УДК 004.91:004.62

Oleksandr Terentiev, Doctor of Technical Sciences, Associate Professor, Principal researcher, Institute of Telecommunications and Global Information Space of the National Academy of Sciences of Ukraine
ORCID ID: https://orcid.org/0000-0002-4288-1753 ***e-mail:*** o.terentiev@gmail.com

Yurii Abroskin, graduate student, Institute of Telecommunications and Global Information Space of the National Academy of Sciences of Ukraine
ORCID ID: https://orcid.org/0009-0009-9828-5596 ***e-mail:*** abroskin21@gmail.com

Volodymyr Duda, graduate student, Institute of Telecommunications and Global Information Space of the National Academy of Sciences of Ukraine
ORCID ID: https://orcid.org/0009-0002-4278-4635 ***e-mail:*** dudavolodimir@gmail.com

Tetyana Prosyankina-Zharova, Doctor of Technical Sciences, Associate Professor, senior researcher, Institute of Telecommunications and Global Information Space of the National Academy of Sciences of Ukraine
ORCID ID: https://orcid.org/0000-0002-9623-8771 ***e-mail:*** t.pruman@gmail.com

Institute of Telecommunications and Global Information Space of the National Academy of Sciences of Ukraine, Kyiv, Ukraine

ANALYSIS OF TEXT ANALYTICS METHODS FOR KNOWLEDGE EXTRACTION FROM UKRAINIAN-LANGUAGE SOCIAL MEDIA

Abstract. *The purpose of the study is to review and systematize current text analytics and natural language processing methods for knowledge extraction from unstructured social media content, with a focus on Ukrainian-language sources.*
A comparative analysis of topic modelling methods (LSA, NMF, LDA, HDP, Top2Vec, BERTopic), ontology construction approaches, OSINT data collection tools, and the F1 evaluation metric for named entity recognition tasks was conducted.
Comparative analysis of four topic modelling methods applied to real Twitter datasets demonstrated that BERTopic (coherence score 0.62) outperforms LDA (0.45) and Top2Vec (0.56) for short texts; the NER-UK 2.0 corpus provides a baseline solution for Ukrainian named entity recognition with an F1 score of 0.89. Theoretically, the selection of methods that take into account the temporal dynamics of topics is justified. Practically, five-block pipeline architecture for knowledge extraction from Ukrainian-language social media is proposed.
The originality of the work lies in the adaptation of the Methontology-based approach to ontology generation for short unstructured Ukrainian-language texts. Further prospects include practical implementation and validation of the proposed pipeline on real Ukrainian social media datasets.
Keywords: *text analytics, data processing, Coherence Score, F1-score, LSA, NMF, LDA, Top2Vec, BERTopic, OSINT.*

О.М. Терентьєв, Ю.Ю. Аброскін, В.О. Дуда, Т.І. Просянкіна-Жарова

Інститут телекомунікацій і глобального інформаційного простору НАН України, м. Київ, Україна

АНАЛІЗ МЕТОДІВ ТЕКСТОВОЇ АНАЛІТИКИ ДЛЯ ВИДОБУВАННЯ ЗНАНЬ З УКРАЇНОМОВНОГО КОНТЕНТУ СОЦІАЛЬНИХ МЕРЕЖ

Анотація. *Мета дослідження полягає в аналізі та систематизації сучасних методів текстової аналітики для видобування знань із соціальних мереж з акцентом на україномовний контент. Було виконано порівняльний аналіз шести методів тематичного моделювання (LSA, NMF, LDA, HDP, Top2Vec, BERTopic), підходів до побудови онтологій та графів знань, інструментів OSINT, а також метрики F1 для оцінювання завдань розпізнавання іменованих сутностей.*

Порівняльний аналіз методів тематичного моделювання на реальних наборах повідомлень показав, що BERTopic (когерентність 0,62) перевищує LDA (0,45) і Top2Vec (0,56) на коротких текстах; корпус NER-UK 2.0 забезпечує базове рішення NER для української мови з точністю F1 = 0,89.

