UDC 004.75 (004.62) https://doi.org/10.20535/2786-8729.6.2025.339127

MATHEMATICAL MODEL OF CLUSTERING OF INFORMATIONAL MESSAGES WITH INDICATORS OF ACTIVITY FOR THE INFORMATION CONTENT BY TONE AND AREAS OF SOCIETY ACTIVITY

Oleksii Pysarchuk
National Technical University of Ukraine
"Igor Sikorsky Kyiv Polytechnic Institute", Kyiv, Ukraine
https://orcid.org/0000-0001-5271-0248

Danylo Baran *
National Technical University of Ukraine
"Igor Sikorsky Kyiv Polytechnic Institute", Kyiv, Ukraine
https://orcid.org/0009-0007-0361-6870

*Corresponding author: danil.baran15@gmail.com

The mathematical model of clustering of information messages has been further developed, which is based on the frequency analysis of their tonality using Natural Language Processing methodologies with the support of large language models; OLAP visualization of clustering results and is distinguished by an established system of indicators of information content activity by areas of society activity with hierarchical compression of incoming Big Data arrays, which determines the database model for their storage. This provides an improvement to the analysis of information messages in global information networks by taking into account many factors in the areas of society activity.

The main idea and goal of the mathematical model for clustering information messages is to implement a sequence of preparation stages for detecting critical activity of the information content in global media. In practice, this is the establishment of a list and the determination of indicator values that measure content activity in primary messages, followed by their transformation into a time series – a systematized dataset. In the conditions of high density of the flow of occurrence, dynamics of development, and transformation of information content, a Big Data structure of information messages is taken into account. Therefore, the clustering model, apart from division by informational features, should provide the hierarchical compression of incoming Big Data arrays.

Research objective: development of a mathematical model of clustering information messages with indicators of information content activity by tone and spheres of activity of society. Research subject: methods of clustering information messages. Research object: process of clustering information messages.

Keywords: Big Data, clustering, Natural Language Processing.

1. Introduction

At present, the global information space is highly popular which is formed by the media, social networks, thematic channels and used as a technological platform by global information network. Analysis of the content of information flows in the global information space allows to assess the moods and information preferences of society. This is important for organizing the effective work of government and business structures for the composition of the services offers, goods, public events, etc.

Analysis of the content of global information networks requires the composition of activity indicators that would adequately reflect the essence and number of information messages in various spheres of society.

The above is relevant for increasing the efficiency, reliability, and completeness of identifying critical information messages.

2. Literature review and problem statement

The classical clustering problem involves the formation of clusters formed by sets of objects with common or similar characteristics / properties. This is implemented by well-known machine learning methods, for example: *k*-means; Support Vector Machine; *k*-nearest neighbors; hierarchical clustering [1, 2]. Clustering approaches using deep learning methods with artificial neural networks are also known [3–5].

The specificity of the task considered in the article is the processing of natural language and the need to measure information content. Natural language clustering is traditionally done by tonality, content, and other methods. Support Vector Machine, modern approaches are built on neural networks, and especially on Large Language Models (LLM) [4, 5].

A wide range of works are devoted to the problems of measuring the activity of information content, in particular [6–8]. Indicators of activity can be the frequency of repetition by keywords, primary information arrays, etc.

Combinatorial analysis of existing approaches allows to assert that the unitary use of each of them does not provide high reliability and completeness of the reflection for the dynamics of information content that affects the individual and society. Therefore, it is advisable to use them combinatorily with certain innovations that would reflect: the possibility of forming objective indicators; taking into account the subjectively directed tone of information messages; reflection of the versatility for the forms, methods and spheres of information messages, the transformation of their relevance over time; having an ability to calculate and control the large arrays.

In connection with the above, the article is devoted to solving the current problem of developing a mathematical model of clustering information messages with indicators of information content activity by tone and spheres of activity of society.

3. The aim and objectives of the research

The purpose of the research is to develop a mathematical model for clustering information messages with indicators of information content activity by tone and areas of society.

To achieve this goal, the following tasks were built and solved:

– development of a mathematical model for clustering information messages;

– verification and evaluation of the effectiveness of the application of the mathematical model of clustering of information messages.

4. Materials and methods for developing a mathematical model of information message clustering

4.1. The object and hypothesis of the study

The main idea behind clustering is due to the need to ensure a comprehensive and adequate display of information content. This is implemented by establishing indicators that reflect information activity for various spheres of society. As a measure of activity, it is proposed to use the frequency of messages in various information sources. The measured frequency forms a set of clusters, depending on the topic and emotional coloring of information messages. The specified process involves: a mathematical description of monitoring objects – information sources; the formation of indicators of information content activity; a formalized description of the infological model of sources, factors and indicators of content activity in global information networks; the development of a mathematical model of information message clustering and its verification.

The list of Key Performance Indicators (KPI) should reflect the content of information messages in different spheres of society, be objectively calculated (defined) with a reference to the information messages disseminated by means of global information networks – through sources of primary messages. That is, to establish the relationship of objective measures (in numerical equivalent) of real messages in the format of natural language, reflecting their content from the set of

spheres of values of society. The above can be described by the infology of the subject area in three stages:

1. Mathematical description of monitoring objects – information sources;
2. Establishment of a system of indicators of the activity of information content;
3. Formalized description of the infological model of sources, factors and indicators of content activity in global information networks.

