DOI: https://doi.org/10.15276/ict.01.2024.25
УДК 004.93

Порівняння текстової інформації з інформаційних джерел на основі алгоритму косинусної подібності

Угрин Дмитро Ілліч[1)]
Д-р техніч. наук, професор каф. Комп'ютерних наук
ORCID: https://orcid.org/0000-0003-4858-4511; d.ugryn@chnu.edu.ua. Scopus Author ID: 57163746300
Каланча Артем Дмитрович[1)]
Аспірант каф. Програмного забезпечення комп'ютерних систем
ORCID: https://orcid.org/0009-0004-1451-7470; kalancha.artem@chnu.edu.ua
[1)] Чернівецький національний університет ім. Ю. Федьковича, вул. Коцюбинського, 2. Чернівці, 58012, Україна

АНОТАЦІЯ

У тезах представлено дослідження, спрямоване на розробку оптимальної концепції аналізу та порівняння інформаційних джерел на основі великих обсягів текстової інформації за допомогою методів обробки природної мови (NLP). Об'єктом дослідження стали новинні Telegram-канали, які використовуються як джерела текстових даних. Було здійснено попередню обробку текстів, включаючи очищення, токенізацію та лематизацію, для формування глобального словника, що складається з унікальних слів усіх інформаційних джерел. Для кожного джерела було побудовано векторне представлення текстів, розмірність якого відповідає кількості унікальних слів у глобальному словнику. Частота вживання кожного слова у текстах каналу відображалася у відповідних позиціях вектора. Застосувавши алгоритм косинусної подібності до пар векторів, було отримано квадратну матрицю, яка демонструє ступінь подібності між різними джерелами. Результати дослідження показують ефективність запропонованого підходу для кількісної оцінки подібності текстових даних з різних джерел. Виявлено необхідність подальшої оптимізації алгоритму, зокрема шляхом параметризації для досягнення балансу між точністю та обчислювальними витратами, а також відокремлення слів з надмірною вагою, таких як специфічні терміни або назви каналів. Запропонований метод може бути застосований для аналізу інформаційних потоків, виявлення взаємозв'язків між джерелами та дослідження соціально-культурного впливу медіа-контенту в умовах сучасного інформаційного середовища.

Ключові слова: інформаційне джерело; текст; подібність; обробка природної мови; попередня обробка тексту; Telegram; векторизація; косинусна подібність

Актуальність даного дослідження обумовлена значним соціально-культурним впливом інформаційних джерел, які формують громадську думку та поведінку. Стрімкий розвиток штучного інтелекту, зокрема в галузі аналізу текстової інформації [1], відкриває нові можливості для розуміння та обробки великих обсягів даних. Крім того, постійна тендеція до збільшення генерації об'ємів даних при відносно повільному збільшенню обчислювальних можливостей стимулює наукову спільноту знаходити ефективні методи обробки даних, аналіз взаємозв'язків та прихованих патернів.

Метою дослідження є розробка концепції оптимального аналізу та порівняння змісту інформаційних джерел на основі значних обсягів текстової інформації за допомогою засобів обробки природної мови (NLP) [2].

Зростання обсягів інформації, яка генерується у соціальних мережах, викликає підвищений інтерес до автоматизованих методів обробки та аналізу текстових даних [3]. Соціальні медіа, такі як Telegram [4], стають важливим джерелом інформації, що вимагає розробки нових підходів до обробки природної мови (NLP) для збору та аналізу інформації. Основна задача полягає в автоматизації процесу порівняння подібності між різними інформаційними джерелами.

Для аналізу новинних повідомлень у цьому дослідженні використовується вибірка з кількох Telegram-каналів. Вибірка формується на основі популярних каналів, що регулярно публікують новини та мають значну кількість підписників. Процес збору даних включає автоматизований витяг текстових повідомлень, а також метаданих, таких як час публікації.

Після етапу збору даних необхідно провести попередню обробку текстової інформації [5, 6, 7], що є надзвичайно важливим кроком у процесі підготовки даних для подальшого аналізу.

