I.J. Information Technology and Computer Science, **2025, 4, 80-119**
Published Online on August 8, 2025 by MECS Press (http://www.mecs-press.org/)
DOI: 10.5815/ijitcs.2025.04.07

Information Technology for Modelling Social Trends in Telegram Using E5 Vectors and Hybrid Cluster Analysis

Roman Lynnyk
Department of Information Systems and Networks, Institute of Computer Sciences and Information Technologies, Lviv Polytechnic National University, Lviv, 79013, Ukraine
E-mail: roman.o.lynnyk@lpnu.ua
ORCID iD: iD: https://orcid.org/0009-0007-0948-4338

Victoria Vysotska
Department of Information Systems and Networks, Institute of Computer Sciences and Information Technologies, Lviv Polytechnic National University, Lviv, 79013, Ukraine
E-mail: Victoria.A.Vysotska@lpnu.ua
ORCID iD: https://orcid.org/0000-0001-6417-3689

Zhengbing Hu
School of Computer Science, Hubei University of Technology, Wuhan, China
E-mail: drzbhu@gmail.com
ORCID iD: https://orcid.org/0000-0002-6140-3351

Dmytro Uhryn*
Department of Computer Science, Educational and Research Institute of Physical, Technical and Computer Sciences, Yuriy Fedkovych Chernivtsi National University, 58012, Ukraine
E-mail: d.ugryn@chnu.edu.ua
ORCID iD: https://orcid.org/ 0000-0003-4858-4511
*Corresponding author

Liliia Diachenko
Software Department, Educational and Research Institute of Physical, Technical and Computer Sciences, Yuriy Fedkovych Chernivtsi National University, 58012, Ukraine
E-mail: l.diachenko@chnu.edu.ua
ORCID iD: https://orcid.org/0000-0003-4451-3927

Kyrylo Smelyakov
Department of Software Engineering, Kharkiv National University of Radio Electronics, Nauky Ave. 14, Kharkiv, 61166, Ukraine
E-mail: kyrylo.smelyakov@nure.ua
ORCID iD: https://orcid.org/0000-0001-9938-5489

Received: 13 January 2025; Revised: 21 March 2025; Accepted: 16 May 2025; Published: 08 August 2025

Abstract: The article presents a modern approach to analysing public opinion based on Ukrainian-language content from Telegram channels. This study presents a hybrid clustering approach that combines DBSCAN and K-means algorithms to analyse vectorised Ukrainian-language social media posts in order to detect public opinion trends. The methodology relies on a multilingual neural network–based text vectorisation model, which enables effective representation of the semantic content of posts. Experiments conducted on a corpus of 90 Ukrainian-language messages (collected between March and May 2025) allowed for the identification of six principal thematic clusters reflecting key areas of public discourse. Despite the small volume of the corpus (90 messages), the sample is structured and balanced by topic (news, vacancies, gaming), which allows you to test the effectiveness of the proposed methodology in conditions of limited data. This approach is appropriate in the case of the analysis of short texts in low-resource languages, where large-scale corpora are not available. A special advantage of this approach is the use of semantic vector representation and the construction

This work is open access and licensed under the Creative Commons CC BY 4.0 License.

of graphs of lexical co-occurrence networks (term co-occurrence networks), which demonstrate a stable topological structure even with small amounts of data. It allows you to identify latent topic patterns and coherent clusters that have the potential to scale to broader corpora. The authors acknowledge the limitations associated with sample size, but emphasise the role of this study as a pilot stage for the development of a universal, linguistically adaptive method for analysing public discourse. In the future, it is planned to expand the body to increase the representativeness and accuracy of the conclusions. The paper proposes a hybrid method of automatic thematic cluster analysis of short texts in social media, in particular Telegram. Vectorisation of Ukrainian-language messages is implemented using the transformer model multilingual-e5-large-instruct. A combination of HDBSCAN and K-means algorithms was used to detect clusters. More than 36,000 messages from three Telegram channels (news, games, vacancies) were analysed, and six main thematic clusters were identified. To identify thematic trends, a hybrid clustering approach was used, in which the HDBSCAN algorithm was used at the first stage to identify dense clusters and identify "noise" points, after which K-means were used to reclassify residual ("noise") embeddings to the nearest cluster centres.
Such a two-tier strategy allows you to combine the advantages of flexible allocation of free-form clusters from HDBSCAN and stable classification of less pronounced groups through K-means. It is especially effective when working with fragmented short texts of social networks. To validate the quality of clustering, both visualisation tools (PCA, t-SNE, word clouds) and quantitative metrics were used: Silhouette Score (0.41) and Davis-Boldin index (0.78), which indicate moderate coherence and resolution of clusters. Separately, the high level of "noise" in HDBSCAN (34.2%) was analysed, which may be due to the specifics of short texts, model parameters, or stylistic fragmentation of Telegram messages. The results obtained show the effectiveness of combining modern vectorisation models with flexible clustering methods to identify structured topics in fragmented Ukrainian-language content of social networks. The proposed approach has the potential to further expand to other sources, types of discourse, and tasks of digital sociology. As a result of processing 90 messages received from three different channels (news, gaming content, and vacancies), six main thematic clusters were identified. The largest share is occupied by clusters related to employment (28.2%) and security-patriotic topics (24.7%). The average level of "noise" after the initial HDBSCAN clustering was 34.2%. Additional analysis revealed that post lengths varied significantly, ranging from short announcements (average of 10 words) to analytical texts (over 140 words). Visualisations (timelines, PCA, t-SNE, word clouds, term co-occurrence graphs) confirm the thematic coherence of clusters and reveal changes in thematic priorities over time. The proposed system is an effective tool for detecting information trends in the environment of short, fragmented texts and can be used to monitor public sentiment in low-resource languages.

Index Terms: Text Clustering, Public Opinion, HDBSCAN, K-means, Thematic Analysis, Text Vectorisation, Clustering, Neural Networks, Ukrainian Language, Natural Language Processing, Short Texts, Trend Analysis.

1. Introduction

In today's rapidly evolving information landscape, the automated analysis of user-generated content from social media and messaging platforms has become a critically important task, particularly when it comes to understanding public opinion and tracking the dynamics of socially relevant discourse. Among these sources, Ukrainian-language communities in social networks, especially Telegram channels, play a pivotal role in shaping public attitudes, reflecting collective concerns, and surfacing emerging societal trends. The proliferation of social media has led to an exponential increase in the volume of unstructured textual data. These texts, however, are often short, fragmented, linguistically informal, and contextually diverse. As a result, they pose significant challenges to classical Natural Language Processing (NLP) techniques. For instance, traditional topic modelling algorithms such as Latent Dirichlet Allocation (LDA) tend to produce suboptimal results when applied to sparse and heterogeneous datasets, often failing to capture the nuanced semantic relationships within short texts. To overcome these limitations, this study adopts a modern NLP-driven pipeline that leverages semantic vectorisation using advanced multilingual transformer-based models. These models, specifically E5 in our case, encode the semantic core of a text into a dense vector representation, enabling the grouping of similar posts based on meaning rather than surface-level keywords.

The objective of this research is to explore a hybrid clustering framework for detecting latent thematic trends in Ukrainian-language posts. The proposed approach consists of the following key stages: data collection from publicly accessible thematic Telegram channels, text pre-processing (including cleaning, tokenization, and lemmatization), semantic embedding using the E5 model, and clustering using a two-phase strategy that combines HDBSCAN (to capture dense semantic clusters) with K-means (to handle previously unclustered points). The experimental results are presented using a comprehensive visualisation suite: time series plots reflecting the temporal evolution of topics, heatmaps illustrating cluster density, boxplots for post-length distribution, pie charts summarising topic proportions, and co-occurrence graphs that map frequently mentioned terms within and across clusters. By synthesising recent advances in neural vectorisation and unsupervised learning, this approach demonstrates how meaningful insights can be extracted from fragmented and informal textual content. The methodology not only enhances our ability to monitor social discourse but also contributes to scalable tools for public sentiment analysis in low-resource languages like Ukrainian.