4.2. Mathematical description of monitoring objects – information sources

The primary data will be considered *information messages* (*IP*) *distributed* by Internet publications (in global information networks) of various orientations, formats, forms of ownership and reliability of messages. The main requirement is the focus on the news content of information messages. These categories of publications will be called *information sources* (*IS*). Obtaining primary information messages – *primary* ("*raw*") *data* is carried out in the process of monitoring information sources. IS monitoring provides for periodic (several times, or once a day – for a specified time) review of the content of information messages on IS *objects of monitoring* (*OM*) and saving the results obtained.

The list of information sources should be representative in number, composition and provide high indicators of reliability (probability of detection and accuracy of forecast) and completeness (set of factors taken into account) of determining the dominant information content in global information networks.

Taking into account the above, a hierarchical structure of monitoring objects has been defined – global media: news sites, information channels of messengers, social networks, etc., with a distribution according to the level of trust / reliability of information / popularity.

The set of IS is divided into five groups, according to the level of verification and authorization of IP in accordance with authority in the media space, belonging to the form of ownership and representation in the global information space. Each group has a different set of OM. The quantitative composition of OM is a variable parameter. The set of IS that are put on observation form a set of monitoring objects:

$$OM = \left[OM_{1i_1}, OM_{2i_2}, OM_{3i_3}, OM_{4i_4}, OM_{5i_5}\right], \quad (1)$$

where, $i_1 = 1 \dots N_{OM_1}$, $i_2 = 1 \dots N_{OM_2}$, $i_3 = 1 \dots N_{OM_3}$ $i_4 = 1 \dots N_{OM_4}$, $i_5 = 1 \dots N_{OM_5}$.

For each group of OM, it is possible to set a scale of weight coefficients proportional to the degree of trust in the IP on the IS:

$$G = \begin{cases} OM_{1i_1} = 9 \dots 10, \\ OM_{2i_2} = 7 \dots 8, \\ OM_{3i_3} = 5 \dots 6, \\ OM_{4i_4} = 3 \dots 4, \\ OM_{5i_5} = 1 \dots 2, \end{cases} \quad (2)$$

where, $i_1 = 1 \dots N_{OM_1}$, , $i_2 = 1 \dots N_{OM_2}$, $i_3 = 1 \dots N_{OM_3}$, $i_4 = 1 \dots N_{OM_4}$, $i_5 = 1 \dots N_{OM_5}$.

Weight coefficients are determined by expert means and in the simplest case take unit values.

The result of monitoring is primary data arrays – information messages (in natural language linguistics) with reference to: OM, date, time. Primary data sets – IP will form Big Data structures over time. The monitoring results will be stored in a database.

4.3. Indicators of information content activity.

The versatility of forms and methods of IP aimed at the consciousness and subconscious of the individual / social groups / community, the massiveness and continuity of IP into the spheres of values of society require a mirror description in the set of IP indicators. We will assume that critical activities from the standpoint of IP focus on seven basic areas of values: economic, political, security, social, law, spiritual, individual rights and freedoms (Fig. 1).

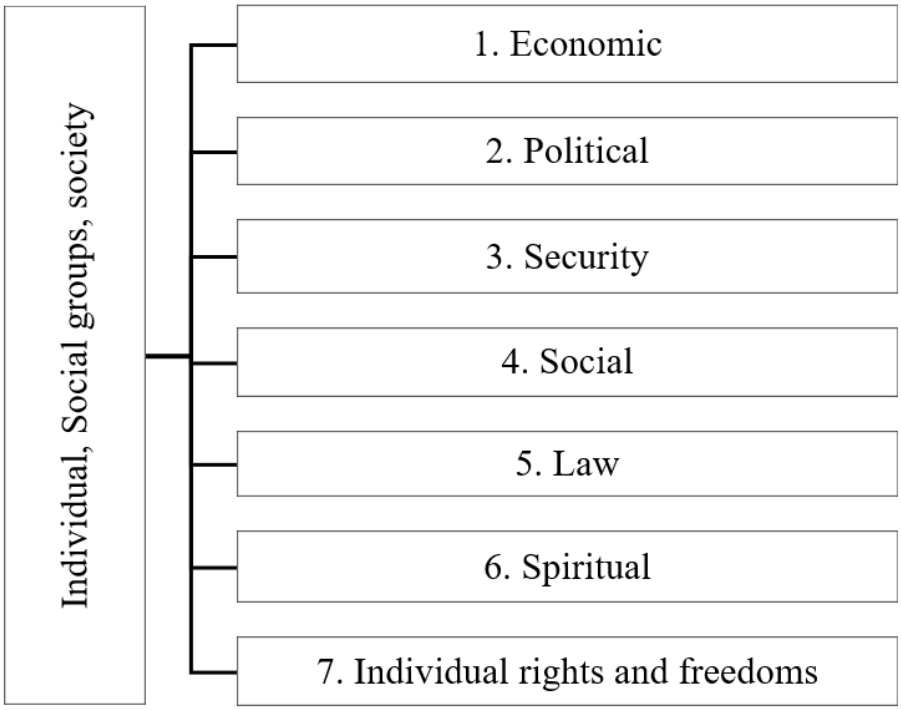

Fig. 1. Spheres of values and activities of society.

The result of observation of an individual OM from the (2) is the primary information message for a specific day / date and time of day:

$$IP_{ij} = \{text_{ij}\}, \tag{3}$$

where, $i = 1 \dots N = \begin{bmatrix} i_1 = 1 \dots N_{OM_1}, \\ i_2 = 1 \dots N_{OM_2}, \\ i_3 = 1 \dots N_{OM_3}, \\ i_4 = 1 \dots N_{OM}, \\ i_5 = 1 \dots N_{OM_5} \end{bmatrix}, j = 1 \dots M = \begin{bmatrix} day_1[t_{\text{morning}}, t_{\text{morning}}, t_{\text{evening}}], \\ day_2[t_{\text{morning}}, t_{\text{morning}}, t_{\text{evening}}], \\ day_3[t_{\text{morning}}, t_{\text{morning}}, t_{\text{evening}}], \\ \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \\ day_M[t_{\text{morning}}, t_{\text{morning}}, t_{\text{evening}}] \end{bmatrix}.$

In the designations (3), the index i – denotes the object of monitoring, j – the time (day – day, time of day) of receiving an information message from the i -th source.

