Issue 17, 2025

https://doi.org/10.23939/sisn2025.17.034

UDC 004.9:519.8

COMPARISON AND CLUSTERING OF TEXTUAL INFORMATION SOURCES BASED ON THE COSINE SIMILARITY ALGORITHM

Zhengbing Hu[1], Dmytro Uhryn[2], Artem Kalancha[3]

[1] Hubei University of Technology,
School of Computer Science, Wuhan, China
[2-3] Yuri Fedkovich Chernivtsi National University,
Department of Computer Systems Software, Chernivtsi, Ukraine
[1] E-mail: drzbhu@gmail.com, ORCID: 0000-0002-6140-3351
[2] E-mail: d.ugryn@chnu.edu.ua, ORCID: 0000-0003-4858-4511
[3] E-mail: kalancha.artem@chnu.edu.ua, ORCID: 0009-0004-1451-7470

This article presents a study aimed at developing an optimal concept for analyzing and comparing information sources based on large amounts of text information using natural language processing (NLP) methods. The object of the study was Telegram news channels, which are used as sources of text data. Pre-processing of texts was carried out, 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 texts was constructed, 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 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. An analysis of the similarity of channels in limited time intervals was conducted, which allowed us to identify trends in changes in their information policies. The model parameters were optimized to ensure maximum channel differentiation, which increased the efficiency of the analysis. Clustering algorithms were applied, which divided the channels into groups according to the degree of lexical similarity. The results of the study demonstrate the effectiveness of the proposed approach for quantitatively assessing the similarity and clustering text data from different sources. The proposed method can be used to analyze information sources, identify relationships between sources, study the dynamics of changes in their activities, and assess the socio-cultural impact of media content.

Keywords: information source, text, similarity, natural language processing, text preprocessing, Telegram, vectorization, cosine similarity, clustering.

Problem Statement

The relevance of this study arises from the increasing socio-cultural influence of information sources, which actively shape public opinion, influence societal behavioral patterns, and determine the information agenda. In the modern digital environment, saturated with an immense number of sources, it is critically important not only to analyze their content but also to understand the relationships between them. The rapid development of artificial intelligence technologies, particularly in natural language processing (NLP), provides powerful tools for studying and modeling these complex interactions.

The significance of this article is further underscored by the proliferation of information sources, many of which duplicate content from others or belong to organized groups of influence that pursue political, economic, or social objectives. These groups leverage media platforms to manipulate public opinion,

disseminate ideologies, or control information flows. This complicates the analysis and comprehension of the true origins of information, as well as its uniqueness, reliability, and purpose, making the study of such sources even more critical.

Another factor contributing to the relevance of this research is the continuous increase in the volume of text data generated online. The rate of information growth far exceeds the pace of advancements in computing capabilities, necessitating the search for new, more efficient approaches to data processing. The demand for computational resource optimization, combined with the need to analyze vast amounts of textual information, drives the scientific community to develop methods that can extract the most significant aspects of information and uncover hidden patterns in big data.

Analysis of Recent Studies and Publications

This research aims to develop a concept for the optimal analysis and comparison of information source content based on large volumes of textual data using natural language processing (NLP) tools (Talakh, 2019). The increasing volume of information generated on social networks has led to growing interest in automated methods for processing and analyzing textual data (Talakh, Holub, Lazarenko, n.d). Social media platforms, such as Telegram (Telegram, 2025), are becoming critical sources of information, necessitating the development of new NLP approaches for data collection and analysis. The primary objective is to automate comparing similarities between different information sources while considering computational complexity and available resources.

To analyze news messages, this study utilizes a sample of several Telegram channels. The sample is selected based on popular channels that regularly publish news and have a significant number of subscribers. The data collection process involves the automated extraction of text messages along with metadata, such as publication timestamps.

Following the data collection stage, it is essential to perform text pre-processing (Chai, 2023) (Mohammad, 2018) (Camacho-Collados, 2018), which is a crucial step in preparing data for further analysis. This stage consists of multiple procedures, each aimed at improving the quality of the input data and enhancing the accuracy of subsequent analytical operations. Pre-processing enables data cleaning, structuring, and compliance with the requirements of the applied analytical methods. The key steps include:

Text Cleaning – in the initial stage, unnecessary characters, punctuation, stop words, and other non-informative elements are removed. This significantly reduces data volume and enhances the focus on key terms. For instance, emoji characters, HTML tags, and special symbols commonly found in messages are eliminated from the text.