Теоретично обґрунтовано вибір методів з урахуванням часової динаміки тем, для подальшого використання в дисертаційному дослідженні. Запропоновано концептуальну архітектуру п'ятиблокового конвеєру, для практичного використання.

Оригінальність дослідження полягає в адаптації загальновідомого підходу під назвою Methontology до генерації онтологій для коротких неструктурованих україномовних текстів.

Перспективи подальшої роботи – практична реалізація та апробація конвеєру на реальних даних україномовних соціальних мереж.

Ключові слова: *текстова аналітика, обробка даних, коефіцієнт узгодженості тем, F1-метрика, LSA, NMF, LDA, Top2Vec, BERTopic, OSINT.*

https://doi.org/10.32347/2411-4049.2026.1.161-170

Вступ

Видобування знань із текстових масивів соціальних мереж є актуальним завданням, яке дозволяє аналізувати суспільні настрої, виявляти тренди та здійснювати моніторинг інформаційного простору. Як зазначають фахівці [1], соціальні платформи накопичують значні обсяги неструктурованого тексту, опрацювання якого потребує застосування комплексу методів інтелектуального аналізу даних. Для україномовного контенту зазначена задача має особливу специфіку через мовні особливості онлайн-комунікації та обмеженість спеціалізованих інструментів до 2022 року.

Завдання автоматизованого аналізу громадської думки через соціальні мережі активно досліджується як для загальних [4, 6], так і для кризових контекстів [5]. Зокрема, дослідження україномовних повідомлень засобами NLP (Natural Language Processing – обробка природньої мови), машинного навчання [2] та виявлення дезінформації в українському медіапросторі [3] демонструють зростаючий інтерес до цієї предметної галузі. Досвід авторського колективу в кластеризації новинних текстів методом SVD (Singular Value Decomposition – сингулярне розкладання) із використанням системи SAS Enterprise Miner [20] (SAS – Statistical Analysis System) також підтвердив практичну цінність автоматизованого текстового аналізу для виявлення тематичних груп і трендів.

Аналіз останніх досліджень і публікацій

Порівняльний аналіз методів тематичного моделювання для коротких текстів із соціальних мереж став предметом низки сучасних досліджень. У дослідженні [8] виконано порівняння таких методів, як LDA (Latent Dirichlet Allocation – латентне розподілення Діріхле), NMF (Non-negative Matrix Factorization – невід'ємна матрична факторизація), Top2Vec (Topic to Vector – від теми до вектору) та BERTopic (тематичне моделювання на основі BERT-овської моделі), на корпусі, що складається з 31 800 текстових повідомлень. У роботі [9] дослідниками проведено аналогічне порівняння на твітах із хештегом #covidtravel. Обидва дослідження [8, 9] продемонстрували узгоджені результати щодо переваг методів BERTopic та NMF над LDA та Top2Vec у разі обробки коротких текстів. Американські дослідники Анкан Саха та Вікас Сіндхвані в роботі [10] розробили підхід Dynamic NMF (Dynamic Non-negative Matrix Factorization) з часовою регуляризацією, призначений для відстеження еволюції тем у соціальних мережах протягом часу.

З позиції видобування знань із соціальних медіа та побудови онтологій показовою є робота [1], у якій для текстів Facebook застосовано комплекс технік кластеризації та тематичного моделювання. Огляд методів побудови онтологій на основі тексту [14] систематизує підходи – від простого вилучення термінів до повністю автоматизованих рішень на базі великих мовних моделей. Для задач побудови графів знань та встановлення семантичних зв'язків між сутностями ефективним визнано підхід, описаний у роботі [15].

У контексті україномовної обробки природної мови ключовим ресурсом є корпус NER-UK 2.0 (NER – Named Entity Recognition) [7], який забезпечує основу для розпізнавання іменованих сутностей в українських текстах різних жанрів, зокрема в дописах соціальних мереж.

