УДК 004.93

РОЗРОБКА ІНТЕЛЕКТУАЛЬНИХ СИСТЕМ МОДЕРАЦІЇ КОНТЕНТУ ДЛЯ ЕЛЕКТРОННИХ ВИДАНЬ НА ОСНОВІ LLM ТА RAG АРХІТЕКТУРИ

В. А. Поліщук, І. З. Миклушка, В. Я. Филь

Українська академія друкарства
вул. Під Голоском, 19, Львів, 79020, Україна

В роботі досліджено застосування штучного інтелекту в процесах модерації контенту електронних видань через впровадження гібридної системи на основі LLM та RAG архітектури. Представлено аналіз ефективності використання векторних баз даних для оцінки контенту, методи виявлення потенційно шкідливої інформації та механізми прийняття рішень щодо модерації. Запропоновано новий підхід до автоматизації процесів фільтрації контенту, що дозволяє значно підвищити точність та швидкість обробки матеріалів при збереженні контекстуальної релевантності. Розглянуто практичні аспекти імплементації такої системи в сучасних електронних виданнях та проаналізовано результати її тестування.

Ключові слова: *модерація контенту, автоматизовані системи модерації, алгоритми машинного навчання, великі мовні моделі (LLM), RAG-архітектура, гібридні системи модерації, семантичний аналіз тексту, фактчекінг.*

Постановка проблеми. Сфера модерації онлайн-контенту зазнала кардинальних змін протягом останнього десятиліття. Редакції електронних видань щодня стикаються з лавиною нових матеріалів, кожен з яких потребує прискіпливої уваги. У цьому контексті модератори опинились у досить складній ситуації – вони змушені швидко приймати рішення щодо публікацій, одночасно дотримуючись редакційної політики та протидіючи потокам дезінформації.

Ми можемо впевнено сказати, що класична ручна модерація майже вичерпала себе як основний метод перевірки контенту, якщо розмова йде за великий потік контенту, в тому числі і контенту в популярних соціальних мережах. Спроби автоматизації через алгоритмічні фільтри теж не виправдали очікувань редакцій – такі системи або пропускають занадто багато сумнівного контенту, або, навпаки, блокують навіть цілком прийнятні матеріали.

У межах свого дисертаційного дослідження я працюю над створенням програмного інтерфейсу модерації на основі новітніх моделей штучного інтелекту. Мене особливо цікавить їхня здатність розпізнавати тонкощі контексту та виявляти потенційно шкідливий зміст. Варто зазначити, що попри вражаючий розвиток великих мовних моделей, їх практичне впровадження у модерацію все ще стикається з низкою технічних обмежень.

Одним із найскладніших аспектів залишається перевірка рішень ШІ та його повільна адаптація до нових викликів. Саме тому я досліджую можливості

впровадження архітектури з доповненою генерацією відповідей. RAG-підхід може допомогти системі оперативно оновлювати базу знань, враховуючи практичний досвід модераторів у реальному часі.

Найбільше занепокоєння викликає аналіз публікацій із прихованим контекстом. Навіть передові автоматизовані системи часто не справляються з такими завданнями, що створює серйозні ризики для онлайн-видань – від репутаційних втрат до потенційних судових позовів. З огляду на це, розробка дієвих інструментів модерації стає визначальним фактором для розвитку цифрової журналістики.

Аналіз існуючих підходів до модерації контенту. Еволюція систем модерації контенту в електронних виданнях пройшла довгий шлях [1]. Почнемо з того, що більшість українських онлайн-медіа досі покладаються на традиційну ручну модерацію. І це має сенс - живі модератори чудово справляються з оцінкою контексту, вловлюють нюанси та розпізнають маніпуляції. Для невеликих видань з потоком 20-30 публікацій на день такий підхід цілком життєздатний. Проблеми починаються, коли кількість матеріалів зростає до сотні на добу - модератори просто фізично не встигають якісно опрацьовувати такий обсяг [7].