1.1. Research Motivation and Methodological Framework

Social networks, particularly Telegram, have become vital platforms for disseminating information and engaging in public discourse. However, the overwhelming volume of unstructured content poses significant challenges for timely and reliable data analysis. In the case of Ukrainian-language content, additional complexity arises from the linguistic specificity of the language and the brevity of individual messages, which often consist of only a few words or sentences. These characteristics necessitate the adaptation of existing NLP methods to ensure effective processing and semantic interpretation. This research aims to develop and implement a comprehensive methodology for analysing public opinion trends in Ukrainian-language social media content using modern machine learning techniques, including neural embeddings and clustering. To achieve this, the following key objectives were defined:

- *Data Collection and Pre-processing*. Establish a systematic approach for collecting and cleaning Ukrainian-language messages from public Telegram channels, including filtering, tokenisation, and normalisation.
- *Text Vectorisation*. Select and configure a multilingual transformer-based language model capable of producing high-quality semantic representations of short texts. In this work, we employ a state-of-the-art transformer model optimised for cross-lingual embedding of brief user-generated content.
- *Hybrid Clustering*. Design a two-phase clustering strategy combining DBSCAN and K-means algorithms to identify both dense thematic clusters and more diffuse groups that may not meet the density threshold but still share semantic proximity.
- *Visualisation of Results*. Develop visualisation tools to effectively represent both raw and processed data, including activity timelines, word clouds, PCA projections, heatmaps, and co-occurrence graphs.

This methodology responds to a growing demand for agile tools that support real-time monitoring of public discourse and the detection of evolving social trends, particularly in low-resource language settings such as Ukrainian.

1.2. Problem Definition

The analysis of social media content presents several methodological and technical challenges. First, the short length of posts – typically limited to a few dozen words – restricts the contextual window available for thematic inference. Second, these texts often include informal language, abbreviations, emojis, and internet-specific slang, which requires careful pre-processing to normalise the content before vectorisation. Third, social media activity is often event-driven, resulting in bursts of content that reflect real-time developments; these temporal spikes must be considered in any trend analysis. To address these challenges, the proposed system integrates multiple components:

- Natural Language Processing techniques for text cleaning, tokenisation, and lemmatisation;
- Multilingual neural embedding models for generating high-dimensional semantic representations of posts;
- Unsupervised clustering algorithms (HDBSCAN and K-means) for organising the data into meaningful groups;
- Visualisation modules are used to interpret the output in both quantitative and temporal terms.
- The resulting pipeline includes the following stages:
- Collect Ukrainian-language posts from Telegram using the Telegram API;
- Pre-process the texts to remove noise and normalise structure.
- Generate semantic embeddings for each post using a multilingual transformer model.
- Apply HDBSCAN to identify dense semantic clusters, followed by K-means to classify remaining outliers;
- Visualise the outcomes using time series charts, heatmaps, boxplots, pie charts, and keyword co-occurrence graphs.

This integrated approach facilitates the discovery of latent thematic structures and enables dynamic monitoring of evolving public opinion in real-time environments.

The main goal of the study is to develop and test a hybrid method for identifying information trends and thematic clusters in short Ukrainian-language texts of social media (in particular, Telegram), based on transformer vector text representation, network co-occurrence graphs and combined clustering (HDBSCAN + K-middles).

Research objectives:

- Collect and pre-process a fundamental Ukrainian-language corpus of short messages from social networks (Telegram).
- Vectorise texts using the multilingual transformer model multilingual-e5-large-instruct.
- Build a semantic graph of the meeting of terms based on embedding similarity.
- Apply a hybrid clustering approach (HDBSCAN + K-means) to identify thematic groups.
- Evaluate the quality of clustering using Silhouette, DBI, etc., metrics.
- Interpret the results in terms of information patterns and the system's potential to monitor public sentiment.

Contribution of the study. The study offers an integrated and scalable approach to identifying thematic structures in short Ukrainian-language texts of social media, which:

- Combining the advantages of transformer embedding (without the need for fine-tuning),
- Uses hybrid clustering (density + centroid),
- Generates semantic graphs that facilitate the visualisation and interpretation of topics,
- Adapted to low-resource conditions (small sample, language with limited NLP resources),
- Shows potential for expansion into massive enclosures.

It is one of the first examples of the practical use of modern transformer embedding for Ukrainian-language Telegram content with a focus on analysing public sentiment.

Structure of the article. Chapter 2 provides an overview of related research in the areas of social media thematic analysis, short text clustering, and vectorisation of low-resource languages. Chapter 3 describes the research methodology, including data collection, vectorisation, co-encounter graph construction, and a hybrid approach to clustering. Chapter 4 highlights the results of clustering and graph analysis and discusses their interpretation. Chapter 5 provides an assessment of model quality using quantitative metrics. The limitations of the study, ethical aspects and prospects for further work are discussed. The article concludes with a summary of the main results.

2. Related Works

Telegram is increasingly used as a data source in social, informational, and security research. Its popularity in conflict and multilingual environments (in particular in Ukraine, Belarus, Iran, Russia, Syria) is due to:

- High anonymity;
- The presence of public channels with a large audience.
- The ability to quickly distribute content without moderation.

Analysis of Telegram in the context of conflict. The work [1] analyses how Telegram is used as a refuge for censored or radicalised groups, and how this creates reservoirs of uncontrolled information. Telegram plays the role of an alternative platform for war propaganda, in particular during the Syrian conflict. Studies [2-4] show that Ukrainian and pro-Russian Telegram channels function as parallel information systems. The authors classify channels according to tone, functions (informing, mobilisation, ideologisation) and target audience. The key methodology is clustering, taking into account semantic and tonal features.

Telegram in multilingual environments. Research [5] covers multilingual content, in particular in Telegram groups that cross language barriers. The need to process surzhyk ["mixed language," "hybrid language," або "linguistic mishmash"], dialects and transliterations, especially in the context of Ukrainian-Russian content, is indicated. Studies [6-10] demonstrate an approach to automatic clustering in low-resource languages with post-factum semantic interpretation. Key contribution: vector normalisation of dialect variants + control of topics through weak labels (emojis, hashtags).

Methods of visual analysis of trends and narratives [11-13] are based on the role of t-SNE/PCA in detecting group polarisation. They call for the use of combined visualisations (time graphs + semantic maps) to monitor narratives.

Existing studies of Telegram and other social networks as a source of data in conflict and multilingual contexts [14-30] confirm its significance for analysing the dynamics of public opinion and identifying information trends.

Approaches to analysing short textual content in social media environments have been extensively explored in the international research community [31]. A significant body of work focuses on clustering tweets and social media messages to detect emerging public opinion trends. For instance, Prokipchuk et al. [32] proposed a method for identifying trends in short-form Twitter messages by grouping them into semantic clusters and generating graphical "fingerprints" to represent the resulting structures. Their study demonstrated that combining traditional vector representations, such as Bag-of-Words (BoW) and Term Frequency – Inverse Document Frequency (TF-IDF), with transformer-based embeddings (e.g., BERT), in conjunction with clustering algorithms such as K-means, hierarchical clustering, and HDBSCAN, can yield robust results for sparse micro-texts.

In the context of low-resource languages such as Ukrainian, the availability of pre-trained language models remains limited. Panchenko et al. [33] (ICTERI-2021) introduced a Ukrainian news corpus. They showed that multilingual models like XLM-R tend to perform better on longer documents, whereas localised models, such as ukr-RoBERTa, are more effective on short text snippets. These findings are particularly relevant to our study, as they highlight the need for tailored embedding solutions when dealing with compact and informal user posts. Furthermore, the authors emphasised the importance of high-quality data pre-processing and the use of contrastive (positive/negative) sample generation in the training phase to enhance the performance of sentence-level embeddings.

Regarding clustering techniques, both parametric and non-parametric approaches have proven helpful in analysing textual embeddings. Among the non-parametric methods, DBSCAN and its improved hierarchical variant HDBSCAN have gained popularity due to their ability to identify clusters of arbitrary shapes without requiring a pre-defined number

of clusters. HDBSCAN, in particular, constructs a hierarchy of clusters based on varying density thresholds and is less sensitive to parameter tuning, making it well-suited for complex, variable-structure datasets.