Мета дослідження

Метою дослідження є аналіз та порівняння існуючих методів текстової аналітики за кількісними та якісними критеріями, а також обґрунтування вибору підходів для розробки інформаційної технології видобування знань з україномовного контенту соціальних мереж.

Теоретичні основи дослідження

Методи тематичного моделювання

Тематичне моделювання дає змогу автоматично виявляти приховані смислові структури в текстових корпусах. Методи LSA та NMF належать до алгебраїчних методів, які здійснюють факторизацію матриці «термін–документ». Метод NMF застосовує TF-IDF-зважування (Term Frequency – Inverse Document Frequency) та обмеження невід'ємності, що забезпечує вищу точність виявлення тем порівняно з методом LSA [9]. У свою чергу метод LDA є імовірнісною байєсівською моделлю, у якій документ розглядається як суміш тем, а тема – як розподіл над словниковим запасом. Для відстеження змін тем у часі розроблено Dynamic Topic Model (DTM), що розширює LDA шляхом урахування часових зрізів [10].

BERTopic підхід базується на трансформерних моделях для отримання щільних векторних представлень документів, після чого застосовує алгоритм UMAP (Uniform Manifold Approximation and Projection – рівномірна апроксимація та проекція багатовиду) для зменшення розмірності та HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise – алгоритм ієрархічної щільнісної просторової кластеризації з урахуванням шуму) для кластеризації. Порівняно з LDA та NMF, метод BERTopic краще справляється з короткими та семантично неоднозначними текстами соціальних мереж [8, 9]. Top2Vec має подібний підхід, проте поступається BERTopic у чіткості розмежування тем і генерує більшу кількість перекриваючих кластерів [8].

Метрика Coherence Score для оцінювання тематичних моделей

Для об'єктивного порівняння методів тематичного моделювання без залучення ручної експертної оцінки використовується метрика Coherence Score (коефіцієнт узгодженості тем). Задачу автоматичного оцінювання узгодженості тем уперше систематично сформулювали Девід Ньюман, Джей Хан Лау, Карл Грізер та Тімоті Болдуін в своїй роботі [11], встановивши, що міра на основі взаємної точкової інформації (PMI – Pointwise Mutual Information), розрахована на корпусі всесвітньо відомої Інтернет-енциклопедії Wikipedia, досягає рівня кореляції з оцінками людей ρ = 0,78 (за коефіцієнтом кореляції Спірмена) для новинних текстів. Пізніше дослідники Міхаель Редер, Андреас Бот та Олександр Хіннебург в своєму дослідженні [12] запропонували уніфікований фреймворк, що охоплює всі відомі міри когерентності. Найкраща з виявлених ними мір – C_v, яка поєднує непряму косинусну подібність із нормалізованою PMI (NPMI – Normalized Pointwise Mutual Information) та ковзним вікном, досягаючи середньої кореляції з людськими оцінками 0,731.

Загальна формула метрики коефіцієнта узгодженості тем UCI (скорочено C_uci від UCI coherence), для окремої теми, описаної множиною з N ключових слів $W = \{w_1, \dots, w_N\}$, визначається як нормована сума попарних PMI-оцінок [12]:

$$C_{UCI} = \frac{2}{N \cdot (N-1)} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} PMI(w_i, w_j), \qquad (1)$$

де

$$PMI(w_i, w_j) = \log \frac{P(w_i,\ w_j)+\varepsilon}{P(w_i) \cdot P(w_j)}, \qquad (2)$$

Значення ймовірностей, $P(w_i)$ – появи i −го слова та $P(w_i,\ w_j)$ – сумісної появи одночасно i −го та j −го слів, оцінюються на основі статистики спільної зустрічальності слів у ковзному вікні на великому корпусі текстів. Значення коефіцієнта узгодженості тем належить до інтервалу [0; 1], де більше значення відповідає кращій інтерпретованості теми. На матеріалі двох реальних наборів даних коротких текстів BERTopic досяг найвищого значення 0,62, в той час як LDA лише 0,45 [13].