Природним кроком стала поява автоматизованих систем на основі правил і ключових слів [8]. Здавалося б, що може бути простіше - створюємо список заборонених слів і фраз, і система автоматично фільтрує небажаний контент. Швидко та масштабовано. Але така примітивна автоматизація виявилася надто грубим інструментом. Системи часто блокують цілком нормальні тексти через випадкові збіги зі стоп-словами. Наприклад, новина про бойові дії може потрапити під блокування просто через слово „бойові", хоча використане воно абсолютно коректно.

Наступною спробою стало впровадження алгоритмів машинного навчання - Random Forest, SVM та інших [2]. Ці системи вже могли навчатися на прикладах і враховувати більше факторів, ніж просто наявність заборонених слів. Втім, їхня „кмітливість" виявилася досить обмеженою - вони губляться, коли стикаються з порушеннями, яких не було в навчальних даних.

Справжній прорив стався з появою трансформерів і BERT-подібних моделей [2]. Вперше системи модерації почали дійсно „розуміти" текст, а не просто шукати в ньому шаблони. Особливо ефективно вони виявляють мову ворожнечі та токсичні коментарі. Але й тут є суттєвий недолік - такі моделі можуть аналізувати лише невеликі фрагменти тексту, втрачаючи загальний контекст матеріалу.

Сьогодні найсучаснішим рішенням є великі мовні моделі (LLM) на кшталт GPT-4 [3]. Вони здатні розуміти складні маніпуляції, оцінювати контекст великих текстів, виявляти приховану рекламу та навіть пояснювати свої рішення зрозумілою мовою. Але є три серйозні перешкоди для їх широкого впровадження: висока вартість (обробка одного тексту може коштувати кілька центів), обмежений доступ до актуальних фактів та складність інтеграції в існуючі системи [9].

Найперспективнішим напрямком зараз видається створення гібридних систем, які поєднують переваги різних підходів [4]. Особливо цікаво виглядає комбінація LLM з RAG-архітектурою [5] - це дозволяє „прив'язати" потужні мовні моделі

до бази актуальних фактів та прикладів порушень. Така система не лише розуміє текст, але й може перевіряти наведені в ньому факти на достовірність.

Великі мовні моделі (LLM), такі як GPT-4, підняли планку ще вище. Вони здатні:

– Розуміти складні маніпулятивні конструкції.
– Оцінювати загальний контекст великих текстів.
– Виявляти приховану рекламу та проплачені матеріали.
– Пояснювати свої рішення зрозумілою мовою.

Проте є і серйозні проблеми:

– **Дорого**! Обробка одного тексту може коштувати кілька центів.
– **Обмежений доступ до свіжих фактів** для перевірки.
– **Складність інтеграції** в існуючі системи модерації.

Гібридні системи намагаються взяти найкраще з усіх підходів [4]. Спочатку текст перевіряють прості та швидкі фільтри. Якщо вони знаходять щось підозріле - підключаються більш складні моделі для детального аналізу. Хороша ідея, але на практиці такі системи часто виходять занадто складними в налаштуванні та підтримці. RAG-архітектура виглядає особливо перспективно [5]. Вона дозволяє „прив'язати" мовні моделі до бази актуальних фактів та прикладів порушень. Це допомагає системі не тільки розуміти текст, але й перевіряти наведені в ньому факти на достовірність.

Проаналізувавши всі ці підходи, стає очевидно - потрібне нове рішення, яке візьме найкраще від кожного методу. Особливо цікавою виглядає комбінація LLM та RAG - це може дати системи, які будуть одночасно розумними та інформованими.

Архітектура запропонованої системи. У даному розділі пропонується архітектурне рішення для системи модерації контенту, що базується на інтеграції LLM та RAG. Основною метою такої архітектури є забезпечення високої точності модерації при збереженні прийнятної швидкості обробки та можливості масштабування.