Text messages (3) are expanded into the results database table. Thus, over time, a Big Data array of primary information messages in natural language is formed. In the future, the process of forming indicators of information content activity is carried out through the following stages:

– primary processing of information messages;

– clustering of the first level – by spheres Figure 1;

– clustering of the second level – by the tone (negative, neutral, positive) of messages in each sphere of values and activities of society;

Setting the frequency (number) of messages of the corresponding tonality in each area of the list by spheres Figure 1.

The listed stages are implemented recurrent in time for the flow of information messages (3). This allows you to form a data structure of the Time Serie type, frequencies, tonality, flow of information messages for different spheres of values of society.

Thus, a cause-and-effect (infological) relationship / model of sources, factors and indicators of content activity in global information networks is established with their decomposition by spheres

and tonality. This allows you to establish a numerical measure of content activity at the level of quantitative and qualitative display of messages in the global information space.

Primary processing of information messages – involves a *pipeline of preparatory stages of NLP* (*Natural Language Processing*): filtering; normalization; tokenization; removal of stop words; lemmatization. The result is processed information arrays: filtered message with content preservation – $IP_{ij,filter}$; lemmatized message – $IP_{ij,lemma}$:

$$IP_{ij,filter} = f_{NLP,filter}(text_{ij}), \tag{4}$$

$$IP_{ij,lemma} = f_{NLP,lemma}(text_{ij}), \tag{5}$$

where, $f_{NLP,filter}$ – denotes text message filtering operations, $f_{ij,lemma}$ – NLP pipeline operations: filtering; normalization; tokenization; removing stop words; lematization.

Information arrays $IP_{ij,lemma}$ – are used for operational frequency and probabilistic analysis. Information messages $IP_{ij,filter}$ are subject to further in-depth processing by clustering them and forming indicators of information content activity. In the future, the results of in-depth processing have *OLAP* (*Online Analytical Processing*) format for visualization and analysis.

Clustering of the first level – *spherical* involves establishing the belonging of information messages to the spheres of values and activities of society (Fig. 1). The input information is the results of the primary filtering of the received information messages (2.4), which give spherical clusters according to the transformation model:

$$S_{area} = f_{NLP,S_{clustering}}(IP_{ij,filter}), \tag{6}$$

where, $S_{area} = [S_1, S_2, S_3, S_4, S_5, S_6, S_7]$ – reflect sets of information messages clustered by spheres of values and activities of society (Fig. 1) information messages (4); $area = 1 \dots 7$; $f_{NLP,S_{clustering}}$ – the operation of clustering incoming messages in natural language by spheres of Figure 1 with NLP process technologies.

Clustering of the second level – *spherical tonality* involves establishing the tone of information messages in each sphere of values (Fig. 1) by categories: negative, neutral, positive. The input information for clustering of the second level is sets S_{area}. The results of the second level of clustering can be represented by the transformation:

$$S_{area,tonality} = f_{NLP,S_{tonality}}(S_{area}), \tag{7}$$

where, $S_{area,tonality} = \begin{bmatrix} S_{area,tonality} = \text{negative}, or \\ S_{area,tonality} = \text{neutral}, \ or \\ S_{area,tonality} = \text{positive}. \end{bmatrix}$ – sets of information messages clustered by tone S_{area}; $area = 1 \dots 7$: $f_{NLP,S_{tonality}}$ – the operation of clustering incoming text messages in natural language by tonality with NLP process technologies (the specifics of the implementation are also revealed below).

Secondary processing of clustered information messages – *frequency analysis of clusters of spherical tonality is* taken as input information by the formed clusters (6). The task of the stage is to establish the frequency / number of messages of each of the three keys from the sphere of values of Figure 1 for each moment of time of discrete moments of the time j of OM monitoring. The result of the stage is the frequencies:

$$F_{j,area,tonality} = \begin{bmatrix} \hat{F}_{,area,tonality} = \text{negative} \\ \hat{F}_{,area,tonality} = \text{neutral} \\ \hat{F}_{,area,tonality} = \text{positive} \end{bmatrix}, \tag{8}$$

where, $\hat{F}$ – denotes the average daily and OM frequency of occurrence of negative / neutral / positive information messages from the sphere $area$ – list of Figure 1; , $j = 1 \ldots M$.

It is worth noting that the determination of the average value of the frequency (8) occurs according to the trivial expression:

$$\hat{F}_{j,area,tonality} = \sum_{l=1}^{l=3} \begin{bmatrix} F_{j,area,tonality} = \text{negative} \\ F_{j,area,tonality} = \text{neutral} \\ F_{j,area,tonality} = \text{positive} \end{bmatrix}, \quad (9)$$

where, $l = 3 = t_{\text{morning}}, t_{\text{morning}}, t_{\text{evening}}$.