Метрика F1 для оцінювання результатів NLP

Ключовою метрикою для оцінювання задач обробки природної мови, зокрема розпізнавання іменованих сутностей (NER), є міра F1. Вона поєднує точність (Precision, позначається як P) та повноту (Recall, позначається як R) в одному показнику. Пітер Крістен, Девід Дж. Хенд та Нішаді Кіріелле в своєму дослідженні [19] у ґрунтовному огляді зазначають, що F1 є гармонійним середнім P та R і обчислюється за формулою:

$$F_1 = \frac{2 \cdot P \cdot R}{P + R} = \frac{2 \cdot TP}{2 \cdot TP + FP + FN}, \qquad (3)$$

де TP (true positives) – кількість вірно ідентифікованих сутностей; FP (false positives) – хибно визначені сутності; FN (false negatives) – пропущені сутності. Значення $F1 \in [0; 1]$, де 1 відповідає ідеальному результату. Ця метрика вважається стандартом оцінювання NER-систем на наборах даних CoNLL (https://huggingface.co/datasets/eriktks/conll2003) та ACE (https://catalog.ldc.upenn.edu/LDC2006T06) [19], що використовуються фахівцями для тестування моделей. Для корпусу NER-UK 2.0 [7] базова модель на основі RoBERTa-large досягає зваженого значення F1 = 0,89 на 13 категоріях іменованих сутностей в українських текстах.

Побудова онтологій та графів знань

Онтологія в контексті інформаційних систем являє собою формальний опис понять предметної галузі та відносин між ними. У роботі [14] Олександр Маедхе та Штеффен Штааб виокремлюють п’ять ступенів автоматизації процесу побудови онтологій: від ручного проєктування в програмі Protege [18] до повної автоматизації на основі великих мовних моделей [16]. Для забезпечення структурованого збереження даних та їхньої сумісності використовується технологічний стек RDF (Resource Description Framework – стандарт для представлення даних у вигляді графу знань), OWL (Web Ontology Language – мова для опису онтологій) та SPARQL (SPARQL Protocol and RDF Query Language – мова запитів для роботи з даними в форматі RDF). Реалізація послідовних фаз за методологією Methontology (від специфікації до імплементації) може бути відтворена через формування структурованих інструкцій (промптів) для сучасних LLM (великих мовних моделей) [16].

Природним інструментом для візуалізації виявлених сутностей та існуючих між ними зв’язків виступають графи знань [15]. В умовах аналізу соціальних платформ вузлами графа стають відповідні концепти (особи, локації, організації чи події), а ребрами – встановлені семантичні відношення. Поєднання тематичного моделювання з графовими технологіями дозволяє не лише окреслювати коло тем, а й фіксувати концептуальні взаємозв’язки між ними.

Результати дослідження

Порівняльний аналіз методів тематичного моделювання

На основі проведеного огляду літератури складено порівняльну характеристику шести методів тематичного моделювання за п’ятьма якісними критеріями (таблиця 1).

Таблиця 1. Якісне порівняння методів тематичного моделювання

Метод	**Підхід**	**Динаміка у часі**	**Придатність для коротких текстів**	**Підтримка укр. мови**
LSA	Матрична факторизація (SVD)	Обмежена	Слабка	Потребує адаптації
NMF	Матрична факторизація (TF-IDF)	Обмежена	Добра	Потребує адаптації
LDA	Баєсівський генеративний	Через DTM	Середня	Так (з корпусом)
HDP	Непараметричний байєсівський	Часткова	Середня	Так (з корпусом)
Top2Vec	Векторні вкладення	Обмежена	Добра	Так (mBERT)
BERTopic	Трансформерні моделі	Висока	Висока	Так (uk-BERT, mBERT)

Кількісне порівняння методів тематичного моделювання здійснено на підставі даних дослідження з роботи [11], виконаного на двох реальних наборах даних, що містять короткі тексти. Перший набір даних складається з 29 200 коротких відгуків користувачів державного порталу ОАЕ, другий набір даних складається з 1 600 готельних відгуків платформи TripAdvisor. Оцінювання проводилось при двох значеннях кількості тем (k = 5 та k = 10) за метрикою коефіцієнта узгодженості тем. Отримані результати порівняння наведені в таблиці 2.