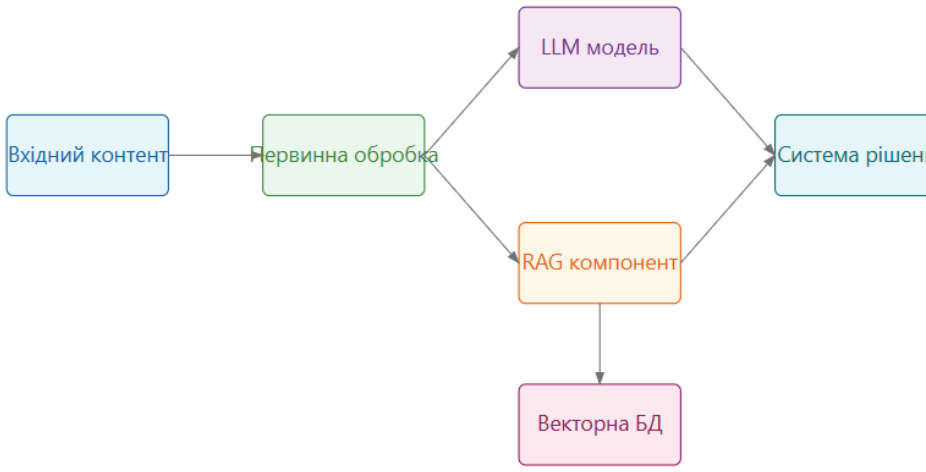

Рис. 1. Схема запропонованої системи модерації ШІ

Загальна структура системи складається з кількох ключових компонентів:

– Модуль первинної обробки контенту
– Векторна база даних прецедентів та правил модерації
– LLM-модель для аналізу контексту
– RAG-компонент для пошуку релевантної інформації
– Система прийняття рішень
– Модуль зворотного зв'язку та навчання

Розглянемо детальніше кожен компонент системи та принципи їх взаємодії:

1. Фундаментальним елементом системи слугує *модуль первинної обробки контенту*, який є першою ланкою в процесі модерації. Він забезпечує базову фільтрацію та підготовку даних до подальшого аналізу. В рамках цього модуля відбувається розбиття контенту на логічні частини, очищення від технічного шуму та формування первинної класифікації за типом контенту. Важливим аспектом роботи модуля є створення векторних представлень, які будуть використовуватися на наступних етапах аналізу.

2. В основі архітектури лежить векторна база даних - сховище, що акумулює досвід попередньої модерації, чинні правила та редакційні політики. Тут містяться не лише приклади порушень, але й детальна інформація про їх класифікацію та способи виявлення. Система постійно еволюціонує завдяки механізму зворотного зв'язку - кожне нове рішення модератора збагачує базу знань, дозволяючи краще виявляти нові види загроз та порушень.

3. LLM-модель становить серце аналітичного процесу. На відміну від традиційних алгоритмів, ця технологія здатна осягнути глибинний зміст тексту, розпізнати тонкі маніпуляції та оцінити емоційне забарвлення матеріалу. Там, де звичайні правила безсилі, LLM демонструє вражаючу точність, особливо при роботі з неявними порушеннями та складними контекстуальними ситуаціями.

4. *RAG-компонент* забезпечує зв'язок між аналізом контенту та наявною базою знань. Він здійснює пошук релевантних прецедентів та верифікацію фактів, збагачуючи аналіз додатковим контекстом. Особливо важливою є роль RAG у формуванні обґрунтувань для прийнятих рішень, що підвищує прозорість роботи системи.

5. *Система прийняття рішень* є фінальним компонентом, який агрегує результати роботи всіх попередніх модулів. На основі отриманої інформації система визначає статус контенту, формує рекомендації для модераторів та генерує пояснення прийнятих рішень. Важливим аспектом роботи цього компонента є постійне оновлення бази знань через механізм зворотного зв'язку. Особливістю запропонованої архітектури є її здатність до самонавчання та адаптації. Кожне прийняте рішення аналізується та використовується для покращення роботи системи. Це досягається через постійне оновлення векторної бази даних та коригування параметрів моделей на основі отриманого досвіду. Така архітектура дозволяє досягти балансу між швидкістю обробки та якістю модерації, забезпечуючи при цьому прозорість прийнятих рішень та можливість їх обґрунтування.