Tokenization – this process involves breaking text into individual words, referred to as tokens. Tokenization is essential for building subsequent models, as it represents text as a sequence of tokens. In this study, each Telegram message is converted into an array of tokens, allowing text information to be stored in a structured format.

Formulation of the Article's Objective

The aim of the article is to develop and compare methods for clustering textual information sources using the cosine similarity algorithm, with the goal of improving the processes of text data analysis and grouping for further use in information systems and analytical tools.

Main Results

Applying cosine similarity

The first step in comparing information sources is to construct a global dictionary based on textual data from all sources, capturing all unique words that appear across all channels. The result is a global dictionary $V = [v_1, v_2, \ldots, v_n]$, where each word v_i represents a unique term extracted from the complete text

corpus, and n denotes the total number of unique words. The dictionary is structured as a multidimensional vector, where each word corresponds to a separate dimension, and the numerical value of each dimension reflects the frequency of occurrence of the corresponding word across all sources.

Construction of Vectors for Sources. Once the global dictionary is defined, each individual source (such as a channel, article, or document) can be represented as a vector $K_i = [k_1, k_2, k_3, \ldots, k_n]$, where the dimensionality of this vector corresponds to the size of the global dictionary V. Each element k_i in this vector reflects the frequency of a word corresponding to a specific dimension of the dictionary.

Thus, a vector provides a structured representation of the vocabulary of each information source, enabling the evaluation and comparison of texts through mathematical methods. For instance, to measure the similarity between sources, cosine similarity can be applied (Magara, 2018) (Park, 2020), which quantifies the degree of similarity between vectors based on their orientation in multidimensional space (see Formula 1, where A and B are vectors). This approach allows for an accurate assessment of the content and thematic closeness of different information sources, thereby facilitating conclusions about their interconnections or shared information sources (see Fig. 1). In this figure, the vectors represent two distinct textual sources. The X and Y axes correspond to word from the global dictionary in each source, illustrating how each dimension captures word usage patterns that inform the cosine similarity measure. This visualization specifically exemplifies the conceptual methodology rather than presenting an actual comparative result between specific Telegram channels.

(1)

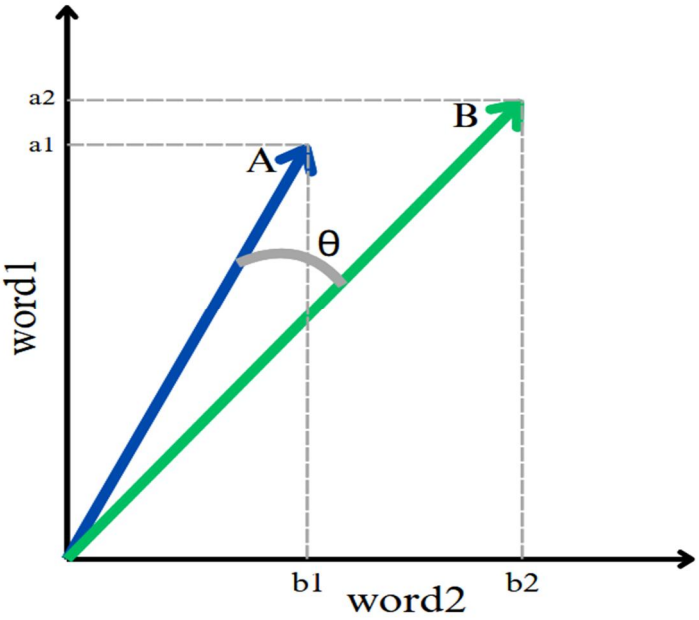

Fig. 1. Comparison of similarity vectors of two information sources, built on the basis of the dictionary used

Cosine similarity takes on a value from -1 to 1, where:

1 – means complete identity of vectors (and, accordingly, complete similarity of channels),

0 – lack of similarity (vectors are orthogonal),

–1 – complete opposite of vectors (and, accordingly, the content of channels is completely different).

Vector Dimensionality Reduction. Since the dictionary may contain tens of thousands of unique words, the resulting vectors exhibit high dimensionality. To enhance the efficiency of analysis and reduce computational costs, dimensionality reduction methods are applied by reducing the size of the initial dictionary.

Based on the analysis results, hypotheses can be formulated regarding the relationships and influence of different Telegram channels. For example, if two channels exhibit high similarity over a specific period, this may indicate interaction or the use of common information sources. Additionally, it is possible to examine how changes in authorship or topics influence text similarity. A sudden shift in a channel's vocabulary at a certain point in time may signal changes in editorial policy or audience composition.