Цей етап включає низку процедур, кожна з яких спрямована на поліпшення якості вхідних даних та підвищення точності наступних аналітичних операцій. Попередня обробка дозволяє очищати дані від зайвих елементів, структурувати їх та забезпечити відповідність вимогам методів аналізу, що використовуються:

Очищення тексту – на першому етапі відбувається видалення непотрібних символів, пунктуації, стоп-слів та інших елементів, які не несуть інформаційного навантаження. Це дозволяє значно зменшити обсяг даних і зосередитися на ключових словах. Наприклад, з тексту можуть бути видалені символи емодзі, HTML-теги та спеціальні символи, що часто зустрічаються у повідомленнях.

Токенізація – процес токенізації включає розбиття тексту на окремі слова, які називаються токенами. Токенізація є необхідною для побудови подальших моделей, оскільки дозволяє представити текст у вигляді послідовності токенів. У нашому випадку, кожне повідомлення з Telegram-каналу перетворюється на масив токенів, що дозволяє зберегти текстову інформацію у структурованому вигляді.

Побудова словника – на основі очищених даних будується глобальний словник $V = [v_1, v_2, \ldots, v_n]$, що включає всі унікальні слова, які зустрічаються у текстах усіх каналах (де n - це кількість унікальних слів). Словник представляється у вигляді вектора, де кожне слово преставляє окремий вимір, а значення виміру – кількість траплянь слова у всій множині джерел, що дозволяє у подальшому створювати векторні представлення лексичного складу.

Побудова векторів для джерел – для кожного джерела будується вектор $K_i = [k_1, k_2, k_3, \ldots, k_n]$, розмірність якого дорівнює розмірності глобальному словнику V. Кожна простір у векторі відповідає простору глобального словника, а значення просторі – частота вживання певного слова у тексті каналу. Таким чином, вектор відображає особливості словникового запасу інформаційного джерела та може бути використаний для порівняння з іншими векторами, побудованими на основі інших інформаційних джерел.

Таким чином, вектор представляє собою структуроване відображення словникового запасу кожного інформаційного джерела, дозволяючи оцінювати та порівнювати тексти за допомогою математичних методів. Наприклад, для вимірювання подібності між джерелами може використовуватися косинусна подібність [8], яка дозволяє оцінити ступінь подібності між векторами на основі їхнього напрямку у багатовимірному просторі (див. формулу 1, де A, B – вектори). Це дозволяє точно визначити, наскільки близькими є контент та тематика різних інформаційних джерел, що, в свою чергу, дає можливість робити висновки щодо їхньої взаємодії чи спільних інформаційних джерел (Рис. 1).

$$\cos(\theta) = \frac{A \cdot B}{||A|| \cdot ||B||} \tag{1}$$

Косинусна подібність набуває значення від *-1* до *1*, де:

- *1* – означає повну ідентичність векторів (і, відповідно, повну подібність каналів),
- *0* – відсутність подібності (вектори ортогональні),
- *-1* – повну протилежність векторів (і, відповідно, вміст каналів повністю різний).

Зменшення розмірності векторів: оскільки словник може містити десятки тисяч унікальних слів, вектори мають високу розмірність. Для підвищення ефективності аналізу та зменшення обчислювальних витрат пропонуються наступні методи зменшення розмірності.

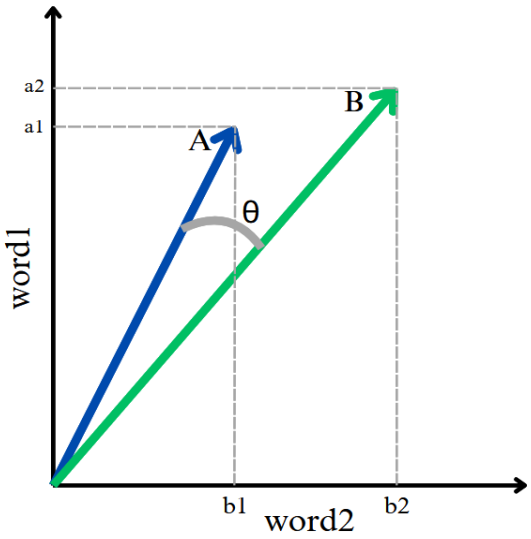

Рис.1. **Порівняння подібності векторів двох інформаційних джерел, побудованих на основі використовуваного словника**