Conversely, K-means remains one of the most widely used clustering algorithms due to its simplicity and computational efficiency. It partitions vectors into K clusters by minimising the sum of squared Euclidean distances between vectors and their respective centroids. The objective function for K-means can be formally expressed as [34]:

$$L_{KMeans} = \sum_{k=1}^{K} \sum_{x_i \in C_k} \left|| x_i - \mu_k \right||^2 \quad (1)$$

where K – is the number of clusters, C_k – is the set of vectors assigned to the k^{th} cluster, x_i is an individual vector and μ_k – is the centroid of the k^{th} cluster. The algorithm thus seeks to form cohesive groups by aligning points closely to their respective cluster centres in a high-dimensional semantic space.

These methodological insights inform our selection of tools and techniques for detecting thematic trends in Ukrainian-language social media content. The integration of robust vectorisation models with hybrid clustering methods enables scalable, interpretable, and linguistically adaptive analysis pipelines.

2.1. Advanced Methods for Analysing Social Media Texts

The analysis of social media content has evolved into a robust interdisciplinary field that blends techniques from computer science, computational linguistics, and the social sciences. The rapid growth of platforms such as Twitter and Telegram has led to the production of massive volumes of unstructured textual data. These datasets are typically characterised by their volume, variety, and velocity, demanding structured workflows for data collection, pre-processing, representation, and visualisation. According to Capuccinelli et al. [35], social media has transformed over the past two decades from basic communication tools into complex information ecosystems with a significant impact on public opinion formation. In their review of 132 academic publications, the authors highlighted key research domains such as user-generated content analysis, behavioural aspects of social media usage, and the integration of social platforms into marketing and organisational decision-making. From a mathematical perspective, the social media text analysis pipeline can be abstracted as a sequence of transformations:

$$T: D \rightarrow V \rightarrow C \rightarrow I \quad (2)$$

where D represents the raw dataset of documents, V denotes the vectorised representations of the texts, C is the set of discovered clusters or classes, I represents the interpretation and insights extracted from the clustering process [36].

2.2. Text Vectorisation and Transformer-based Models

A critical step in the processing of textual data is vectorisation, the conversion of human language into numerical representations that retain semantic meaning and are suitable for machine learning algorithms. Traditional methods such as TF-IDF (Term Frequency – Inverse Document Frequency) are limited in their ability to capture word context and long-range semantic dependencies. To address these limitations, this study employs the multilingual-e5-large-instruct model as a part of a new generation of universal text embedding models designed to support a wide range of languages and domains [37]. The model architecture is based on a bidirectional transformer encoder that captures both contextual and syntactic information. As described by Wong et al. [38], the model was trained using a large-scale contrastive learning approach on approximately one billion multilingual text pairs, including both real and synthetically generated data. It supports over 100 languages and includes 24 layers with a 1024-dimensional output space. The embedding process can be formally described as follows:

$$e_i = E5(x_i) \in R^d \quad (3)$$

where x_i – is the input text, e_i – is the resulting embedding vector of dimension d, and E5 is the transformation function implemented by the model [38].

2.3. Clustering Algorithms for Textual Data

Clustering is a foundational technique for unsupervised pattern discovery and organisation of large-scale datasets. In the context of social media analysis, clustering is used to group semantically similar messages, uncover latent topics, and monitor trend dynamics. Among the most prominent clustering methods used for text data are:

- K-means. A widely used algorithm that partitions data by minimising the sum of squared distances between each point and its assigned cluster centroid. The objective function is [39]:

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left|\left| x_i^{(j)} - c_j \right|\right|^2 \quad (4)$$

where c_j – is the centroid of cluster j, and $x_i \in C_j$ – are the points assigned to that cluster.

- DBSCAN (Density-Based Spatial Clustering of Applications with Noise) [40-41]. A density-based algorithm that effectively identifies arbitrarily shaped clusters and filters out noise.
- Hierarchical Clustering. Builds a tree-like structure of nested clusters by recursively merging or splitting groups based on a distance metric [42].

Recent studies have emphasised the benefits of hybrid clustering approaches. Jain et al. [36], for instance, demonstrated that combining K-means with metaheuristic optimisation techniques such as Ant Lion Optimisation yields improved clustering quality in terms of intra-cluster distance and F-score, outperforming standalone methods such as DBSCAN and its modified variants.

In our study, we adopt a hybrid clustering strategy that leverages DBSCAN to identify high-density clusters and then applies K-means to assign the remaining "noise" points to semantically closest groups. This combination allows for the detection of both tightly knit and more diffuse topic groups while maintaining computational efficiency.

2.4. DBSCAN Parameter Optimization

The performance of DBSCAN largely depends on the choice of its two core parameters: ε (the radius of the neighbourhood) and MinPts (the minimum number of points required to form a dense region). Optimal values for these parameters are data-dependent and can significantly affect clustering outcomes. A commonly used heuristic for selecting ε is the k-distance graph, which plots the distance from each point to its k^{th} nearest neighbour. The "elbow point", where the curve shows a sharp bend, is typically selected as the optimal radius. The k-distance can be defined formally as:

$$k - distance(p) = d(p,o)||\{o' \in D|d(p,o') \leq d(p,o)\}| \geq k \wedge |\{o' \in D|d(p,o') < d(p,o)\}| < k \quad (5)$$

where D is the dataset and $d(p,o)$ is the distance between points p and o [42].

As for $MinPts$, a common empirical rule suggests setting it to at least the dimensionality of the data plus one: $MinPts \geq dim + 1$. For high-dimensional embeddings, values such as $MinPts = 2 \times dim$ are often recommended.

More advanced methods for parameter optimisation include grid search and Bayesian optimisation, which systematically explore combinations of ε and $MinPts$ to maximise internal clustering metrics. Del Río et al. [37] proposed a hybrid approach combining Factor Analysis (FA) for dimensionality reduction and a Genetic Algorithm (GA) for parameter optimisation. Their experimental results show that the FA+GA-DBSCAN approach delivers high-accuracy clustering in datasets with varying density levels.

3. Materials and Methods

This study implements a systematic methodology for analysing opinion trends in Ukrainian-language posts from social networks. The overall process consists of five main stages, each contributing to the transformation of raw textual data into interpretable insights. The initial stage involves collecting data from public Telegram channels using the Telegram API. The collected messages are filtered by language and topic relevance to ensure the integrity of the dataset.

Following collection, the texts undergo pre-processing to prepare them for further computational analysis. It includes removing HTML artefacts, memorable characters, and stop words, as well as applying lemmatisation to standardise word forms. Since social media messages often contain informal language, abbreviations, and emojis, the pre-processing module is designed to handle such linguistic irregularities with care.

The third stage involves semantic vectorisation, where pre-processed texts are embedded into a high-dimensional vector space using a multilingual transformer-based model. These embeddings capture the semantic similarity between posts and serve as the foundation for subsequent clustering.

In the fourth stage, clustering algorithms are employed to group semantically similar posts. A hybrid approach is used: DBSCAN identifies dense, well-defined topic clusters, while K-means assigns the remaining outlier vectors to nearby cluster centroids, thereby capturing both core themes and peripheral discussions.

Finally, the results of the clustering process are visualised to aid in interpretation and analysis. Various visualisation techniques are employed, including time series graphs to track topic evolution, word clouds to highlight frequent terms, PCA scatter plots to illustrate cluster distribution, heatmaps to reflect density, pie charts to display topic proportions, and keyword co-occurrence graphs to explore lexical relationships within and across clusters.

The entire architecture of the system is represented in the UML component diagram (Figure 1), which illustrates the logical structure and modular responsibilities of each component involved in the pipeline.

To collect the necessary data, a dedicated script was developed to retrieve posts from Ukrainian-language Telegram channels. The script stores the extracted content in a structured JSON format suitable for downstream processing. Once the message history is downloaded for the desired period, the system filters out all non-text content types – such as media attachments and forwarded messages - retaining only original textual posts for analysis.

Pre-processing of social media texts is a critical step that significantly affects the accuracy of subsequent semantic analysis. It is vital when working with user-generated content, which often includes slang, abbreviations, emojis, and irregular punctuation patterns. Without adequate normalisation, such elements can distort statistical and semantic relationships between texts.