However, the average OM frequency is defined as a weighted average value – proportional to the level of confidence in the monitored object, for example, on a scale of weighting factors of G_l type (2) in accordance with the expression:

$$\hat{F}_{j,area,tonality} = \sum_{l=1}^{l=N_{OM}} G_l \begin{bmatrix} \hat{F}_{area,tonality} = \text{negative} \\ \hat{F}_{area,tonality} = \text{neutral} \\ \hat{F}_{area,tonality} = \text{positive} \end{bmatrix}, \quad (10)$$

where, $l = N_{OM} = N_{OM_1} + N_{OM_2} + N_{OM_3} + N_{OM_4} + N_{OM_5}$.

Thus, sets (8) reflect the causal (infological) relationship / model of sources, factors and indicators of content activity in global information networks and is a data structure of the Time Series type – frequencies of the tonality of the flow of information messages for different spheres of values of society.

4.4. Formalized description of the infological model of sources, factors and indicators of content activity in global information networks

The infological model is aimed at ensuring the process of identifying critical activity of information content in global media, which can change moods and lead to dangerous actions of society. Infology allows you to form a multifactorial system of indicators – measurements – digitization of social processes that occur under the influence of target information. *The infological model is formalized in accordance with the introduced mathematical descriptions and models* (1)–(10) *and is presented in the format of the scheme Figure 2.*

The cause-and-effect relationships of the infological model have the following linguistic description.

Information content, its discussions, which are present in global information networks as news sites, information channels of messengers, social networks, are subject to constant round-the-clock monitoring. This is implemented through observation of a representative list of monitoring objects (1) with established measures of trust in them (2). Monitoring objects form the first *stratum* (*layer*) *of monitoring* objects, which has a direct reflection of the real global information space. This imitates the perception of information content by the object of information influence – a person (individual, social group, society). The execution of the objects of monitoring form the first level of the scheme Figure 2.

The result of OM monitoring is a Big Data array of information messages (3) – primary data presented in natural language. IP arrays are stored in the appropriate database. Information messages obtained as a result of OM observation form *the second stratum* (*layer*) *of information messages* (the second level of the scheme Figure 2).

The strata of information processing (third) is formed by the following processes: primary processing of IP, models (4), (5); clustering of the first level – spherical (6); clustering of the second level – spherical tonality (7).

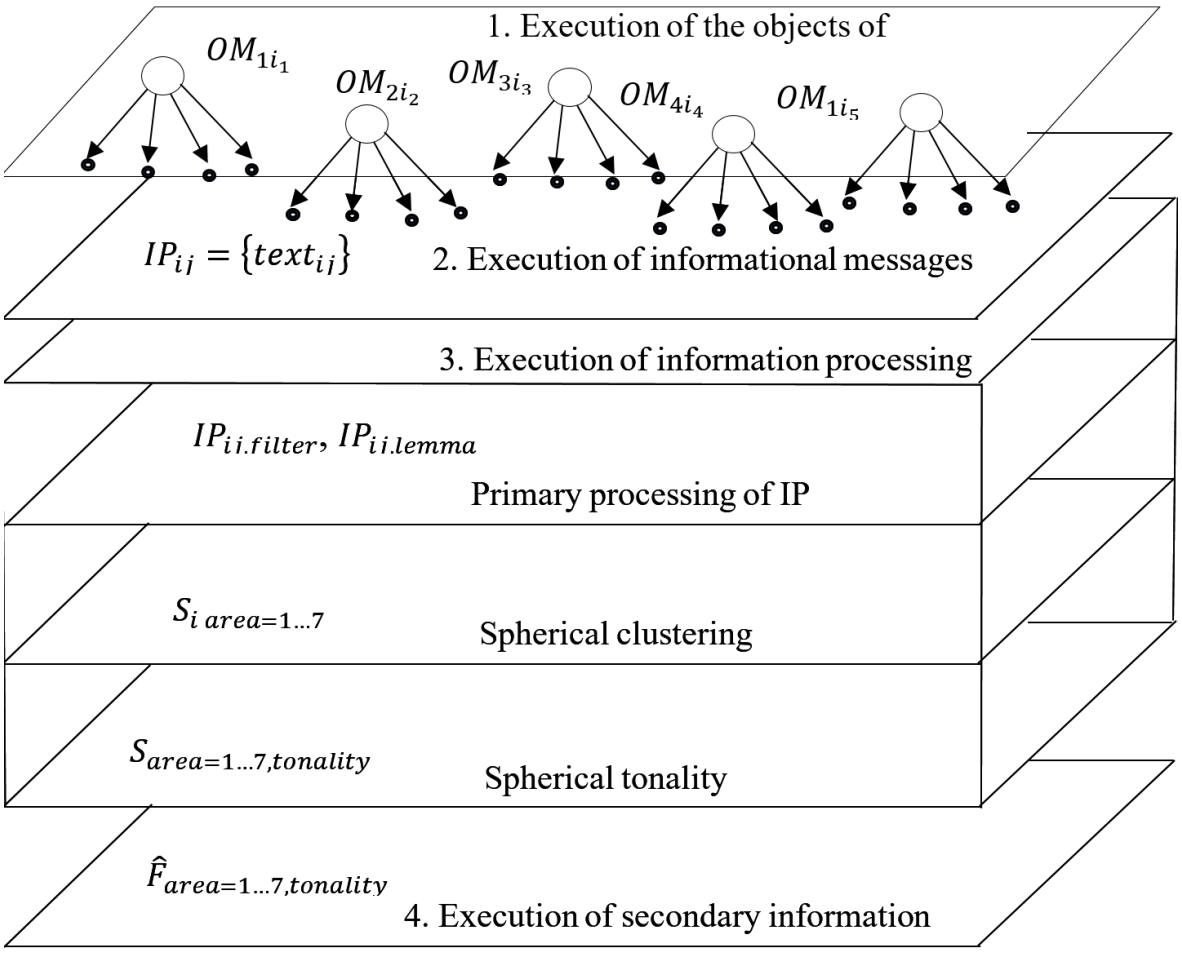

Fig. 2. Strata (layers) of the infological model of sources, factors and indicators of content activity in global information networks

Hierarchical compression of incoming Big Data arrays of IP forms the fourth *stratum of secondary information processing*. The result is a data structure of the Time Serie type – the frequency of the tonality of the flow of information messages for different areas of value (Fig. 1).