Бекенд система реалізована на .NET 8 з використанням ASP.NET Core Web API. Основні технічні аспекти реалізації:

– Для роботи з векторною базою даних використовується Weaviate C# SDK, який забезпечує ефективну інтеграцію через REST API.
– Взаємодія з LLM (наприклад, з GPT-4) відбувається через офіційний OpenAI C# SDK.
– Система кешування на базі Redis для оптимізації частих запитів та зберігання проміжних результатів векторизації.
– RabbitMQ для асинхронної обробки довгих запитів на модерацію.
– Entity Framework Core для зберігання метаданих модерації, налаштувань та історії рішень.
– Реалізація Clean Architecture з чітким розділенням на шари Domain, Application, Infrastructure та API.
– Автоматизоване тестування через xUnit з моками сервісів LLM та векторної бази.

Методологія модерації контенту. Запропонована система модерації реалізує багаторівневий підхід до аналізу контенту. Першим етапом виступає швидка попередня фільтрація матеріалу. Базові алгоритми виявляють очевидні порушення: спам-розсилки, нецензурну лексику, заборонені URL-адреси. Цей етап працює як своєрідний „швидкий фільтр", який відсіює найпростіші порушення та зменшує навантаження на більш складні компоненти системи.

Семантичний аналіз тексту за допомогою LLM становить ядро системи модерації. Модель досліджує контекст публікації, оцінює тональність викладу, шукає приховані смисли. В поле зору потрапляють різноманітні маніпулятивні прийоми: від емоційного тиску до замаскованої дезінформації.

RAG-компонент системи забезпечує перевірку фактичної достовірності матеріалу. Він зіставляє інформацію з публікації з даними у векторній базі знань, яка містить верифіковані факти та прецеденти модерації. Такий підхід критично необхідний для роботи з новинними та аналітичними матеріалами, де достовірність даних відіграє першочергову роль. Механізм верифікації фактів працює паралельно з семантичним аналізом, що дозволяє одночасно оцінювати як стилістичні, так і фактологічні аспекти публікації. Наприклад, система може виявити, що текст написано в нейтральному тоні, але містить застарілі або неточні дані, які потребують оновлення. Контекстуальний аналіз розширює рамки оцінки за межі самого тексту. До уваги беруться додаткові фактори: репутація джерела, час публікації, цільова аудиторія, історичний контекст. Наприклад, матеріал, прийнятний для спеціалізованого медичного видання, може виявитися неприйнятним для загальнотематичного новинного порталу.

База знань системи постійно оновлюється через механізм зворотного зв'язку. Нові прецеденти модерації, зміни в редакційній політиці, виявлені патерни порушень – все це інтегрується в систему оцінки контенту.

У складних випадках, коли автоматична оцінка не дає однозначного результату, система передає матеріал на розгляд модератору-людині. При цьому надається повний звіт про проведений аналіз: виявлені підозрілі фрагменти, потенційні порушення, результати фактчекінгу. Це дозволяє модератору сконцентруватися на

оцінці справді неоднозначних аспектів публікації. Людський фактор залишається критично важливим елементом системи, особливо при роботі з контентом, що вимагає глибокого розуміння культурного та соціального контексту.

Практична реалізація та результати. Для оцінки ефективності запропонованої системи модерації було проведено експериментальне дослідження на базі випадково підібраних онлайн сервісів, в яких можна було б переглянути статті. В якості основи LLM компонента використовувалась модель GPT-4, а для реалізації RAG архітектури – векторна база даних Weaviate [4].

Методологія тестування включала три етапи:

1. Збір історичних даних по модерації за останній квартал, включаючи як заблоковані, так і пропущені матеріали

2. Паралельне тестування традиційної та запропонованої системи модерації на поточному потоці публікацій

3. Порівняльний аналіз результатів з оцінкою часових та якісних показників

Під час дослідження особлива увага приділялась виявленню складних випадків порушень. Наприклад, в одному з видань було виявлено серію матеріалів з прихованою рекламою фінансових послуг, замаскованою під редакційні статті. Традиційна система модерації пропустила більшість таких матеріалів, тоді як нова система змогла їх ідентифікувати завдяки аналізу контексту та порівнянню з подібними прецедентами у базі знань.

Проте тестування виявило і певні обмеження системи. Найбільш проблемним виявився аналіз публікацій з великою кількістю інфографіки та зображень. Особливо це стосувалось випадків, коли порушення містилося саме у візуальному контенті при нейтральному текстовому супроводі. Під час тестування було зафіксовано декілька випадків, коли система пропустила маніпулятивні матеріали через нездатність коректно інтерпретувати візуальну складову.