Таблиця 2. Значення метрики коефіцієнта узгодженості тем для різних методів тематичного моделювання [13]

Метод	Перший набір даних, k = 5	Перший набір даних, k = 10	Другий набір даних, k = 5	Другий набір даних, k = 10
LSA	0,5	0,53	0,46	0,39
NMF	0,49	0,5	0,33	0,31
LDA	0,45	0,4	0,42	0,48
PAM	0,49	0,44	0,32	0,31
Top2Vec	0,56	0,54	0,38	0,33
BERTopic	0,62	0,56	0,58	0,6

З таблиці 2 видно, що BERTopic демонструє найвище значення коефіцієнта узгодженості тем на обох датасетах (0,62 та 0,58 при k = 5), що підтверджує його перевагу для коротких неструктурованих текстів. Примітно, що метод NMF показує близькі до BERTopic результати на першому наборі даних (0,49), однак суттєво поступається на другому наборі даних (0,33) – тобто його ефективність залежить від характеру вхідних даних. LDA демонструє відносно стабільні результати на обох наборах даних і зберігає перевагу для часового аналізу завдяки розширенню DTM [10]. Окремо варто зазначити, що для задачі розпізнавання іменованих сутностей в україномовних текстах корпус NER-UK 2.0 забезпечує базовий рівень F1 = 0,89 на 13 категоріях сутностей [7], що підтверджує достатність наявної NLP-інфраструктури для реалізації запропонованої архітектури.

Аналіз підходів до побудови онтологій

Порівняння підходів до побудови онтологій здійснювалося за трьома критеріями: рівень автоматизації, адаптованість до неструктурованого контенту та підтримка україномовних текстів. Ручний підхід на основі використання програми Protege [18] забезпечує найвищу якість, але потребує значних зусиль експерта з відповідної предметної області. NER-технологічні процеси на базі використання NER-UK 2.0 [7] є частково автоматизованим та безпосередньо придатним для україномовних текстів. LLM-підхід [16] із структурованими Methontology-промптами є найбільш автоматизованим, проте розроблявся для структурованих англомовних документів і потребує адаптації для коротких дописів у соціальних мережах.

Концептуальна архітектура інформаційної технології

На основі проведеного аналізу запропоновано концептуальну архітектуру у вигляді п’ятиблокового конвеєру для видобування знань з україномовного контенту соціальних мереж (рис. 1).

Блок 1. Збір даних (OSINT-модуль). Автоматизований збір текстового контенту та метаданих з платформ Telegram, YouTube, Facebook через офіційні API. Разом із текстом фіксуються метрики залученості (перегляди, реакції, коментарі, репости) та часова мітка публікації.

Блок 2. Попередня NLP-обробка. Токенізація, лематизація, морфологічний аналіз та видалення стоп-слів із використанням інструментарію для укр. мови: моделей на базі BERT та корпусу NER-UK 2.0 [7] для розпізнавання іменованих сутностей.

Блок 3. Тематичне моделювання та кластеризація. Виявлення латентних тематичних структур із використанням LDA/HDP або BERTopic. Для аналізу динаміки застосовується часово-зрізовий підхід [10]: корпус розбивається на часові вікна, а міжвіконне зіставлення дозволяє відстежувати еволюцію тематик.

Блок 4. Побудова онтологій та графів знань. На основі виявлених тем та іменованих сутностей автоматично будуються онтологічні структури у форматі OWL за принципами Methontology [16]: (1) глосарій термінів, (2) таксономія концептів, (3) ситуативні зв’язки. Графи знань формуються на основі семантичної близькості та спільної зустрічальності сутностей у темах.