4.5. Mathematical model of clustering of information messages

Mathematical model of clustering of information messages – formed from partial models of each of the stages:

1. Construction of an infological model of sources, factors and indicators of content activity in global information networks: monitoring of the global information space; establishing the belonging of news to the spheres of values of society; establishing the tone of information messages by spheres of values; counting the number of information messages by tone in each of the spheres of values and the formation of time series;

2. Primary processing of information messages – *pipeline of NLP preparatory stages*: filtering; normalization; tokenization; removal of stop words; lemmatization;

3. Clustering of the first level – *spherical* involves establishing the belonging of information messages to the spheres of values and activities of society (Fig. 1);

4. Clustering of the second level – *spherical tonality* involves establishing the tone of information messages in each sphere of values and activities of society (Fig. 1) by categories: negative, neutral, positive;

5. Secondary processing of clustered information messages – *frequency analysis of clusters of spherical tonality*. The task of the stage is to establish the frequency / number of messages of each of

the three tonalities from the sphere of values Figure 1 for each moment of time j – discrete moments of time as a frequency $\hat{F}$ – average per day and by OM frequency of appearance of negative / neutral / positive information messages from the sphere of society.

5. Results of verification and evaluation of the effectiveness of the application of the mathematical model of clustering of information messages

To determine the functional suitability for practical application, verify the structure and content and evaluate the effectiveness of the proposed mathematical model for clustering information messages, a program module has been developed. The module is implemented in python with the following libraries: spacy, nltk, re, matplotlib, seaborn, Textblob, vaderSentiment.vaderSentiment, nltk.sentiment.vader, GoogleTranslator.

Monitoring took place on three daily sections: morning 08:00–10:00, midday 12:00–14:00, evening 17:00–20:00. Verification and evaluation of the effectiveness of the proposed mathematical model was carried out based on the monitoring results obtained within 2 weeks. That is, according to a sample of 42 dimensions, which contains more than 1000 informational messages.

Evaluation of the effectiveness of the results of IP clustering by sentiment was implemented on the statistics of 100 messages for each area of society. Comparison of automated solutions was carried out according to the agreed assessment of experts. The experts were officials – consumers of clustering results. The calculations showed the probability of correct identification $P = 0{,}87$ and the probability of a $F = 0{,}2$.

Based on the data obtained, spatio-temporal OLAP was implemented – analysis of the data obtained. The results of visualization in the format of OLAP cubes are shown in the graphs of Figure 3.

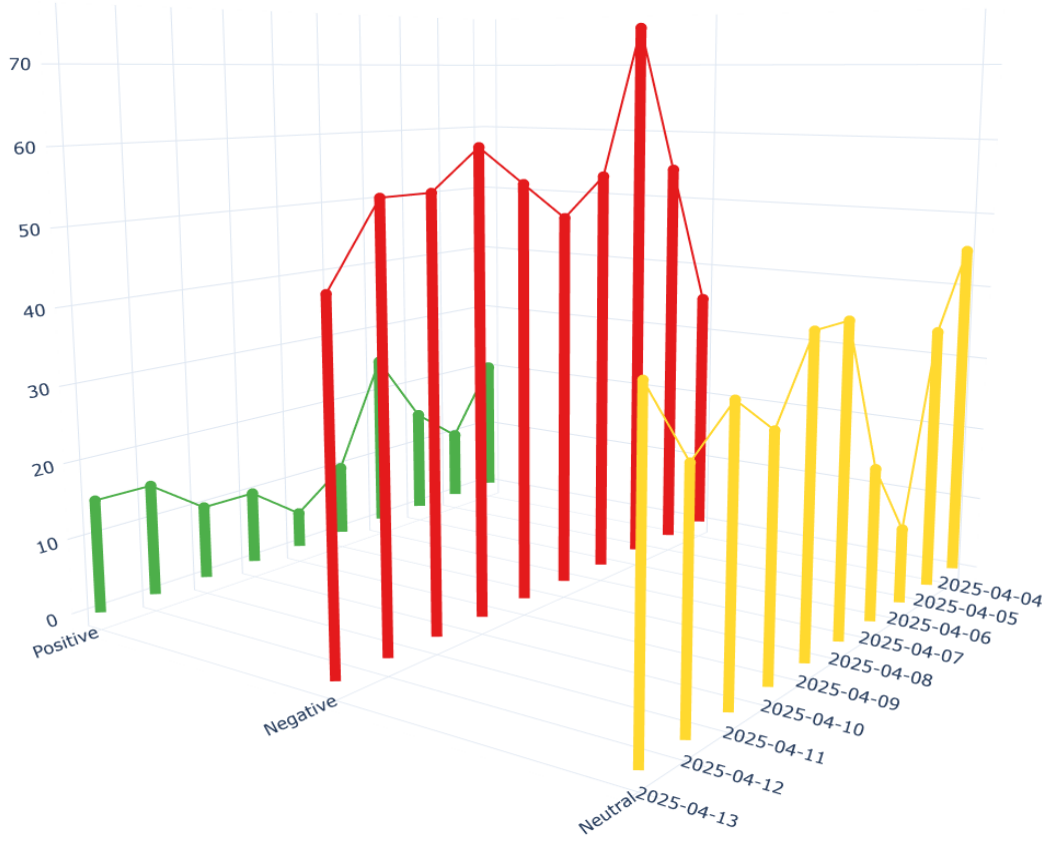

Fig. 3 (*a*). OLAP – analysis of received and generated data: Change in time of tonality in the morning