This study focuses on the application of NLP methods for comparing information sources based on textual data. The primary emphasis is on vectorization and the use of cosine similarity to assess textual similarity. The proposed methods enable the efficient processing of large volumes of textual data, yielding valuable insights for further analysis.

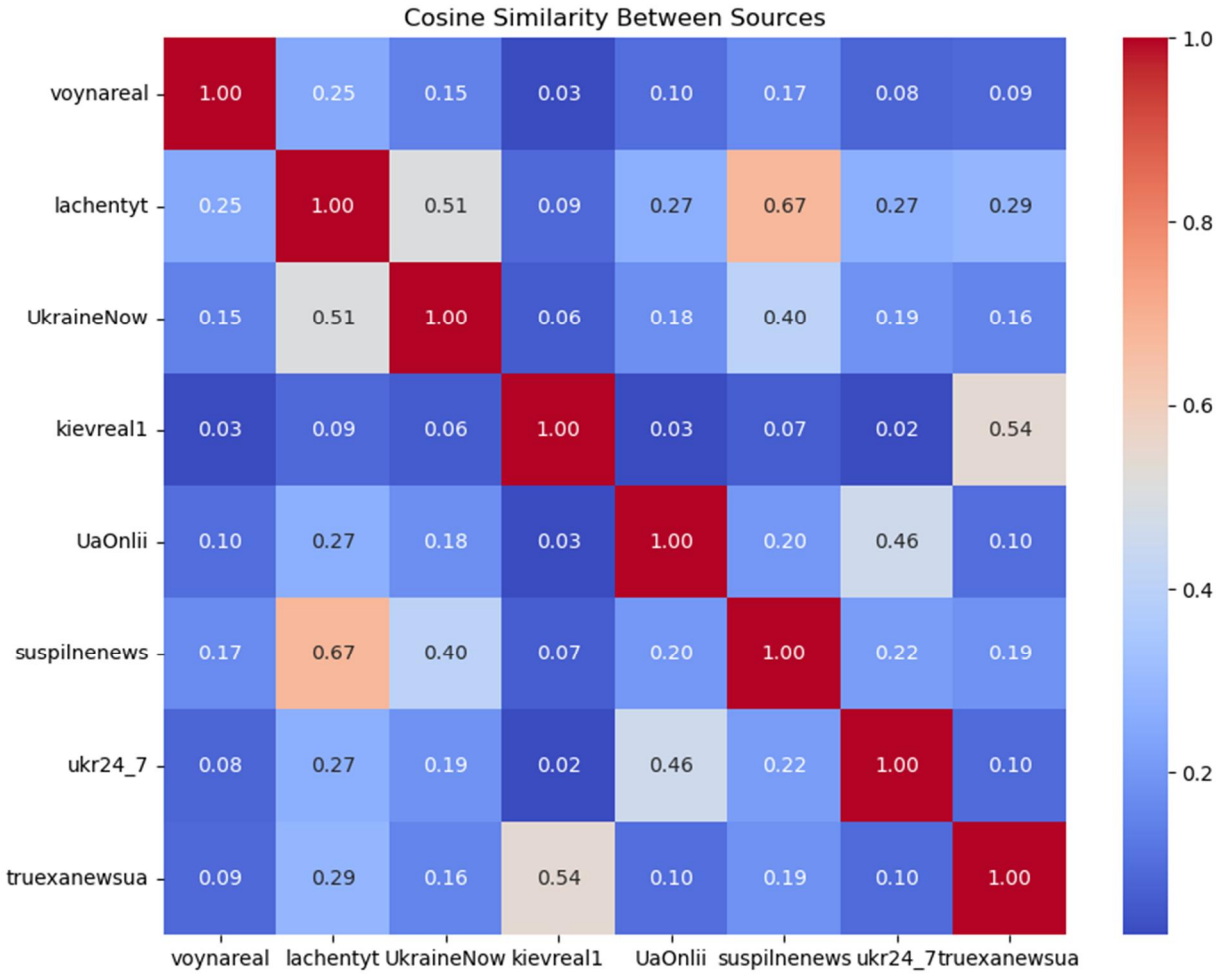

Fig. 2. Heat map. Similarity of sources based on their lexical composition

In analyzing textual data from Telegram channels, choosing the right similarity measure significantly affects clustering outcomes. Two common methods – cosine similarity and Jaccard similarity – were compared.