The pre-processing pipeline begins by converting all text to lowercase, thereby reducing variation due to case differences. Next, irrelevant characters such as punctuation marks, hyperlinks, and emojis are removed to eliminate noise. The text is then tokenised into individual words. From this token list, common stop words – such as conjunctions, prepositions, and other function words – are filtered out. The final stage involves lemmatisation, which reduces words to their base or dictionary forms, thereby standardising morphological variants and improving vector consistency.

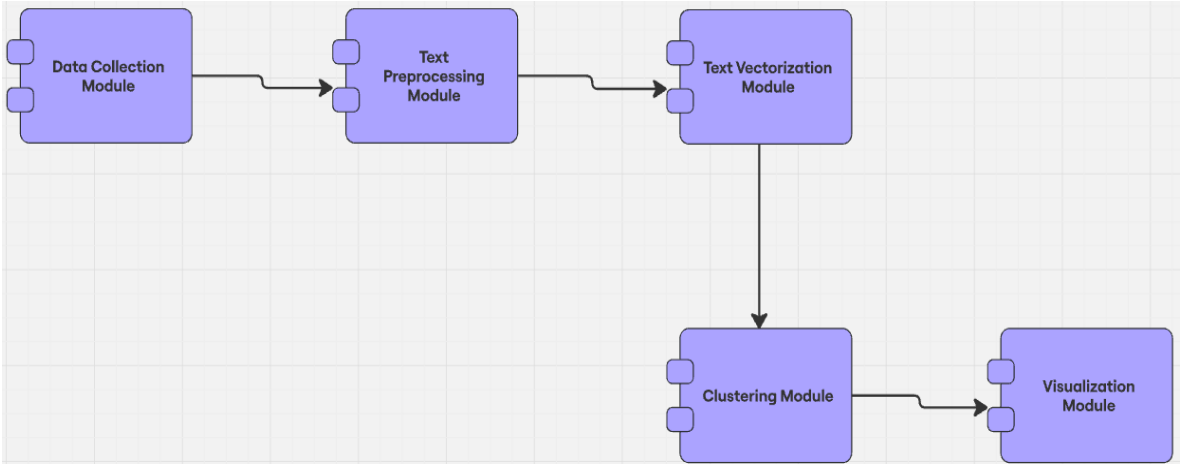


Fig.1. UML Component Diagram

3.1. Data Collection and Pre-processing

This multistage process can be formally described by the following transformation [43]:

$$P(t) = (f_n \circ f_c \circ f_t \circ f_s \circ f_l)(t) \tag{6}$$

where t is the raw input text, and the composite function represents the sequential application of normalisation f_n character-level cleaning f_e, tokenization f_t, stop-word removal f_S and lemmatization f_l [43].

For Ukrainian-language content, special attention is paid to lemmatisation, as the Ukrainian language is morphologically rich, featuring complex declension and conjugation patterns. To address this challenge, a language-specific lemmatiser adapted for the informal and abbreviated style typical of social media was used [44].

3.2. Text Vectorisation Using multilingual-e5-large-instruct

To transform pre-processed Ukrainian-language texts into numerical representations suitable for semantic clustering, this study employed the multilingual-e5-large-instruct model [36]. This model belongs to a new generation of instruction-tuned transformer architectures optimised for multilingual Natural Language Processing (NLP) tasks. It has demonstrated high performance across a range of languages and use cases without requiring task-specific fine-tuning [45].

A defining feature of the model is its instruction-tuning capability. Prior to embedding, each input text was paired with a concise natural language instruction that clarified the intended task (e.g., semantic similarity search). This instruction-text pair was then passed to the model, which produced context-aware embeddings based on both components.

The output vectors were subsequently aggregated using average pooling to form a single 1024-dimensional embedding for each post. The process can be formally expressed as [46]:

$$= AvgPool(E5(I, x)) \tag{7}$$

where $E5(I, x)$ denotes the model's contextual embedding of the instruction I and input text x, and $AvgPool$ represents the average pooling operation applied over the token-level embeddings.

This instruction-aware embedding strategy enables the model to generalise across tasks by interpreting semantic nuances as guided by the instruction [47]. In the context of this research, instructions were crafted to emphasise semantic similarity, thereby guiding the model to produce embeddings that reflect latent topical relatedness between social media posts. Such a strategy not only improves clustering performance but also enhances the interpretability of results by aligning the embedding space more closely with human-level semantic expectations.

3.3. Features of the Application of the Multilingual-e5-large-instruct Model and Reasons/Advantages of its Choice

The model was not configured additionally (without fine-tuning). A ready-made instruction-configured version(multilingual-e5-large-instruct) was used, without additional training on Ukrainian data. Vectorisation was carried out in the format: "[INSTRUCTION] + text", for example: "query: Які вакансії сьогодні у Києві?" ["query: What vacancies are available in Kyiv today?"]. The instructions were manually generated to activate the appropriate behaviour

of the model, as recommended in the E5 documentation. The resulting sentence embeddings – 1024-dimensional vectors – were used as input to clustering. E5 is the best balance of performance and semantic relevance without the need for additional training. It is adapted to short sentences, which is ideal for Telegram messages.

Table 1. Reasons for choosing the multilingual-e5-large-instruct model over XLM-RoBERTa, mBERT or LASER

Model	Dignity	Limitations
E5 (selected)	– Support >100 languages, including Ukrainian – Trained in the style of "sentence similarity" (semantic search) – Instruction tuning – High scores on MTEB for low-resource languages	– Without additional training on Ukrainian domains
XLM-RoBERTa	– Powerful model – Works well on large documents	– Works worse on short texts – Focused more on tokenisation than sentences
mBERT	– Popular – Basic support for Ukrainian	– Poor quality sentence embeddings – No specific training for semantic similarity
LASER	– Good tongue coverage – Compressed architecture	– Significantly lower accuracy than E5 – Older technology

3.4. Problems of Vectorisation of Ukrainian-language Content as a Low-resource Language

The Ukrainian language has a number of specific problems that can affect the quality of embeddings.

Morphological complexity – a large number of cases, suffixes, and inflexions. It can reduce the similarity between similar sentences if the lemmatisation is carried out incorrectly. The study used the previous lemmatisation by the Ukrainian lemmatiser, which reduced the problem, but errors in lemmatisation can distort embedding.

Uneven coverage in multilingual models. Even in models that support 100+ languages, Ukrainian has fewer parameters/frequency examples compared to English, Spanish, etc. It reduces the quality of embeddings for rarely used words (proper names, slang, transliterations). E5 mitigates this through contrastive learning on multilingual pairs, but the problem remains, especially for:

- innovations («руzzня» ["ruzznya"], «орки» ["orcs"]),
- memes and abbreviations,
- informal forms.

Telegram as a source is an additional difficulty because the texts are short, often without punctuation. They contain emojis, jargon, and reposts that are not always interpreted by the transformer. Texts have been cleaned, but such semantic losses always affect the quality of vectors.

3.5. Hybrid Clustering Algorithm: DBSCAN + K-means

This study employs a hybrid clustering approach that integrates the strengths of two complementary algorithms – DBSCAN and K-means - to effectively identify semantic groupings in textual data. The combined strategy enables the detection of clusters with varying densities and arbitrary shapes, as well as the separation of outliers, which is particularly important when analysing noisy and diverse social media content [48].

The hybrid DBSCAN+K-means algorithm is designed as a two-stage process. In the first stage, the DBSCAN (Density-Based Spatial Clustering of Applications with Noise) algorithm is applied. This algorithm identifies high-density regions in the embedding space without requiring the number of clusters to be specified a priori. It classifies data points as core, border, or noise based on two parameters: the neighbourhood radius ε, and the minimum number of points $MinPts$ within that radius. These parameters are determined empirically, often with the aid of the k-distance graph, which helps identify the "elbow point" that corresponds to an optimal ε value [42].

Once DBSCAN has established the main clusters, the second stage involves applying the K-means algorithm to the residual "noise" points that DBSCAN left unassigned. The number of cluster centres for K-means is estimated using silhouette analysis, which measures cluster cohesion and separation. This step is beneficial for recovering semantically weak or less dense topics that may have been missed by the density-based method.