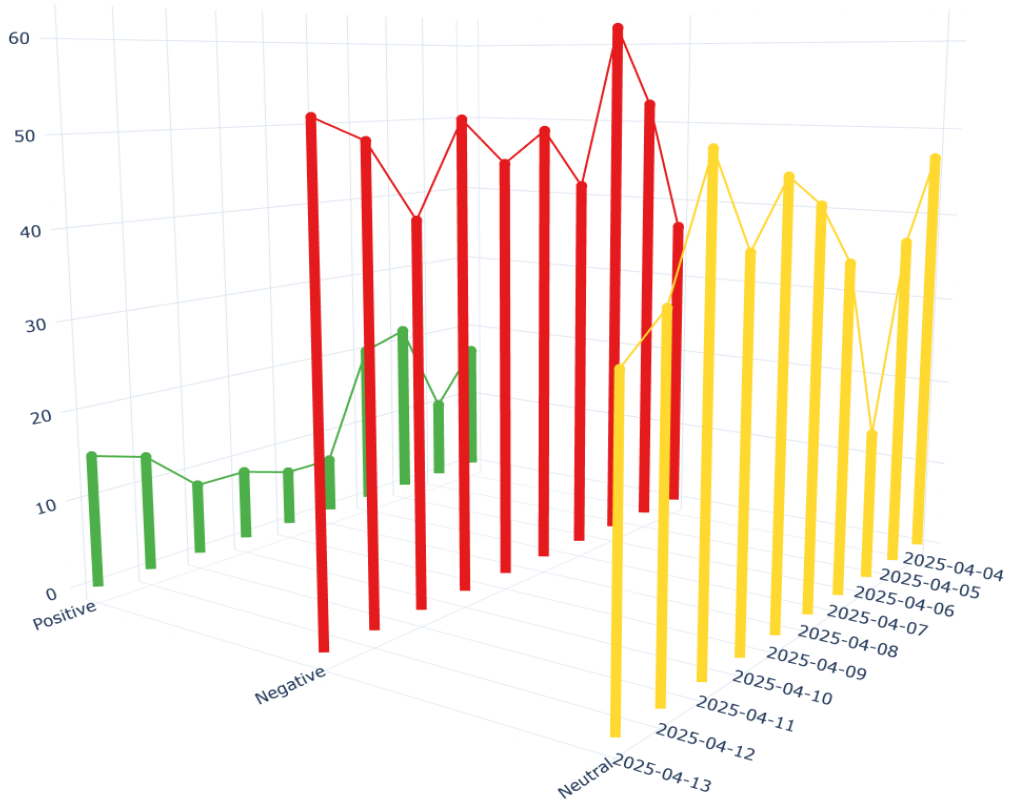

Fig. 3 (*b*). OLAP – analysis of received and generated data: Change in time of tonality for lunch

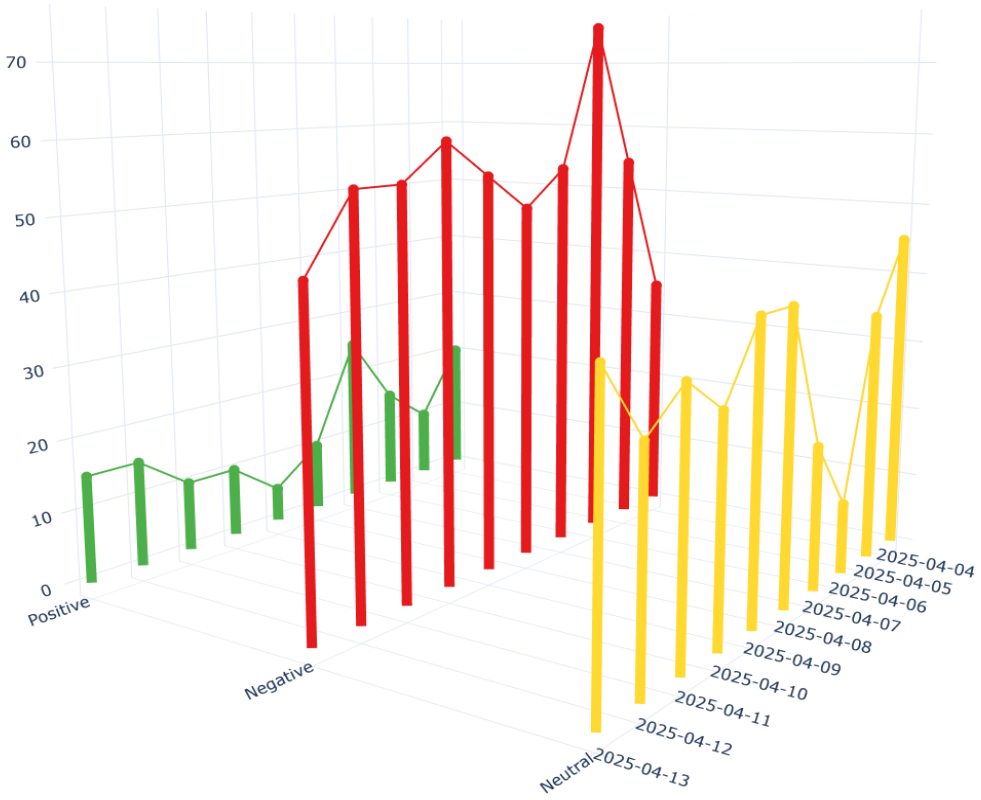

Fig. 3 (*c*). OLAP – analysis of received and generated data: Change in time of key for the evening