Cosine similarity assesses the angle between vectors representing word frequency and usage patterns, effectively capturing lexical alignment. Jaccard similarity, however, only considers the overlap of unique words, neglecting word frequency entirely, thus providing a more superficial comparison.

Empirical results demonstrate clear differences. Cosine similarity had a lower mean similarity (0.210), higher standard deviation (0.165), and higher average absolute deviation (0.127). This indicates greater sensitivity and more precise distinctions between channels. Conversely, Jaccard similarity showed higher mean similarity (0.283) and lower variability (standard deviation of 0.097 and average absolute deviation of 0.082), reducing its effectiveness in differentiating channels.

Therefore, cosine similarity is better suited for clustering textual data from Telegram channels, as it offers more precise and meaningful clusters. Future research could further validate these findings using larger datasets or different textual contexts.

In the experiment, textual data from eight Telegram channels were analyzed, each containing a different number of messages (see Table 1). These Telegram channels were selected as they represent popular Ukrainian-language sources primarily focused on political and social news. Their selection ensures sufficient data volume, thematic relevance, and coverage of diverse perspectives essential for analyzing lexical similarities and clustering effectiveness. The text vectorization algorithm described above was applied for the analysis. Specifically, a minimum token occurrence threshold was set at 220 occurrences. As a result, a K_i vector of dimension 81 was generated for each channel, where each component of the vector represented the frequency of a specific word from the global dictionary within the channel's text.

Table 1

Information sources and their data volume

Source	Number of messages
UaOnlii	2014
UkraineNow	1670
kievreal1	1565
lachentyt	934
suspilnenews	1088
truexanewsua	1474
ukr24_7	1063
voynareal	2307

The next step in the analysis was to apply the cosine similarity algorithm to each pair of the resulting vectors. This allowed us to create a correlation matrix, where each element reflected the degree of similarity between the two information sources (see Fig. 2). The matrix clearly demonstrates the relationships between the channels based on their lexical composition.

Tuning of a parameter for creating a cosine similarity dictionary

In any text data analysis, a key task is to optimize the parameters (Daelemans, 2003) that influence the efficiency of the model. In our case, when analyzing cosine similarity, the critical parameter was the minimum frequency threshold for token occurrences considered in the calculations. This threshold determines which words in the text should be regarded as significant and which should be ignored due to their low frequency.

Eliminating rare terms helps minimize the negative impact of noise on the model. Including words that appear only once or twice would significantly increase the dimensionality of the vector space, leading to higher computational complexity in subsequent operations. This, in turn, would hinder the clarity and reliability of the results. Moreover, such rare tokens are unlikely to contribute meaningful information for thematic analysis or classification tasks.

When selecting a minimum word frequency threshold, it is crucial to find a balance. Setting the threshold too high would leave too few terms in the dictionary, reducing the lexical space to only the most frequent words, which may not be sufficiently representative of the studied channels. Additionally, some of these common words may be absent from certain sources, complicating cross-source analysis and comparison. This could result in the loss of important features that might significantly enhance the model and improve the accuracy of conclusions.

Conversely, setting the threshold too low would expand the lexical set with rare or secondary words, increasing computational complexity and introducing excessive noise into the data. As a result, the model could become overwhelmed by an abundance of unnecessary terms, wasting resources on their processing and complicating result interpretation.

Therefore, the optimal minimum frequency threshold should be high enough to prevent the dictionary from being overloaded with unnecessary, infrequent terms, yet low enough to avoid excessive reliance on only the most frequent words. Striking this balance ensures the preservation of lexical diversity while minimizing noise and maintaining conditions for effective and meaningful comparisons between different channels.

At the beginning of the analysis, the minimum frequency threshold for term occurrences in texts was initially set at 220 uses per word. However, based on the similarity matrix constructed using cosine similarity, this parameter was automatically adjusted to maximize the arithmetic mean of the matrix (excluding its main diagonal). Since the arithmetic mean of the matrix is directly proportional to channel similarity estimates, this adjustment ensured that the selected threshold provided the most informative representation of inter-channel similarity.

To refine the threshold, a grid search (Stokes, 2021) was performed over the range of 10 to 300 with a step size of 10. The analysis determined that a minimum frequency threshold of 30 was the optimal compromise. This value retained a sufficient number of terms to adequately capture the lexical specificity of the texts while avoiding the excessive inclusion of irrelevant words. This approach ensures a balance between model accuracy, computational efficiency, and the representativeness of the constructed dictionary.

