олег Володимирович*** – д.т.н., професор кафедри автоматизації та інтелектуальних інформаційних технологій, Вінницький національний технічний університет, м. Вінниця.

***Marshuk Yurii I.*** **–** student of group 1IST-22b, Faculty of Intelligent Information Technologies and Automation, Vinnytsia National Technical University, Vinnytsia, Ukraine..

***Bisikalo Oleh V.*** **–** Dr. Sc., Professor at the Department of Automation and Intelligent Information Technologies, Vinnytsia National Technical University, Vinnytsia, Ukraine.