Блок 5. Візуалізація та інтерпретація. Відображення виявлених тем, часової динаміки та концептуальних зв'язків. Інтеграція метрик залученості дозволяє ранжувати теми за ступенем резонансності в аудиторії.

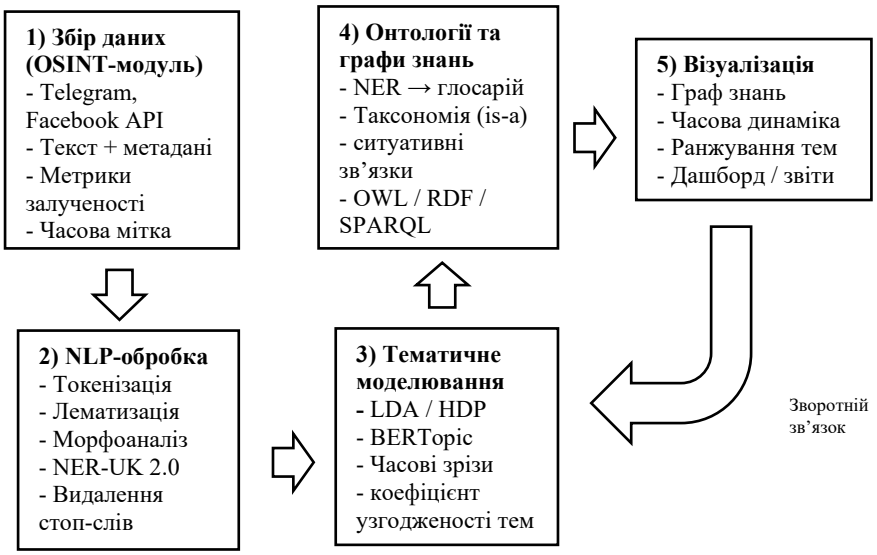

Рис. 1. Концептуальна архітектура конвеєрної системи для видобування знань з україномовного контенту соціальних мереж

Висновки

У статті проведено аналіз та порівняння методів текстової аналітики для задачі видобування знань з україномовного контенту соціальних мереж. Отримано наступні результати:

1. за результатами кількісного порівняння [13] встановлено, що BERTopic (коефіцієнт узгодженості тем = 0,62) перевищує LDA (0,45) та Top2Vec (0,56) для коротких неструктурованих текстів, що обґрунтовує вибір BERTopic як основного методу тематичного моделювання в запропонованій архітектурі;

2. показано, що корпус NER-UK 2.0 [7] із базовим F1 = 0,89 є практичним інструментом для розпізнавання іменованих сутностей у 13 категоріях українських текстів, включно з контентом соціальних мереж;

3. обґрунтовано, що LLM-підхід до автоматичної побудови онтологій [14] на базі відомої методики Methontology є перспективним, але потребує адаптації під специфіку коротких україномовних текстів соціальних мереж – на відміну від структурованих технічних документів, для яких він розроблявся;

4. запропоновано концептуальну архітектуру п'ятиблокового конвеєру (OSINT, NLP-обробка, тематичне моделювання, онтології, візуалізація), що інтегрує метрики залученості аудиторії та дозволяє відстежувати часову динаміку тем.

Практичне значення полягає в тому, що запропонована архітектура може стати основою для розробки системи моніторингу україномовного інформаційного простору. Перспективи подальших досліджень – практична реалізація конвеєру, дослідження ефективності BERTopic із моделями uk-BERT на реальних україномовних даних соціальних мереж та розробка спеціалізованих онтологій для обраних предметних областей.