Cosine similarity based on time intervals

Given that some Telegram channels may change their authors or topics over time, it is essential to analyze textual changes over specific periods. Temporal analysis enables the identification of shifts in a channel's vocabulary over time and assesses how these changes influence its similarity to other channels.

Breakdown into Time Intervals. To analyze temporal variations, all messages within a channel can be segmented into time intervals, such as months, quarters, or half-years. This approach allows the construction of a vector for each interval, facilitating comparisons between different periods.

Analysis of Similarity Dynamics. By constructing vectors for different time intervals, it is possible to examine how the similarity between channels evolves. This analysis provides insights into potential topic shifts, audience changes, or mutual influences among channels.

For this study, two-time intervals were selected: 30 and 90 days. This segmentation allows for analysis at different levels of granularity: shorter intervals capture rapid changes, while longer intervals reveal broader trends. The computational approach relies on the vector representation of texts for each channel within a given time interval. Specifically, tokens from messages within a channel for a defined period are transformed into a K_i vector, followed by the computation of similarity matrices for each studied period.

30-day intervals allow you to capture short-term changes, such as seasonal events, elections, or information campaigns. For example, during a political crisis, a change in the editorial policy of channels may be reflected in an increase or decrease in the similarity of texts with other sources, which helps to track these processes in almost real-time.

90-day intervals focus on long-term changes, such as a gradual change in thematic focus or a shift in emphasis to new topics. They help to identify general trends and strategic changes in the behavior of media platforms.

The biggest change was observed for the channels “lachentyt” and “UkraineNow” (see Table 2). Data for Table 2 were derived from the similarity matrices calculated for two consecutive 90-day periods. These matrices were generated using cosine similarity applied to text vectors of channels segmented by these intervals. Their mutual similarity level increased from 0.394 to 0.497, but at the same time, the mentioned channels increased their similarity indicators with other channels. This may indicate the coordination of thematic directions, synchronization of information campaigns, or at least the loss of exclusivity of these channels. These changes may be related to events that require consolidated coverage, such as political or economic crises.

Table 2

The largest changes in channel similarity between two 90-day intervals

Channel pair	Similarity values changing	Difference
lachentyt – UkraineNow	0.394 – 0.497	+0.103
lachentyt – voynareal	0.207 – 0.275	+0.068
UkraineNow – ukr24_7	0.139 – 0.199	+0.06
UkraineNow – voynareal	0.099 – 0.156	+0.056
UkraineNow – suspilnenews	0.325 – 0.377	+0.051

Channel Clustering

A similarity matrix was used for the analysis, which reflects the degree of similarity between channels. Based on it, the clustering algorithms Affinity Propagation (Guan, 2011) and Spectral Clustering (Janani, 2019) were applied, which allowed to identification of groups of channels with similar characteristics.

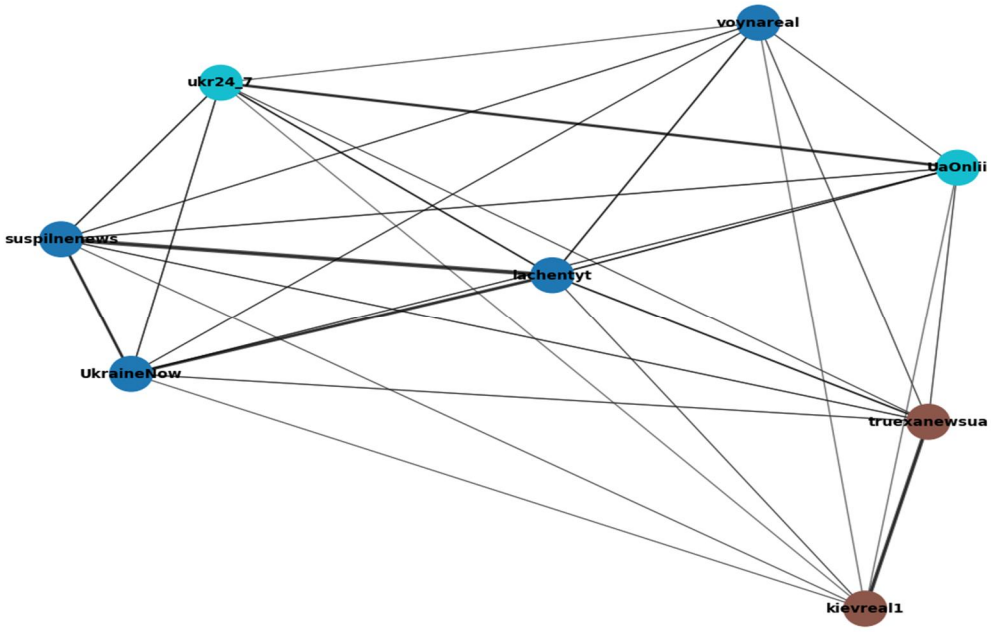

Fig. 3. Clustering using Affinity Propagation

Affinity Propagation uses the similarity matrix as input data to determine the “cluster centers” by iteratively exchanging messages between all points. This algorithm automatically determines the number of clusters by evaluating which points best represent the others. Thanks to this approach, the method efficiently processed the similarity matrix and formed three groups of channels.