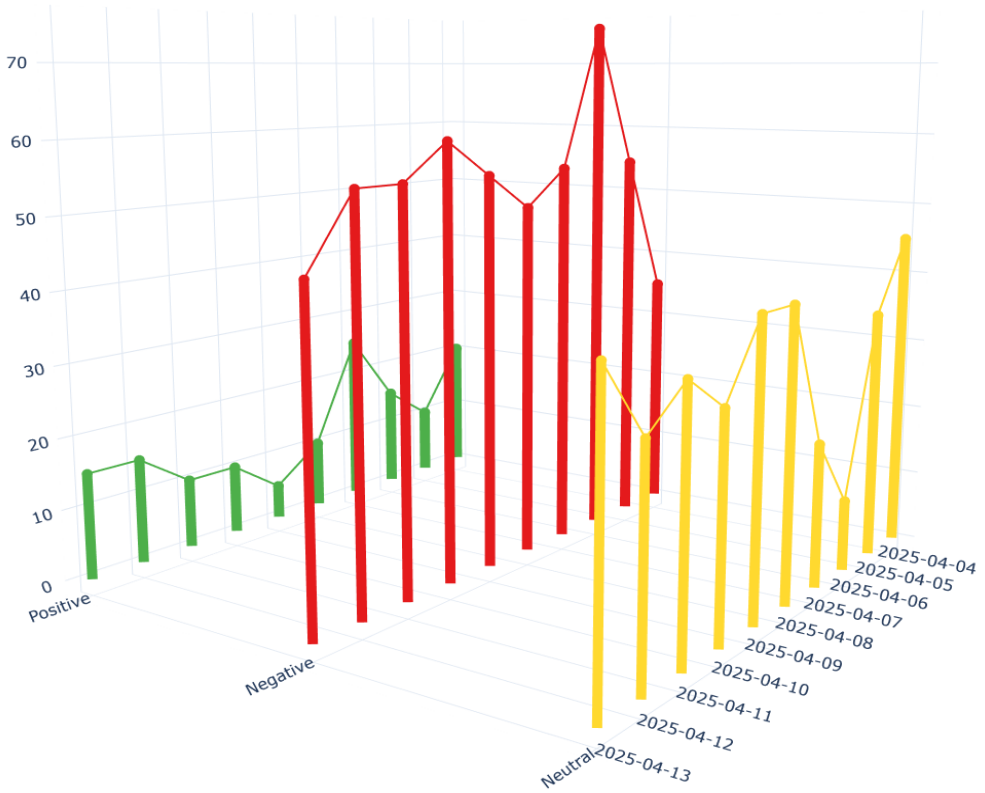

Fig. 3 (*d*). OLAP – analysis of received and generated data: Average key change per day

The peculiarity and advantage of the research results in Fig. 3 is their objective focus and analysis of real data. That is, conducting a full-scale experiment.

6. Discussion of the obtained results

The analysis of the graphs in Figure 3 allows us to form the following conclusions. At the time of monitoring in the information space, the number of negative messages prevails over the positive ones. This is observed throughout the monitoring period. The ratio of positive / negative / neutral messages during the day has a relatively stable balance. Over time, the change in content activity by tone has an oscillating monotonic trend with random noise. Peak (abnormal) number of negative messages is observed on the facts of high-profile events – for the monitoring period, in particular. Evening 04.04.25 – air attack on Kryvyi Rih (05.04.25 – morning anomalies in the number of negative information messages) [9]. Morning 06.04.25 – Kyiv, air night attack by missiles and UAVs, [10]. Moreover, the second event against the background of the first has a lower level of activity.

Thus, the formed indicators of the activity of information content as the frequency of tonality of information messages reflect real events and dynamics of information messages, and therefore are suitable for practical application

Conclusion

In the course of research, the mathematical model of clustering of information messages was further developed, which is based and is different on the frequency analysis of their tonality using NLP methodologies with the support of large language models; OLAP visualization of clustering results and is distinguished by an established system of indicators of information content activity by areas of society activity with hierarchical compression of incoming Big Data arrays. This provides an increase in the completeness of the analysis of information messages in global information networks by taking into account many factors in the areas of society activity.

The calculations showed the following indicators of the effectiveness of clustering by the tone of information messages: probability of correct identification / efficiency $P = 0{,}87$; probability of missing / loss $F = 0{,}2$. These performance indicators are acceptable for the practical use of clustering results.

The results of the OLAP analysis proved that at the time of monitoring in the information space, the number of negative messages prevails over the positive ones. This is observed throughout the monitoring period. The positive / negative / neutral messages ratio during the day has a relatively stable balance. Over time, the change in content activity in tone has a fluctuating monotonic trend with random noise. The peak (abnormal) number of negative reports is observed on the facts of confirmed resonant events.

Thus, the formed indicators of information content activity as the frequency of tonality of information messages reflect real events and dynamics of information messages, and therefore are suitable for practical application. Therefore, the proposed solutions for clustering information messages are effective and suitable for practical application.

References

[1] O. Pysarchuk, A. Gizun, A. Dudnik, V. Griga, T. Domkiv, and S. Gnatyuk, "Bifurcation prediction method for the emergence and development dynamics of information conflicts in cybernetic space," *in Proc. 1st Int. Workshop on Cyber Hygiene & Conflict Management in Global Information Networks (CyberCon-2019)*, Kyiv, Ukraine, 2019, pp. 692–709. [Online]. Available: https://ceur-ws.org/Vol-2654/paper54.pdf. Accessed: Aug. 1, 2025.