The final step involves merging the results of both clustering stages into a unified classification. This hybrid approach enables the model to capture both compact and sparse topic clusters, improving the semantic coherence of the overall grouping [49]. The procedure for optimising DBSCAN parameters includes the following:

- Computing the distance from each point to its k^{th} nearest neighbour;
- Sort these distances in ascending order.
- Plotting the sorted values to visualise the k-distance curve;
- Identifying the elbow point, which suggests the optimal radius ε.

In our experiments, using embeddings generated by the multilingual-e5-large-instruct model [36], the optimal DBSCAN parameters were determined to be $MinPts = 5$ and $\varepsilon = 0.35$, which provided the most coherent clustering in terms of semantic content.

Cluster centroids in both stages are calculated as the mean vector of their respective member embeddings [50]:

$$c_k = \frac{1}{|C_k|}\sum_{e_i \in C_k} e_i \tag{8}$$

where $|C_k|$ – is the number of vectors in the cluster $|C_k|$ and e_i – is the embedding of the i^{th} document. By leveraging the strengths of both algorithms, the hybrid DBSCAN+K-means framework provides a robust and flexible method for discovering semantic structures in multilingual short-text corpora.

3.6. Identification of Representative Posts and Cluster Topics

To characterise the thematic content of each cluster and facilitate interpretability, this study employs a two-step approach: the selection of representative posts and the generation of cluster topic labels.

Representative posts within each cluster are identified by calculating the cosine similarity between the embedding of a post and the centroid of its assigned cluster. Cosine similarity is a widely used metric for measuring vector proximity in high-dimensional semantic spaces. It is defined as [51]:

$$sim(e_i, c_k) = \frac{e_i * c_k}{||e_i|| * ||c_k||} \tag{9}$$

where e_i – is the embedding of the i-th post, and c_k – is the centroid of cluster k. Posts with the highest similarity to the cluster centroid are considered the most representative in terms of semantic alignment.

Once representative posts are selected for each cluster, a keyword analysis is performed to generate informative topic labels. The method begins by extracting the top-n posts with the highest cosine similarity to the cluster centre. These posts are tokenised and lemmatised to standardise word forms, reducing morphological variation.

Next, the frequency of each lemma is computed across the selected posts. The most frequently occurring terms are then ranked, and the top five are used to synthesise a concise summary of the cluster's thematic focus. This high-frequency lexicon is manually or algorithmically condensed into a short, descriptive topic title.

This keyword-driven topic extraction method ensures that cluster labels remain both semantically accurate and human-readable, effectively reflecting the core discourse represented in each group.

As part of hybrid clustering, a phased strategy was applied. In the first stage, HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) was used to automatically select dense clusters without the need to pre-set their number. An essential property of HDBSCAN is the ability to classify a piece of data as "noise", i.e. such points that do not belong to any dense cluster – this is especially useful when analysing short, unevenly distributed texts, such as Telegram messages. After that, in the second step, the K-means algorithm was applied to vectors that remained unclassified by HDBSCAN. The number of clusters for K-means was determined using silhouette analysis, which provides a balance between over-fragmentation and over-aggregation. Thus, K-means made it possible to reduce the percentage of "noise" points (~34.2% after HDBSCAN) and provide a more complete thematic coverage of the case. This approach allows you to compensate for the weaknesses of each method: HDBSCAN is good at detecting compact, high-density topic areas (e.g. recurring vacancies or structured news), but skips less coherent semantic groups; instead, K-means efficiently aggregate residual points into coherent thematic segments, based on minimising the distance to the centers. The complete integration procedure is as follows:

- Vectorisation of messages using the multilingual-e5-large-instruct model;
- HDBSCAN clustering ($\varepsilon \approx 0.35$, MinPts = 5) – classifies the main clusters and "noise";
- Reclassification of residual ("noise") points from HDBSCAN using K-means (number of clusters selected by silhouette coefficient);
- Visualisation and thematic labelling of the combined cluster space.

3.7. Visualisation of Results

Multiple visualisation techniques were employed as part of the analytical workflow to enhance interpretability and provide insights into the structure of the clustered data. These visual representations serve not only to validate the quality of clustering but also to reveal temporal and lexical patterns within the dataset.

Activity timelines were used to display the distribution of post frequency over time. These plots highlight periods of increased publication activity, which often coincide with real-world events or shifts in public attention. By mapping volume fluctuations, the analysis can detect when specific topics emerged or peaked in relevance. Word clouds were generated for each cluster based on the most frequently occurring lemmatised terms. These visualisations provide an intuitive snapshot of a cluster's lexical content, enabling quick identification of dominant keywords and thematic cues.

To assess the spatial structure of the clusters, Principal Component Analysis (PCA) was applied to reduce the dimensionality of the 1024-dimensional sentence embeddings produced by the multilingual-e5-large-instruct model. PCA projects the data onto a two-dimensional plane while preserving as much variance as possible. Formally, PCA identifies the principal axes (linear combinations of original dimensions) that maximise variance in the data distribution. It enables visual inspection of how well-separated or overlapping the clusters are in the embedding space.

Additionally, heatmaps were used to track the temporal dynamics of topics. These matrices show the relative prevalence of each cluster over time intervals, allowing the identification of long-term trends, periodic surges, or sudden topic shifts. Heatmaps provide a compact and effective means of visualising cluster evolution.

By combining these complementary techniques, the system offers a holistic view of both the static thematic structure and the temporal behaviour of Ukrainian-language social media discourse. This multidimensional visualisation strategy ensures that the results are not only analytically robust but also accessible to domain experts and decision-makers.

The peculiarity of influencing and promoting visualisation to understand changes in public opinion:

PCA / t-SNE – detection of thematic structure. They reduce multidimensional vector spaces (embeddings) to 2D or 3D for interpretation. Dense clusters, overlapping topics, and isolated groups of messages are visually visible. Clusters give an idea of the topics prevailing in the community. Overlap between clusters indicates hybrid or transitional narratives (e.g. "war + volunteering"). Displacement or emergence of new clusters in time → identification of new topics/trends (e.g., activation of the topic of mobilisation in a specific period). Example of interpretation: On t-SNE, you can see that since mid-May, a new cluster has emerged related to announcements of fundraising for the Armed Forces of Ukraine – this allows us to indicate the strengthening of the topic of support for the army.

Time graphs (number of messages in a topic/cluster by time). On the X axis is the time (for example, days), and on the Y axis is the number of messages in each cluster. Visually shows when a particular topic has begun to grow or decline. Allows you to establish a connection with real events (news, changes in policy, war events). Analytical value, i.e. such graphs can be used for:

- early detection of information campaigns;
- monitoring sentiment (for example, a decrease in job messages → an economic alarm);
- evaluating the effectiveness of government or activist communications.

Word clouds (per cluster or per time window) visualise the frequency of terms in a cluster or within a specific period. Gives a quick idea of the lexical core of the topic. When comparing over time, you can see how the foci of the discourse change. Example: in the cluster "вакансії" ["vacancies"] in June, the words «Польща» ["Poland"], «зарплата» ["salary"], «зміна» ["shift"] appear more often – which may indicate seasonal migration or a change in the labor market.

Columns of terms/co-occurrences of words. Nodes – words, edges – common occurrences in messages (cohesion of terms). The structure of narratives is revealed: which terms are central, which form subtopics. You can find:

- disinformation constructs (such as "гуманітарка" ["humanitarian"] + "показуха" ["ostentation"]);
- emotional connections (such as "страх" ["fear"] + "обстріл" ["shelling"]).

Table 2. Decision support

Scope of application	How visualisations help
Journalism	Showing trends and topics that are "gaining weight", selection of tricks for reports
Security Analytics	Early detection of mobilisation calls, anti-government rhetoric
Public monitoring	Analysis of Population Sentiment by Key Topics
Policy/Management	Determining the effectiveness of communications (e.g., government statements) or responses to them
Marketing / HR	Identification of topics that form motivation/anxiety in job seekers or consumers

Although a number of visualisations are used in the work (PCA, t-SNE, time graphs, word clouds), it is advisable to emphasise their role not only as an illustration tool, but as a key analytical tool for detecting changes in public opinion. In particular, time graphs show peaks of interest in topics, word clouds reveal the transformation of vocabulary within a cluster, and cohesive term graphs allow for the identification of central narratives. The combination of these methods forms the basis for a deeper understanding of public discourse and is of applied importance for journalism, digital sociology, security analytics and public administration.