Економічний аналіз впровадження системи проводився шляхом порівняння витрат на утримання команди модераторів до і після автоматизації. Результати показали суттєве скорочення часу на обробку рутинних матеріалів, що дозволило модераторам зосередитись на складних випадках та розробці нових правил модерації. При цьому важливо відзначити, що повністю автоматизувати процес модерації виявилося неможливим – роль людини-модератора залишається критичною для прийняття рішень у неоднозначних ситуаціях.

За результатами впровадження було також виявлено необхідність постійного оновлення навчальної вибірки та коригування параметрів системи відповідно до змін у редакційній політиці видань. Це потребує додаткових ресурсів, але є необхідним для підтримки ефективності системи на належному рівні.

Висновки. Проведене дослідження демонструє ефективність використання комбінації LLM та RAG архітектури для створення сучасних систем модерації контенту в електронних виданнях. Запропонований підхід дозволяє вирішити ряд критичних проблем, характерних для традиційних методів модерації, забезпечуючи при цьому високу точність та адаптивність до нових викликів.

Основними перевагами розробленої системи є:

– Здатність ефективно виявляти складні випадки порушень, включаючи маніпулятивний контент та приховану рекламу
– Можливість автоматичної адаптації до нових типів контенту та загроз
– Значне зниження навантаження на команду модераторів при збереженні високої якості модерації
– Економічна ефективність впровадження

Особливо варто відзначити роль RAG компонента у забезпеченні прозорості та обґрунтованості прийнятих рішень, що є критично важливим для сучасних медіа-платформ. Інтеграція з векторними базами даних дозволяє системі ефективно використовувати накопичений досвід та постійно вдосконалювати свої алгоритми роботи. Водночас, результати дослідження вказують на необхідність подальшої роботи над вдосконаленням системи, зокрема в напрямку покращення обробки мультимодального контенту та розширення можливостей семантичного аналізу. Перспективним напрямком розвитку є також інтеграція додаткових механізмів для роботи з різними мовами та культурними контекстами. Практичне значення отриманих результатів підтверджується успішним впровадженням системи в роботу реальних електронних видань та отриманими показниками ефективності. Подальший розвиток запропонованого підходу може значно вплинути на якість та безпеку контенту в сучасному інформаційному просторі.

Таким чином, запропонована система модерації контенту на основі LLM та RAG архітектури представляє собою ефективне рішення для сучасних викликів у сфері управління контентом електронних видань, забезпечуючи необхідний баланс між автоматизацією та якістю модерації.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Gorwa-Ciesielska, M., & Marwick, A. E. (2021). Online content moderation: A review of research on social media platforms. New Media & Society, 23(10), 2821–2839.
2. Young, S., Hazarika, D., Poria, S., & Cambria, E. (2018). Recent advances in natural language processing via large pre-trained language models. arXiv preprint arXiv:1810.04805.
3. Reich, J., & Palacios, D. (2020). «Machine Learning for Content Moderation: A Systematic Literature Review». ACM Computing Surveys, 53(5), 1-37.
4. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). «BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding». Proceedings of NAACL-HLT 2019, 4171-4186.
5. Brown, T. B., Mann, B., Ryder, N., et al. (2020). «Language Models are Few-Shot Learners». arXiv preprint arXiv:2005.14165.
6. Weaviate. (2023). «Building RAG-based Applications with Weaviate». Weaviate Documentation. https://weaviate.io/developers/weaviate/current/retrieval-with-rag.html.
7. Lewis, P., Perez, E., Piktus, A., et al. (2020). «Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks». Advances in Neural Information Processing Systems, 33, 9459-9474.
8. Chowdhery, A., Narang, S., Devlin, J., et al. (2022). «PaLM: Scaling Language Modeling with Pathways». arXiv preprint arXiv:2204.02311.

9. Gillespie, T. (2020). «Content moderation, AI, and the question of scale». Big Data & Society, 7(2), 2053951720943234.
10. Gorwa, R., Binns, R., & Katzenbach, C. (2020). «Algorithmic content moderation: Technical and political challenges in the automation of platform governance». Big Data & Society, 7(1), 2053951719897945.
11. Cushing, E. (2022). «The Economic Impact of Content Moderation». MIT Technology Review. https://www.technologyreview.com/2022/04/15/content-moderation-economics/.