[2] S. Pitafi, T. Anwar, and Z. Sharif, "A taxonomy of machine learning clustering algorithms, challenges, and future realms," *Appl. Sci.*, vol. 13, no. 6, p. 3529, 2023. https://doi.org/10.3390/app13063529

[3] O. Puchkov, D. Lande, and I. Subach, "Methods of creating, clustering and visualization of correlation networks determined by the dynamics of thematic information flows," *Inf. Technol. Secur.*, vol. 3, no. 1, pp. 6–16, 2025. https://doi.org/10.20535/2411-1031.2025.13.1.328753

[4] D. Lande and L. Strashnoy, "Advanced Semantic Networking Based on the Large Language Models: Monograph". Kyiv, Ukraine: Engineering, 2025, 258 p. ISBN: 978-617-8180-02-7.

[5] S. Al-Yazidi, J. Berri, M. Al-Qurishi, and M. Al-Alrubaian, "Measuring reputation and influence in online social networks: a systematic literature review," *IEEE Access*, vol. 8, pp. 105824–105851, 2020. https://doi.org/10.1109/ACCESS.2020.2999033

[6] V. Ananthaswamy and B. Seethalakshmi, "Mathematical analysis of information dissemination model for social networking services," *Am. J. Model. Optim.*, vol. 3, no. 1, pp. 26–34, 2015. [Online]. Available: https://pubs.sciepub.com/ajmo/3/1/4/

[7] J. Dong, B. Chen, L. Liu, C. Ai, and F. Zhang, "The analysis of influencing factors of information dissemination on cascade size distribution in social networks," *IEEE Access*, vol. 6, pp. 54185–54194, 2018. [Online]. Available: https://ieeexplore.ieee.org/document/8472147

[8] P. Cui, B. Yin, and B. Xu, "The application of social recommendation algorithm integrating attention model in movie recommendation," *Sci. Rep.*, vol. 13, no. 1, pp. 169–188, 2023. [Online]. Available: https://www.nature.com/articles/s41598-023-43511-1. Accessed: Aug. 1, 2025.

[9] "Vybukhy u Kryvomu Rozi 4 kvitnia 2025 roku: shcho vidomo," Fakty.ua, 5 kvit. 2025. [Online]. Available: https://fakty.com.ua/ua/proisshestvija/20250405-vybuhy-u-kryvomu-rozi-4-kvitnya-2025-roku-shho-vidomo/ [in Ukrainian]. Accessed: Aug. 1, 2025.

[10] "Vybukhy u Kyievi u trokh raionakh stolytsi: stalysia pozhezhi cherez raketnu ataku RF," Fakty.ua, 6 kvit. 2025. [Online]. Available: https://fakty.com.ua/ua/proisshestvija/20250406-vybuhy-u-kyyevi-u-troh-rajonah-stolyczi-stalysya-pozhezhi-cherez-raketnu-ataku-rf/ [in Ukrainian]. Accessed: Aug. 1, 2025.

УДК 004.75 (004.62)

МАТЕМАТИЧНА МОДЕЛЬ КЛАСТЕРИЗАЦІЇ ІНФОРМАЦІЙНИХ ПОВІДОМЛЕНЬ З ІНДИКАТОРАМИ АКТИВНОСТІ ІНФОРМАЦІЙНОГО КОНТЕНТУ ЗА ТОНАЛЬНІСТЮ І СФЕРАМИ ДІЯЛЬНОСТІ СОЦІУМУ

Олексій Писарчук
Національний технічний університет України
«Київський політехнічний інститут імені Ігоря Сікорського», Київ, Україна
https://orcid.org/0000-0001-5271-0248

Данило Баран
Національний технічний університет України
«Київський політехнічний інститут імені Ігоря Сікорського», Київ, Україна
https://orcid.org/0009-0007-0361-6870

Отримала подальший розвиток математична модель кластеризації інформаційних повідомлень, яка базується на частотному аналізі їх тональності з використанням методологій обробки природної мови з підтримкою моделей великих мов; OLAP-візуалізації результатів кластеризації та відрізняється усталеною системою показників активності інформаційного контенту за сферами діяльності суспільства з ієрархічним стисненням вхідних масивів великих даних, що визначає модель бази даних для їх зберігання. Це забезпечує покращення аналізу інформаційних повідомлень у глобальних інформаційних мережах шляхом врахування багатьох факторів у сферах діяльності суспільства.

Основною ідеєю та призначення математичної моделі кластеризації інформаційних повідомлень є реалізація послідовності етапів підготовки до виявлення критичної активності інформаційного контенту в глобальних медіа. На практиці це створення списку та визначення значень показників, що відображають активність контенту в первинних повідомленнях, з подальшим їх перетворенням у часовий ряд – систематизований набір даних. В умовах високої щільності потоку виникнення, динаміки розвитку та трансформації інформаційного контенту враховується структура великих даних інформаційних повідомлень. Отже, модель кластеризації, окрім поділу за інформаційними ознаками забезпечувати ієрархічне стиснення вхідних масивів великих даних.

Мета дослідження: розробка математичної моделі кластеризації інформаційних повідомлень з показниками активності інформаційного контенту за тональністю та сферами діяльності суспільства. Предмет дослідження: методи кластеризації інформаційних повідомлень. Об'єкт дослідження: процес кластеризації інформаційних повідомлень.

Ключові слова: великі дані, кластеризація; природна мова.