Spectral Clustering also works with the similarity matrix, transforming it into a spectral space using eigenvalues and eigenvectors. In this new space, data points are grouped according to their proximity, for example, using the K-means method (Abualigah, 2016). Thanks to this technique, the algorithm is well suited for detecting clusters with nonlinear boundaries, as in the case of our matrix analysis.

Applying the Affinity Propagation method to the general channel similarity matrix for the entire period, we obtained 3 clusters as a result (see Fig. 3). Each vertex of the graph in the figure represents a separate channel, and the thickness of the edges between the channel vertices is the degree of similarity.

Now that the predicted number of clusters has been determined, other clustering algorithms can be applied, such as Spectral Clustering, where the number of clusters is explicitly specified. The results of this clustering also produced three clusters (see Fig. 4), consistent with the previous clustering results obtained using Affinity Propagation.

The clustering results provide valuable insights and serve as a foundation for further detailed analysis. In particular, grouping channels based on lexical similarity can help assess the potential affiliation of a given channel with an information influence group. For example, if a channel falls within the same cluster as well-known propaganda sources or organized influence networks, this may indicate similarities in their information policies, target audiences, or methods of influence.

For deeper analysis, comparisons can be made with reference sources previously identified as part of hostile propaganda efforts. Using clustering results, various parameters such as thematic focus, stylistic features, and the prevalence of specific vocabulary typical of propaganda channels can be evaluated. This approach enables the identification of hidden connections between channels, even in cases where no direct affiliation is apparent.