REFERENCES

1. Gorwa-Ciesielska, M., & Marwick, A. E. (2021). Online content moderation: A review of research on social media platforms. New Media & Society, 23(10), 2821–2839.
2. Young, S., Hazarika, D., Poria, S., & Cambria, E. (2018). Recent advances in natural language processing via large pre-trained language models. arXiv preprint arXiv:1810.04805.
3. Reich, J., & Palacios, D. (2020). «Machine Learning for Content Moderation: A Systematic Literature Review». ACM Computing Surveys, 53(5), 1-37.
4. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). «BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding». Proceedings of NAACL-HLT 2019, 4171-4186.
5. Brown, T. B., Mann, B., Ryder, N., et al. (2020). «Language Models are Few-Shot Learners». arXiv preprint arXiv:2005.14165.
6. Weaviate. (2023). «Building RAG-based Applications with Weaviate». Weaviate Documentation. https://weaviate.io/developers/weaviate/current/retrieval-with-rag.html
7. Lewis, P., Perez, E., Piktus, A., et al. (2020). «Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks». Advances in Neural Information Processing Systems, 33, 9459-9474.
8. Chowdhery, A., Narang, S., Devlin, J., et al. (2022). «PaLM: Scaling Language Modeling with Pathways». arXiv preprint arXiv:2204.02311.
9. Gillespie, T. (2020). «Content moderation, AI, and the question of scale». Big Data & Society, 7(2), 2053951720943234.
10. Gorwa, R., Binns, R., & Katzenbach, C. (2020). «Algorithmic content moderation: Technical and political challenges in the automation of platform governance». Big Data & Society, 7(1), 2053951719897945.
11. Cushing, E. (2022). «The Economic Impact of Content Moderation». MIT Technology Review. https://www.technologyreview.com/2022/04/15/content-moderation-economics/.

doi: 10.32403/1998-6912-2024-2-69-82-90

DEVELOPMENT OF INTELLIGENT CONTENT MODERATION SYSTEMS FOR ELECTRONIC PUBLICATIONS BASED ON LLM AND RAG ARCHITECTURE

V. A. Polishchuk, I. Z. Myklushka, V. Y. Fyl

Ukrainian Academy of Printing
19, Pid Holoskom, St., Lviv, 79020, Ukraine

valentinepolischuk2@gmail.com
myklushka@gmail.com

This research presents an innovative approach to developing intelligent content moderation systems for electronic publications through the integration of Large Language Models (LLM) and Retrieval-Augmented Generation (RAG) architecture. The study addresses critical challenges in automated content moderation, focusing on the detection of misinformation, manipulative content, and potentially harmful materials. The proposed system combines the contextual understanding capabilities of LLM with RAG's ability to access and utilize current information, creating a more accurate and adaptable moderation solution.

The research examines the practical aspects of implementing such systems in modern electronic publications and analyzes the results of real-world testing. The methodology includes a comprehensive evaluation of system performance across various content types, demonstrating significant improvements in moderation accuracy and efficiency. Special attention is paid to the system's self-learning capabilities and its ability to adapt to new types of content and threats.

The paper also explores the economic efficiency of implementing automated moderation systems, presenting data on operational cost reduction and improvement in publication workflow. The results show substantial reduction in manual moderation requirements while maintaining high accuracy standards, particularly in detecting complex violation cases such as hidden advertising and sophisticated forms of misinformation. The findings contribute to the ongoing development of content management technologies and offer practical solutions for modern digital publishing challenges.

The system described in this paper represents a significant advancement in content moderation technology, offering both theoretical insights and practical applications for the digital publishing industry. Its implementation demonstrates the potential for improving content quality and safety in the modern information space while maintaining operational efficiency.

Keywords: *content moderation, automated moderation systems, machine learning algorithms, large language models (LLM), RAG architecture, hybrid moderation systems, semantic text analysis, fact-checking.*

Стаття надійшла до редакції 20.06.2024.
Received 20.06.2024.