СПИСОК ЛІТЕРАТУРИ / REFERENCES

1. Salloum, S. A., Al-Emran, M., & Shaalan, K. (2017). Mining social media text: Extracting knowledge from Facebook. *International Journal of Computing and Digital Systems*, 6(2), 73–81. https://www.researchgate.net/publication/314095118_Mining_Social_Media_Text_Extracting_Knowledge_from_Facebook
2. Prokipchuk, O., Vysotska, V., Pukach, P., Lytvyn, V., Uhryn, D., Ushenko, Yu., & Hu, Z. (2023). Intelligent analysis of Ukrainian-language tweets for public opinion research based on NLP methods and machine learning technology. *International Journal of Modern Education and Computer Science*, 15(3), 70–93. https://doi.org/10.5815/ijmecs.2023.03.06
3. Vysotska, V., Przystupa, K., Kulikov, Yu., Chyrun, S., Ushenko, Yu., Hu, Z., & Uhryn, D. (2025). Recognizing fakes, propaganda and disinformation in Ukrainian content based on NLP and machine-learning technology. *International Journal of Computer Network and Information Security*, 17(1), 92–127. https://doi.org/10.5815/ijcnis.2025.01.08
4. Vysotska, V., Mazepa, S., Chyrun, L., Brodyak, O., Shakleina, I., & Schuchmann, V. (2022). NLP tool for extracting relevant information from criminal reports or fakes/propaganda content. *Proceedings of the IEEE 17th International Conference on Computer Sciences and Information Technologies (CSIT)*, 93–98. https://doi.org/10.1109/CSIT56902.2022.10000563
5. Ozyurt, B., & Akcayol, M. A. (2023). A deep learning-based sentiment analysis approach (MF-CNN-BILSTM) and topic modeling of tweets related to the Ukraine-Russia conflict. *Applied Soft Computing*, 143, 110404. https://doi.org/10.1016/j.asoc.2023.110404
6. Liao, H., Wang, C., Gu, Y., & Liu, R. (2025). A text data mining-based digital transformation opinion thematic system for online social media platforms. *Systems*, 13(3), 159. https://doi.org/10.3390/systems13030159
7. Chaplynskyi, D., & Romanyshyn, M. (2024). Introducing NER-UK 2.0: A rich corpus of named entities for Ukrainian. *Proceedings of the Third Ukrainian Natural Language Processing Workshop (UNLP) @ LREC-COLING 2024*, 23–29. https://aclanthology.org/2024.unlp-1.4
8. Ramamoorthy, T., Kulothungan, V., & Mappillairaju, B. (2024). Topic modeling and social network analysis approach to explore diabetes discourse on Twitter in India. *Frontiers in Artificial Intelligence*, 7, 1329185. https://doi.org/10.3389/frai.2024.1329185
9. Egger, R., & Yu, J. (2022). A topic modeling comparison between LDA, NMF, Top2Vec, and BERTopic to demystify Twitter posts. *Frontiers in Sociology*, 7, 886498. https://doi.org/10.3389/fsoc.2022.886498
10. Saha, A., & Sindhwani, V. (2012). Learning evolving and emerging topics in social media: A dynamic NMF approach with temporal regularization. *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining (WSDM '12)*, 693–702. https://doi.org/10.1145/2124295.2124376
11. Newman, D., Lau, J. H., Grieser, K., & Baldwin, T. (2010). Automatic evaluation of topic coherence. *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL*, 100–108. https://aclanthology.org/N10-1012
12. Röder, M., Both, A., & Hinneburg, A. (2015). Exploring the space of topic coherence measures. *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining (WSDM '15)*, 399–408. https://doi.org/10.1145/2684822.2685324
13. Krishnan, A., & Kennedyraj. (2023). Exploring the power of topic modeling techniques: A comparative analysis. *arXiv preprint* arXiv:2308.11520. https://arxiv.org/abs/2308.11520
14. Maedche, A., & Staab, S. (2001). Ontology learning from text: A survey. *IEEE Intelligent Systems*, 16(4), 72–79. https://doi.org/10.1007/3-540-45399-7_30
15. Ji, S., Pan, S., Cambria, E., Marttinen, P., & Philip, S. Y. (2021). A survey on knowledge graphs: Representation, acquisition, and applications. *IEEE Transactions on Neural Networks and Learning Systems*, 33(2), 494–514. https://doi.org/10.1109/TNNLS.2021.3070843