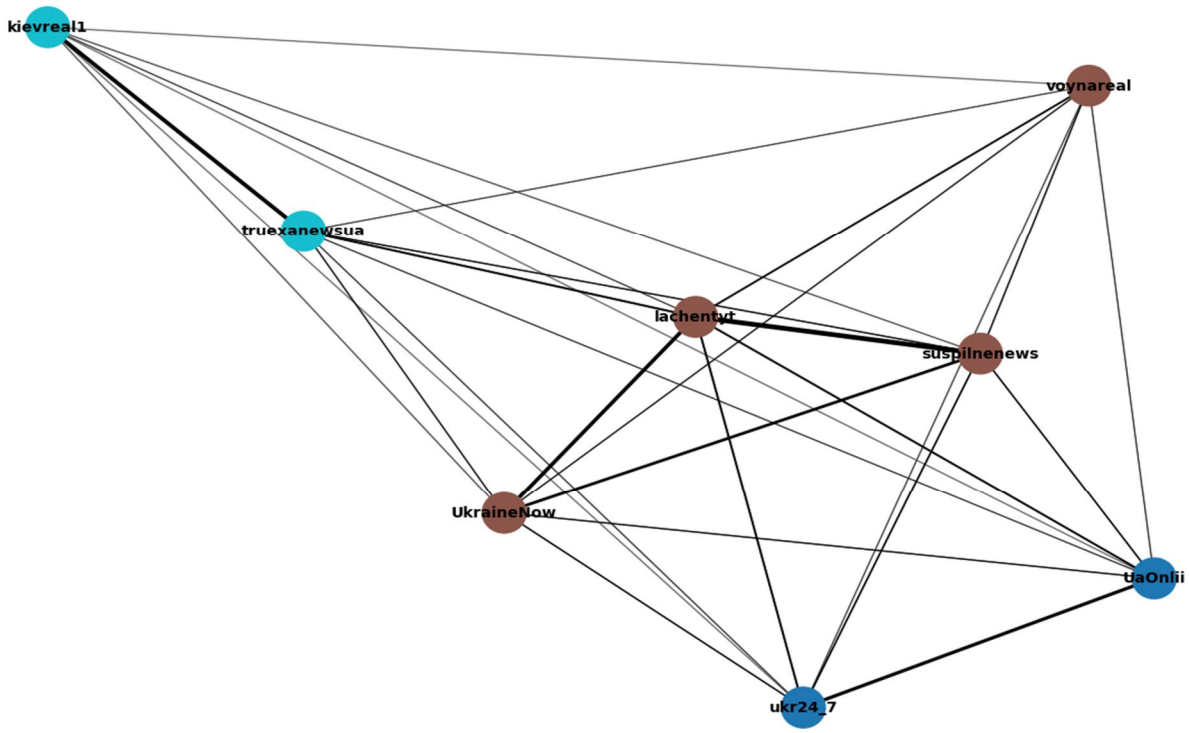

Fig. 4. Clustering using Spectral Clustering

Furthermore, analyzing the cluster assignment of a channel provides insights into its information strategy. For instance, similarity to channels with an overt propaganda orientation may indicate the deliberate adoption of similar narratives or presentation techniques. Such findings serve as a basis for monitoring potentially manipulative sources, identifying emerging trends, and adapting appropriate countermeasures.

Although the last figure 3 is similar to figure 4, in the latter the edges are grouped in space, which gives a better representation of the channel groups and their connections.

Conclusions

This study proposed and implemented an algorithm for comparing information sources (channels), enabling an efficient analysis of their lexical similarity. Optimization of model parameters ensured maximum differentiation between channels. The analysis of channel similarity within limited time intervals provided insights into the dynamics of changes in information policies, including thematic shifts, changes in editorial focus, and adaptations to external events. Additionally, channel clustering was conducted, grouping channels based on their lexical characteristics.

Ideas for future research:

1. Removal of commonly used words. Improve the approach to pre-processing texts by removing commonly used words that do not carry a significant semantic load, for example, such as “subscribe”, “channel name”, or other markers characteristic of the technical language of channels. This will reduce noise in the data and improve the quality of clustering.
2. Calculating trend graphs of channels using the cosine similarity algorithm and moving averages (Chiarella, 2006).
3. Data expansion. Increase the number of channels under study, which will allow for obtaining a more representative data set. This will open up opportunities for analyzing broader relationships between sources, as well as for identifying new clusters and trends.
4. Cluster analysis. Based on the created clusters, conduct additional analysis to find characteristics that are defining for each group. The identification of such features can be used to create a binary classification model (Dogra, 2022), which will automatically determine which cluster a new channel belongs to.

REFERENCES

1. Abualigah, L. M., Khader, A. T., & Al-Betar, M. A. (2016). Multi-objectives-based text clustering technique using K-mean algorithm. *2016 7th International Conference on Computer Science and Information Technology (CSIT),* 1-6. https://doi.org/10.1109/csit.2016.7549464
2. Camacho-Collados, J. (2018). On the Role of Text Preprocessing in Neural Network Architectures: An Evaluation Study on Text Categorization and Sentiment Analysis. *eprint arXiv, 1707(01780)*, 1-4. https://doi.org/10.48550/arXiv.1707.01780
3. Chai, C. (2023). Comparison of text preprocessing methods. *Natural Language Engineering, 29*(3), 509-553. https://doi.org/10.1017/S1351324922000213
4. Chiarella, C., He, X.-Z., & Hommes, C. (2006). *A dynamic analysis of moving average rules. Journal of Economic Dynamics and Control, 30*(9), 1729–1753. https://doi.org/10.1016/j.jedc.2005.08.014
5. Daelemans, W., Hoste, V., De Meulder, F., Naudts, B. (2003). Combined Optimization of Feature Selection and Algorithm Parameters in Machine Learning of Language. In: Lavrač, N., Gamberger, D., Blockeel, H., Todorovski, L. (Eds.) *Machine Learning: ECML 2003. Lecture Notes in Computer Science*, 2837 https://doi.org/10.1007/978-3-540-39857-8_1
6. Dogra, V., Verma, S., Kavita, Chatterjee, P., Shafi, J., Choi, J., & Ijaz, M. F. (2022). A Complete Process of Text Classification System Using State-of-the-Art NLP Models. *Computational Intelligence and Neuroscience, 2022*, 1–26. https://doi.org/10.1155/2022/1883698
7. Guan, R., Shi, X., Marchese, M., Yang. C., & Liang, Y. (2011). Text Clustering with Seeds Affinity Propagation. *IEEE Transactions on Knowledge and Data Engineering*, *23*(4), 627-637. https://doi.org/10.1109/TKDE.2010.144
8. Janani, R., & Vijayarani, Dr. S. (2019). Text document clustering using Spectral Clustering algorithm with Particle Swarm Optimization. *Expert Systems with Applications. 134,* 192-200. https://doi.org/10.1016/j.eswa.2019.05.030
9. Magara B. M., Ojo S. O. & Zuva T. (2018). A comparative analysis of text similarity measures and algorithms in research paper recommender systems. *Conference on Information Communications Technology and Society (ICTAS)*, 1-5. https://doi.org/10.1109/ICTAS.2018.8368766
10. Mohammad, F. (2018). Is preprocessing of text really worth your time for online comment classification? *eprint arXiv, 1806(029908)*, 1-5. https://doi.org/10.48550/arXiv.1806.02908
11. Park, K., Hong, J. S., & Kim, W. (2020). A Methodology Combining Cosine Similarity with Classifier for Text Classification. *Applied Artificial Intelligence, 34*(5), 396–411. https://doi.org/10.1080/08839514.2020.1723868
12. Stokes, E. (2021, December 11). NLP with Pipeline & GridSearch - Towards Data Science. *Medium.* https://towardsdatascience.com/nlp-with-pipeline-gridsearch-5922266e82f4
13. Talakh, M.V. (2019). PART 7. USING TEXT MINING FOR THE ANALYSIS OF SOCIAL NETWORKS. In Ushenko, Y., Ostapov, S. & Golub, S., (Eds.), *INFORMATION TECHNOLOGIES Part 1. Application in computer vision, recognition and intelligent monitoring systems Yuriy Ushenko, Serhiy Ostapov, Serhiy Golub* (pp. 157-173). LAP LAMBERT Academic Publishing.
14. Talakh, M.V., Holub, S. & Lazarenko Y. (n.d.). Intelligent monitoring of software test automation of Web sites. *International Scientific and Practical Conference* “*Intellectual Systems and Information Technologies*”, 46-51.
15. Telegram (2025). *Telegram APIs.* Retrieved April 8, 2025, from https://core.telegram.org/api