Це досягається шляхом зменшення розміру початкового словника. Існує два основних підходи до цього:

Видалення рідковживаних слів. Слова, які мають найменшу частотність у всіх каналах, можуть бути видалені. Наприклад, це можуть бути слова, які повторюються не більше п'яти разів у кожному каналі за весь період спостереження.

Видалення рівномірно розподілених слів. Слова, які рівномірно або майже рівномірно розподілені по всіх каналах, також можуть бути видалені, оскільки вони не несуть значущої інформації для порівняння.

Враховуючи, що деякі Telegram-канали можуть змінювати своїх авторів або тематику протягом часу, важливо аналізувати зміни у тексті протягом певного періоду. Часовий аналіз дозволяє визначити, як змінювався словниковий запас каналу у різні періоди та як ці зміни впливають на його подібність до інших каналів.

Розбиття на часові інтервали. Для проведення аналізу часових змін всі повідомлення каналу можна розбити на часові інтервали, наприклад, місяці або квартали. Це дозволить побудувати вектор для кожного інтервалу та порівнювати їх між собою.

Аналіз динаміки подібності. На основі векторів для різних часових інтервалів можна аналізувати, як змінювалась подібність між каналами протягом часу. Це дозволить робити висновки про можливі зміни у тематиці або аудиторії каналу, а також про взаємний вплив каналів один на одного.

На основі результатів аналізу можна формулювати гіпотези про взаємозв'язок та вплив різних Telegram-каналів. Наприклад, якщо два канали демонструють високу подібність у певний період часу, це може свідчити про те, що вони взаємодіяли або використовували спільні джерела інформації. Також можна дослідити, як зміна авторів або тематики впливає на подібність текстів. Якщо у певний момент часу відбувається різка зміна у словниковому запасі каналу, це може бути пов'язано зі зміною редакційної політики або аудиторії.

Тим не менш, розглянемо застосування методів NLP для порівняння інформаційних джерел на основі текстової інформації. Основна увага приділяється векторизації та використанню косинусної подібності для порівняння текстів на подібність. Запропоновані методи дозволяють ефективно обробляти великі обсяги текстової інформації та отримувати цінні результати для подальшого аналізу.

Таблиця. **Інформаційні джерела та їхній об’єм даних**

Джерело	Кількість повідомлень
UaOnlii	2014
UkraineNow	1670
kievreal1	1565
lachentyt	934
suspilnenews	1088
truexanewsua	1474
ukr24_7	1063
voynareal	2307

У проведеному експерименті було здійснено аналіз текстової інформації з восьми Telegram-каналів, кожен з яких мав різний обсяг вмісту, виражений у кількості повідомлень (див. Таблицю). Для аналізу застосовували описаний вище алгоритм векторизації текстів. Зокрема, було використано параметр мінімальної кількості траплянь токена, який встановлено на рівні 220 траплянь. В результаті для кожного каналу було сформовано вектор K_i розмірності 81, де кожна компонента вектора відображала частоту вживання певного слова з глобального словника у тексті каналу.

Наступним етапом аналізу було застосування алгоритму косинусної подібності для кожної пари отриманих векторів. Це дозволило створити квадратну матрицю, де кожен елемент відображав ступінь подібності між двома інформаційними джерелами (Рис. 2). Матриця наочно демонструє взаємозв'язки між каналами на основі їхнього лексичного складу.