16. Tupayachi, J., Xu, H., Omitaomu, O. A., Camur, M. C., Sharmin, A., & Li, X. (2024). Towards next-generation urban decision support systems through AI-powered construction of scientific ontology using large language models. *arXiv preprint* arXiv:2405.19255. https://doi.org/10.48550/arXiv.2405.19255
17. Boutaleb, A., Picault, J., & Grosjean, G. (2024). BERTrend: Neural topic modeling for emerging trends detection. *arXiv preprint* arXiv:2411.05930. https://arxiv.org/abs/2411.05930
18. Noy, N. F., Sintek, M., Decker, S., Crubezy, M., Fergerson, R. W., & Musen, M. A. (2001). Creating semantic web contents with Protege-2000. *IEEE Intelligent Systems*, 16(2), 60–71. https://doi.org/10.1109/5254.920601
19. Christen, P., Hand, D. J., & Kirielle, N. (2023). A review of the F-measure: Its history, properties, criticism, and alternatives. *ACM Computing Surveys*, 56(3), 73. https://doi.org/10.1145/3606367
20. Mühlroth, C., & Grottke, M. (2022). Artificial intelligence in innovation: How to spot emerging trends and technologies. *IEEE Transactions on Engineering Management*, 69(2), 493–510. https://doi.org/10.1109/TEM.2020.2989214
21. Terentiev, O. M., Duda, V. O., & Abroskin, Yu. Yu. (2025). Analiz tekstovoi informatsii z metoiu klasteryzatsii ta vyiavlennia hrup ekonomichnykh novyn shchodo auktsioniv Ministerstva finansiv iz zaluchennia zovnishnoho finansuvannia [Analysis of textual information for clustering and identification of groups of economic news on Ministry of Finance auctions for external financing]. Development of Education, Science and Business: Results 2025: *Proceedings of the International Scientific and Practical Internet Conference*, December 18–19, 2025, 511–513. FOP Marenichenko V.V., Dnipro, Ukraine. ISBN 978-617-8293-60-4. ISSN 2664-4819. http://www.wayscience.com/wp-content/uploads/2025/12/Conference-Proceedings-December-18-19-2025.pdf (in Ukrainian)

Стаття надійшла до редакції 13.01.26, надійшла після рецензування 16.02.26, прийнята 06.03.26

The article was received 13.01.26, received after revision 16.02.26, accepted 06.03.26

Терентьєв Олександр Миколайович
доктор технічних наук, доцент, провідний науковий співробітник, Інститут телекомунікацій і глобального інформаційного простору НАН України
Адреса робоча: бульв. Чоколівський, 13, Київ, Україна, 03186
ORCID ID: https://orcid.org/0000-0002-4288-1753 ***e-mail:*** o.terentiev@gmail.com

Аброскін Юрій Юрійович
аспірант, Інститут телекомунікацій та глобального інформаційного простору НАН України
Адреса робоча: бульв. Чоколівський, 13, Київ, Україна, 03186
ORCID ID: https://orcid.org/0009-0009-9828-5596 ***e-mail:*** abroskin21@gmail.com

Дуда Володимир Олександрович
аспірант, Інститут телекомунікацій і глобального інформаційного простору НАН України
Адреса робоча: бульв. Чоколівський, 13, Київ, Україна, 03186
ORCID ID: https://orcid.org/0009-0002-4278-4635 ***e-mail:*** dudavolodimir@gmail.com

Просянкіна-Жарова Тетяна Іванівна
доктор технічних наук, доцент, старший науковий співробітник, Інститут телекомунікацій і глобального інформаційного простору НАН України
Адреса робоча: бульв. Чоколівський, 13, Київ, Україна, 03186
ORCID ID: https://orcid.org/0000-0002-9623-8771 ***e-mail:*** t.pruman@gmail.com