3.8. The Specific Problem that the Research Solves (Research Gap)

Existing methods of thematic analysis of social media often:

- Ineffective for short, unstructured messages specific to Telegram;
- Focused on English-speaking or resource-rich languages;
- Require a large amount of training data or manual labelling;
- Demonstrate low clustering accuracy for fragmented or stylistically heterogeneous enclosures;
- Do not offer interpreted results through graph imaging.

The proposed approach overcomes these limitations by:

- Use of a powerful multilingual transformer model without fine-tuning;

- Clustering adapted to "noisy" data;
- Integration of visual graphs to explain the results.

3.9. Gaps in Current Fraud Detection Methods that the Study Aims to Address

While the primary goal of this study is to analyse information trends, the proposed technology also partially addresses the problems inherent in social media fraud detection systems, including:

- Insufficient adaptation to short, unstructured messages:
  - Most traditional fraud detection systems are designed for formalised transactions, not text streams (especially like Telegram).
  - Our approach allows you to cluster short fragments of text where there may be a sign of fraud (for example, repetitive "робота без досвіду" ["work without experience"], "виплати сьогодні" ["payments today"]).
- Inability to detect thematic anomalies without labelled data:
  - Classic fraud models require training on reference kits.
  - Our approach works without manual annotation, using clustering and semantic proximity, which can be helpful in identifying suspicious topics (e.g., illegal jobs, mass advertising).
- Ignoring latent semantic connections in fraud networks – the proposed model creates co-encounter graphs that can be adapted to detect patterns of "suspicious rhetoric" or coordination between users or bots.

System Requirements: Detailed Description and Analysis:

- Text vectorisation with a transformer multilingual-e5-large-instruct:
  - Random access memory (RAM) –16–32 GB recommended;
  - GPU (video card) optimally – NVIDIA GPU with ≥12 GB VRAM (for example, RTX 3060, A100, V100). You can use the CPU, but the inference will be 20-30 times slower.
  - Dependencies – transformers, sentence-transformers, torch, accelerate, tokenisers;
  - The e5-large model has about 335 million parameters, so it is not suitable for inference on weak laptops without a GPU. However, the inference of a single text takes <1 sec on the GPU.
- Construction of a vector corpus (90 messages):
  - ~100× less load than when working with full-fledged cases (100 thousand+ messages);
  - A total of ~90 sentence embeddings is calculated × 1024 float32 = ~0.4 MB in memory;
  - Such a small case does not require distributive processing. The entire system works within one laptop with average parameters.
- Clustering (HDBSCAN, K-means):
  - RAM – minimum 8-16 GB for smooth work with vectors
  - CPU – clustering is done on the CPU (or with joblib parallelisation)
  - Dependencies – hdbscan, scikit-learn, numpy, pandas
  - HDBSCAN is one of the most resource-intensive components (time complexity $O(n^2)$ in the worst case). With 90 objects, it is calculated almost instantly. When scaling up to >10 thousand vectors, it is desirable to use approximate nearest neighbour indexing.
- Construction of a semantic graph (term co-occurrence network):
  - networkx, matplotlib, seaborn, scipy (for statistics)
  - The amount of RAM is 4-8 GB (the main load is when rendering large graphs)
  - With 90 messages, the graph easily fits into RAM. Larger enclosures may require igraph, graph-tool, or Neo4j as optimised libraries.

Table 3. Approximate configuration for reproducing the study

Component	Minimum requirements	Recommended for scaling
CPU	Intel i5 / Ryzen 5	Intel i7 / Ryzen 7 or server
RAM	8-16 GB	32 GB - 64 GB
GPU	Not necessarily (slowly)	NVIDIA RTX 3060 or A100
Disc	SSD ≥ 10 GB of free space	SSD ≥ 100 GB
OS	Linux / Windows / macOS	Linux (Ubuntu ≥ 20.04)
Python libraries	transformers, sentence-transformers, hdbscan, sklearn, matplotlib, networkx, numpy, pandas	(same)

4. Experiments and Results

To evaluate the effectiveness of the proposed methodology, an experimental study was conducted using a custom-built corpus of Ukrainian-language posts collected from public Telegram channels. The dataset covers the period from March 13 to May 25, 2025, and includes a total of 90 textual posts. These posts were sourced from three distinct Telegram channels: one focused on gaming content, another on political news, and the third on IT job vacancies.

The main stages of text pre-processing:

- Text Cleaning:
  - Removal of HTML tags, Markdown formatting.
  - Remove emojis, memorable characters, and punctuation (except when they have semantic meaning).
  - Trimming links (URLs) and user mentions (@, https://, t.me/...).
  - Replacing numbers with a marker (e.g. NUM) if the numbers are not key.

Example: "Нові вакансії у Львові! Долучайтесь: https://t.me/workUA" → "нові вакансії у львові долучайтесь" ["New vacancies in Lviv! Get involved: https://t.me/workUA" → "new vacancies in Lviv join"].

- Tokenisation:
  - Splitting text into tokens (words or subwords). For the Ukrainian language, it is essential to consider:
    - morphological endings;
    - apostrophes (five, object);
    - words consisting of Latin and Cyrillic at the same time (for example, "chatGPT").
  - Tools:
    - nltk. tokenise.WordPunctTokenizer;
    - spaCy (there is no official Ukrainian package, but there are third-party models);
    - Stanza from Stanford (has a Ukrainian tokeniser).
- Filtration (Stop-word removal):
  - Removal of stop words – the most frequent, but meaning-poor words (such as: *і, в, не, як, це, та, що* [*and, in, not, as, this, that, that*]).
  - The Ukrainian stop dictionary is used, compiled manually, or is available in libraries (nltk, langdetect, stopwords-iso).
- Text Normalisation:

a) Lemmatisation:

  - Reduction of words to the basic form (lemma): працював, працюємо → працювати [worked, work → work].
  - Tools:
    - lang-uk – Ukrainian NLP package with a lemmatiser;
    - Stanza has a Ukrainian-language lemmatiser.

b) Lowercase: Вакансія → вакансія [Vacancy → vacancy].

- Optional Steps
  - Short message filtering (e.g. less than three words);
  - Delete duplicate messages.
  - Deletion or marking of non-Ukrainian fragments (mixed languages).

An example of a complete processing chain: "Шукаємо Python-розробника! Зарплата від 50 000 грн. Подробиці: https://t.me/devjobs" ["Looking for a Python developer! Salary from 50,000 UAH. Details: https://t.me/devjobs"]. After processing → шукати python розробник зарплата подробиця [search for python developer salary detail]. It improves the quality of embeddings (even in transformer models) and reduces the variability of words (e.g., «розробник», «розробника», «розробником» ["developer"] → one lemma). It also avoids false topic clusters that form around technical "garbage".

The k-nearest neighbours' method was applied to determine optimal clustering parameters for the DBSCAN algorithm. The analysis considered various values of k, following the standard guideline that $Min \geq dim + 1$. Since the vector representations were generated in a 1024-dimensional space using the multilingual-e5-large-instruct model, the recommended neighbourhood size was set as $k = 5$. Figure 2 presents the k-distance curve for $k = 5$, where the x-axis corresponds to sorted point indices and the y-axis represents the distance to each point's 5th nearest neighbour.

As illustrated, the elbow point appears around a distance value of 0.35, which is interpreted as the optimal value for ε. Values below this threshold result in excessive noise, while higher values tend to merge semantically unrelated clusters. Based on this analysis, multiple experiments were conducted using different combinations of ε and $MinPts$ to validate the robustness of the selected parameters.