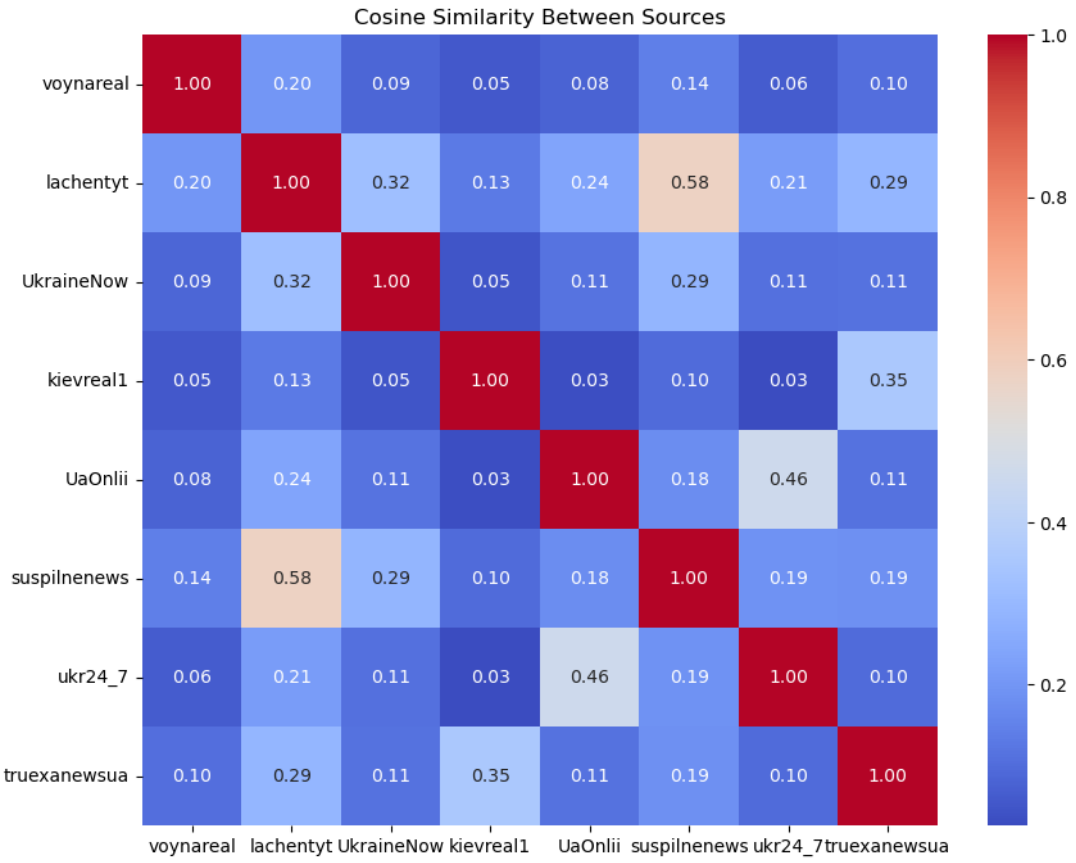

Рис. 2. **Теплова матриця**
Подібність джерел на основі їхнього лексичного складу

Висновки. На основі проведеного дослідження було запропоновано та реалізовано алгоритм оцінки подібності між двома джерелами інформації за допомогою методу косинусної подібності. У подальших дослідженнях планується розширити логіку порівняння векторів, розбивши дані
на рівномірні часові проміжки, параметризувати цей алгоритм, щоб знайти оптимальний баланс між точністю результатів та ефективністю обчислень. Особливу увагу буде приділено відокремленню слів, які можуть мати надмірну вагу в аналізі (наприклад, специфічні терміни, що характерні лише для одного каналу, його назва або девіз), з метою покращення точності оцінки подібності.

СПИСОК ЛІТЕРАТУРИ

1. Johri P., Khatri S. K., Al-Taani A. T., Sabharwal M., Suvanov S., Kumar A. "Natural Language Processing: History, Evolution, Application, and Future Work". *Proceedings of 3rd International Conference on Computing Informatics and Networks. Lecture Notes in Networks and Systems* / Ed. by A. Abraham, O. Castillo, D. Virmani. Springer, Singapore. 2021; 167: 365–375. DOI: https://doi.org/10.1007/978-981-15-9712-1_31.