ПОРІВНЯННЯ ТА КЛАСТЕРИЗАЦІЯ ДЖЕРЕЛ ТЕКСТОВОЇ ІНФОРМАЦІЇ НА ОСНОВІ АЛГОРИТМУ КОСИНУСНОЇ ПОДІБНОСТІ

Чженбін Ху[1], Дмитро Угрин[2], Артем Каланча[3]

[1]Хубейський технологічний університет,
Школа комп'ютерних наук, Ухань, Китай
[2-3]Чернівецький національний університет імені Юрія Федьковича,
кафедра програмного забезпечення комп'ютерних систем, Чернівці, Україна
[1]E-mail: drzbhu@gmail.com, ORCID: 0000-0002-6140-3351
[2]E-mail: d.ugryn@chnu.edu.ua, ORCID: 0000-0003-4858-4511
[3]E-mail: kalancha.artem@chnu.edu.ua, ORCID: 0009-0004-1451-7470

У цій статті представлено дослідження, спрямоване на розроблення оптимальної концепції аналізу та порівняння джерел інформації на основі великих обсягів текстової інформації з використанням методів опрацювання природної мови. Об'єктом дослідження стали канали новин Telegram, які використовуються як джерела текстових даних. Була проведене попереднє опрацювання текстів, включаючи очищення, токенізацію та лематизацію, щоб сформувати глобальний словник, що складається з унікальних слів з усіх джерел інформації. Для кожного джерела було побудовано векторне представлення текстів, розмірність якого відповідає кількості унікальних слів у глобальному словнику. Частота використання кожного слова в текстах каналу відображалася у відповідних позиціях вектора. Застосовуючи алгоритм косинусної подібності до пар векторів, була отримана квадратна матриця, яка демонструє ступінь подібності між різними джерелами. Проведено аналіз схожості каналів на обмежених часових інтервалах, що дозволило виявити тенденції зміни їх інформаційної політики. Параметри моделі були оптимізовані для забезпечення максимальної диференціації каналів, що підвищило ефективність аналізу. Застосовувалися алгоритми кластеризації, які розподіляли канали на групи за ступенем лексичної схожості. Результати дослідження демонструють ефективність запропонованого підходу для кількісної оцінки подібності та кластеризації текстових даних з різних джерел. Запропонована методика може бути використана для аналізу джерел інформації, виявлення взаємозв'язків між джерелами, дослідження динаміки змін їх діяльності та оцінки соціокультурного впливу медіаконтенту.

Ключові слова: джерело інформації, текст, подібність, обробка природної мови, попередня обробка тексту, Telegram, векторизація, косинусна подібність, кластеризація.