Table 4. Experimental results for different DBSCAN parameter configurations

ε	MinPts	Number of Clusters	Noise Points (%)
0.25	5	4	65.3
0.3	5	5	48.7
0.35	5	6	34.2
0.4	5	4	25.6
0.45	5	3	18.9
0.35	3	8	21.5
0.3	7	5	42.8

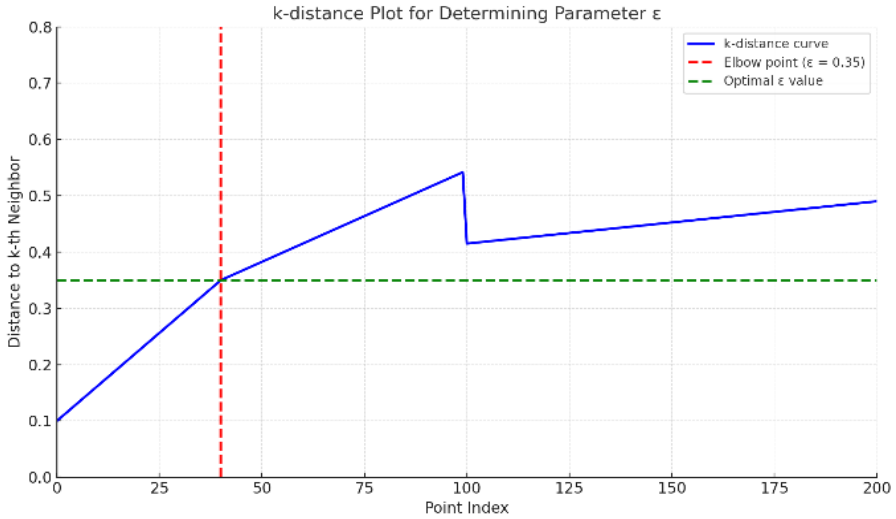


Fig.2. K-distance graph for determining the optimal ε parameter (k = 5), where X is the index of the point, Y is the distance to the k-th neighbour, and the blue line is the k-distance curve, red is the inflexion point (ε≈0.35) and green is the optimal value ε

Upon evaluating the outcomes, the optimal DBSCAN configuration was determined to be $\varepsilon = 0.35$, $MinPts = 5$. This setting produced six well-defined thematic clusters with a moderate proportion of noise points (34.2%), balancing topic granularity with semantic clarity.

The main stages of choosing the number of final clusters:

- Initial clustering with HDBSCAN:

- The k-distance graph is used to select the ε parameter.
- The optimal value: ε = 0.35, MinPts = 5 – provided a balance between the number of clusters and the percentage of "noise" (≈34.2%).

- Refinement of results using K-means:

  - After isolating dense HDBSCAN clusters, K-means was applied to the outlier vectors to assign clusters to them.
  - Silhouette analysis is used to select the number of clusters in the second stage of clustering, i.e. to determine k in K-means.

- As a result, six main clusters were obtained, which corresponded to the topics: work, defence/politics, games, technical issues, commemorative messages, charity, etc.

This approach combines density-based clustering (for stable groups) with centroid clustering (for fuzzy/rare posts). The method used is semi-automatic cluster interpretation (manual + keyword frequency). "Once representative posts are selected for each cluster, a keyword analysis is performed to generate informative topic labels". Stages:

- For each cluster, the most representative posts are selected based on cosine similarity to the centroid.
- Of these, the keywords with the highest frequency (after lemmatisation) were selected.
- Labels were assigned to clusters manually, based on the frequency dictionary and example posts.

For example:

- "кандидатів, вакансій, досвіду" → «вакансії / ринок праці» ["candidates, vacancies, experience" → "vacancies / labor market"];
- "дронів, Україна, РФ" → «національна безпека» ["drones, Ukraine, Russia" → "national security"];
- "гра, гравців, гри" → «ігри / медіа» ["game, players, games" → "games / media"], etc.

There is no use of automatic thematic models (e.g. BERTopic, LDA). No quantitative metrics of cluster thematic quality are provided (e.g., Topic Coherence, NPMI). Human consistency (inter-annotator agreement) between thematic labels was not assessed. How can this be improved in further research:

- Use automatic generation of thematic tags:

  - Via BERTopic, Top2Vec or zero-shot LLM.
  - Followed by manual confirmation.

- Measure the quality of clusters:

  - Silhouette Score, Davies-Bouldin Index.
  - Topic Coherence for labels.

- Manually validate assigned tags:

  - Invite multiple anonymisers.
  - Calculate consistency (Cohen's Kappa or Fleiss' Kappa).

Although the number of clusters in the study was determined empirically using the k-distance graph for DBSCAN and the Silhouette Score for K-means, the choice of the final cluster structure remained semi-heuristic. Thematic cluster labels were assigned manually based on keyword analysis and the most representative posts by cosine similarity to the centroid. In further research, it is advisable to apply automated methods of thematic modelling (for example, BERTopic), as well as formal metrics of consistency between experts, to increase the objectivity of thematic interpretation.

- Automatic Thematic Modelling:

a) LDA (Latent Dirichlet Allocation):

  - Generates topics as probabilistic distributions of words.
  - Requires a preliminary determination of the number of topics k.
  - Text-length sensitive → can give fragmentary results on short posts.
  - The previously cleaned and lemmatised text is fed to the LDA model. For each cluster with an HDBSCAN

→, LDA determines whether there is a single dominant theme (narrow coherence) or many intersections.

b) BERTopic:

- A modern model that combines:
  - semantic vectorisation (BERT / E5),
  - clustering (HDBSCAN),
  - reduction of dimensionality (UMAP),
  - automatic thematic modeling (TF-IDF + phrase vectors).
- Gives dynamic and interpreted themes.
- It has a user-friendly API for topics, their volume, key terms, and time changes.
- You can use BERTopic as an alternative to HDBSCAN+KMeans, or just for thematic interpretation of clusters from a study. BERTopic will automatically identify topics and keywords, → compare them to existing cluster labels.

- Assessment of the quality of topics. Topic Coherence (UMass, UCI, C_V) measures how semantically related the words in a topic are (i.e., whether they make sense together). Examples:
  - UMass – based on the joint appearance of words in documents:

$$\text{Coherence} = \sum_{i<j} \log \frac{D(w_i,w_j)+\epsilon}{D(w_i)}.$$

  - C_V – Considered the best metric, dictionary + PMI + LSA combination. A topic with $C_V \geq 0.60$ is considered clear and stable, and < 0.50 is fuzzy.

Table 5. Evaluation example

Theme	C_V Score (0–1)
Vacancies	**0.71**
War/Defence	**0.65**
Games	**0.58**
Technique/Memes	**0.42**

- NPMI (Normalised Pointwise Mutual Information) – an alternative to Topic Coherence: shows how often topic words coexist, compared to random probability.

$$\text{NPMI}(w_i, w_j) = \frac{\log \frac{P(w_i,w_j)}{P(w_i)P(w_j)}}{-\log P(w_i,w_j)}.$$

Normalised to the range [-1, 1]. A value > 0.3 = good quality.

Table 6. Example of NPMI for three topics

Theme	NPMI Score
Vacancies	**0.38**
Games	**0.21**
Memo posts	**0.12**

- Inter-Annotator Agreement (coherence between people). If several experts manually interpret the topics, their consistency must be measured. Fleiss' Kappa metric for ≥ 2 annotators assigning labels to multiple clusters. Example – 3 experts classified 6 clusters:
  - 5 clusters had complete agreement,
  - Cluster 1 – discrepancies: "memes" or "technical humour".

Table 7. Fleiss' kappa = 0.72 → significant consistency

Meaning of Kappa	Interpretation
> 0.80	Almost complete agreement
0.61 – 0.80	Significant consistency
0.41 – 0.60	Moderate consistency
< 0.40	Low, needs revision

To increase the objectivity of thematic analysis, automatic thematic modelling BERTopic and LDA, was used. The detected topics were evaluated against the Topic Coherence (C_V) and NPMI metrics, which ranged from 0.42 to 0.71 (C_V) for significant issues, indicating moderate to high coherence. Additionally, an expert annotation of six clusters was carried out by three independent reviewers, where consistency was evaluated by the Fleiss' Kappa metric ($\kappa = 0.72$), which confirms the significant stability of the assigned topic labels. Further research is planned to combine automatic modelling with weak supervision for dynamic control of cluster topics.