2. Singh R., Singh S. "Text similarity measures in news articles by vector space model using NLP". *J. Inst. Eng. India*. 2021; Ser. B 102: 329–338. DOI: https://doi.org/10.1007/ s40031-020-00501-5.

3. Talakh M. V. "Part 7. Using text mining for the analysis of social networks". *Information Technologies Part 1. Application in Computer Vision, Recognition and Intelligent Monitoring Systems Yuriy Ushenko, Serhiy Ostapov, Serhiy Golub* / Ed. by Y. Ushenko, S. Ostapov, S. Golub. 2019. p. 157-173.

4. "Telegram (n.d.). Telegram APIs". 2024. – Available from: https://core.telegram.org/api.

5. Camacho-Collados J. Pilevar, M. T. "On the role of text preprocessing in neural network architectures: An Evaluation study on text categorization and sentiment analysis". *eprint arXiv.* 2018. DOI: https://doi.org/10.48550/arXiv.1707.01780.

6. Mohammad F. "Is preprocessing of text really worth your time for online comment classification?". *eprint arXiv.* 2018. DOI: https://doi.org/10.48550/arXiv. 1806.02908.

7. Chai C. "Comparison of text preprocessing methods". *Natural Language Engineering.* 2023. 29 (3): 509-553. DOI: https://doi.org/10.1017/S1351324922000213.

8. Pal S., Chang M., Iriarte M. F. "Summary generation using natural language processing techniques and cosine similarity". *Intelligent Systems Design and Applications. ISDA 2021. Lecture Notes in Networks and Systems* / Ed. by A. Abraham, N. Gandhi, T. Hanne, T. Hong, T. Nogueira Rios, W. Ding. Springer, Cham. 2022; 418: 508-517. DOI: https://doi.org/ 10.1007/978-3-030-96308-8_47.

DOI: https://doi.org/10.15276/ict.01.2024.25
UDC 004.93

Comparison of text information from information sources based on the cosine similarity algorithm

Dmytro I. Uhryn [1)]
Dr. Sc., Professor, Department of Computer Sciences
ORCID: https://orcid.org/0000-0003-4858-4511; d.ugryn@chnu.edu.ua. Scopus Author ID: 57163746300

Artem D. Kalancha [1)]
PhD student, Department of Computer Systems Software
ORCID: https://orcid.org/0009-0004-1451-7470; kalancha.artem@chnu.edu.ua
[1)] Yuri Fedkovich Chernivtsi National University, 2 Kotsyubinsky Str. Chernivtsi, 58012, Ukraine

ABSTRACT

This article presents research aimed at developing an optimal concept of analysis and comparison of information sources based on large volumes of textual information using natural language processing (NLP) methods. The object of the study was Telegram news channels, which are used as sources of text data. Texts were pre-processed, including cleaning, tokenization, and lemmatization, to form a global dictionary consisting of unique words from all information sources. For each source, a vector representation of the texts was built, the dimension of which corresponds to the number of unique words in the global dictionary. The frequency of use of each word in the channel's texts was displayed in the corresponding positions of the vector. By applying the cosine similarity algorithm to pairs of vectors, a square matrix was obtained that demonstrates the degree of similarity between different sources. The results of the study show the effectiveness of the proposed approach for quantitative assessment of the similarity of textual data from different sources. The need for further optimization of the algorithm was identified, in particular by parameterization to achieve a balance between accuracy and computational cost, as well as the separation of words with excessive weight, such as specific terms or channel names. The proposed method can be applied to the analysis of information flows, the identification of relationships between sources and the study of the socio-cultural influence of media content in the conditions of the modern information environment.

Keywords: Information source; text, similarity; natural language processing; text preprocessing; Telegram; vectorization; cosine similarity