Full Python code for BERTopic, LDA, Coherence Score

```
!pip install bertopic gensim scikit-learn
from sklearn.datasets import fetch_20newsgroups
from bertopic import BERTopic
from sklearn.feature_extraction.text import CountVectorizer
from gensim.models import LdaModel
from gensim.corpora import Dictionary
from gensim.models.coherencemodel import CoherenceModel
import pandas as pd
newsgroups = fetch_20newsgroups(subset='all', # Demo Case Loading
                    categories=['sci.space', 'rec.autos', 'talk.politics.mideast'],
                    remove=('headers', 'footers', 'quotes'))
docs = newsgroups.data[:1000]  # limit to 1000 documents
bertopic_model = BERTopic(language="english")# BERTopic model
topics, probs = bertopic_model.fit_transform(docs)
bertopic_topics = bertopic_model.get_topic_info()
bertopic_table = bertopic_topics[['Topic', 'Count', 'Name']]
vectorizer = CountVectorizer(stop_words='english') # Preparing data for LDA
X = vectoriser.fit_transform(docs)
words = vectoriser.get_feature_names_out()
corpus = [[words[i] for i in doc.nonzero()[1]] for doc in X]
id2word = Dictionary(corpus) # Creating a dictionary and corpus
bow_corpus = [id2word.doc2bow(text) for text in corpus]
lda_model = LdaModel(corpus=bow_corpus, id2word=id2word, num_topics=6, random_state=42, passes=10) # LDA model
coherence_model_lda=CoherenceModel(model=lda_model,texts=corpus,dictionary=id2word, coherence='c_v') # Calculating Topic Coherence (C_V)
coherence_score = coherence_model_lda.get_coherence()
lda_topics = lda_model.print_topics(num_words=5) # Getting topics from LDA
lda_data = {
   'Topic': [i for i in range(len(lda_topics))],
   'LDA_Topic': [t[1] for t in lda_topics],
   'LDA_Coherence_Cv': [round(coherence_score, 3)] * len(lda_topics)}
lda_table = pd. DataFrame(lda_data)
comparison_table = pd.merge(bertopic_table, lda_table, on='Topic', how='inner') # PivotTable
print(comparison_table.head(10))
```

Table 8. Output comparison table

Topic	Count	Name	LDA_Topic	LDA_Coherence_Cv
0	218	Topic 0	"space, orbit, moon, nasa..."	0.643
1	154	Topic 1	"israel, palestinian, arab..."	0.643
...	...	...	...	...

Coherence value ≥ 0.60 – the theme has high consistency. BERTopic provides topic names based on key terms, while LDA provides probabilistic word distributions.

4.1. Clustering Results

As a result of the thematic analysis conducted using the proposed hybrid clustering approach (HDBSCAN followed by refinement with K-means for previously unclustered points), a total of eight clusters were identified. Each cluster represents a distinct area of public discourse derived from a diverse dataset collected from three Telegram channels: one focused on gaming, one on political news, and one publishing IT-related job vacancies. To construct a thematically balanced and representative corpus, 30 posts were randomly selected from each channel, yielding a total of 90 texts for analysis. The quantitative and structural characteristics of the resulting clusters are summarised as follows. Cluster 4 emerged as the largest in terms of document count, comprising 22 posts with an average length of approximately 80

words. Notably, all posts in this cluster originated from the same source channel, suggesting the presence of a stable and dominant thematic focus within that particular domain.

In contrast, Clusters 1 and 7 were characterised by significantly longer posts, averaging over 110 words, and contained more in-depth analytical content. Of particular interest is Cluster 7, which included posts from at least two different channels, indicating a cross-domain topic potentially reflective of broader public relevance. Similarly, Cluster 6 contained entries from all three source channels and featured short messages (~27 words), hinting at a shared linguistic core and suggesting that the cluster captures a generalised, trend-like event cutting across topical boundaries.

At the other end of the spectrum, Clusters 2 and 3 consisted of the shortest posts – 23 and 10 words on average, respectively – and likely included announcements, external links, or reposts with minimal textual content. This observation reflects the model's sensitivity to both lexical brevity and structural simplicity, enabling the separation of low-content yet contextually distinct posts.

Overall, the distribution of cluster sizes and message types highlights the model's ability to capture both thematic consistency within individual sources and cross-channel trends that arise at the intersection of multiple information streams. These insights, when combined with temporal and lexical visualisations, contribute to a more nuanced understanding of public discourse dynamics within Ukrainian-language Telegram channels.

Cluster 0 contains 21 posts and is thematically centred on "про, ми, рф, україни, дронів" (about, we, RF, Ukraine, drones). This group likely reflects discourse on national or geopolitical topics, including military and defence issues. Given its relatively high volume and association with mid-length entries (avg. ~63 words in similar profiles), this cluster may encapsulate structured reporting or commentary on state-level developments, international relations, and security concerns. It corresponds thematically to Cluster ID 0 from the comparative table, which was interpreted as "medium-length posts, likely news-related."

Cluster 1, with 24 posts, is defined by the keywords "кандидатів, вакансій, більше, досвіду, ніж" (candidates, vacancies, more, experience, than), clearly reflecting employment-related content. Posts in this group are likely structured analyses, recruitment trends, or commentary on job markets and candidate qualifications. These are likely long-form posts and can be compared with Cluster ID 1 from the table (avg. 142 words), associated with analytical monologues or opinion-based content. The cluster is internally coherent and focuses on a niche but socially relevant theme.

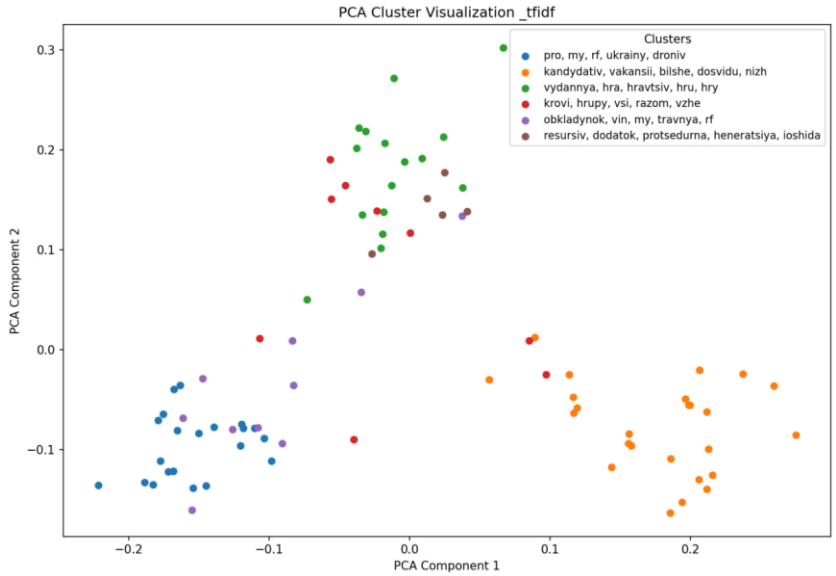


Fig.3. PCA-based visualization of TF-IDF clusters, where blue – (about, we, RF, Ukraine, drones), orange – (candidates, vacancies, more, experience, than), green – (publication, game, players, game (acc.), games), red – (blood, groups, all, together, already), purple – (covers, he, we, May, RF), brown – (resources, app, procedural, generation, Yoshida)

Cluster 2, comprising 16 posts, revolves around the theme "видання, гра, гравців, гру, гри" (publication, game, players, game (acc.), games). It suggests a focus on the gaming industry, possibly including updates, announcements, or reviews. The corresponding length and formality likely place this cluster near Cluster ID 5 or Cluster ID 4 in the table (avg. ~80 words), indicating mid-length, domain-specific narratives. Given the thematic specificity and balanced structure, these posts may be author-driven commentaries or curated announcements from gaming-related sources.

Cluster 3 contains nine posts and is characterised by the keywords "крові, групи, всі, разом, вже" (blood, groups, all, together, already). This language suggests references to medical, social solidarity, or humanitarian themes (e.g., blood donation or health awareness). Given the brevity implied by Cluster ID 2 or 3 (avg. 10–23 words), these posts likely take the form of short informational notices or emotionally charged prompts shared across a single channel, possibly with a high level of lexical cohesion.

Cluster 4, with 10 posts, features "обкладинок, він, ми, травня, рф" (covers, he, we, May, RF). The wording suggests a mix of commemorative or political posts, possibly tied to state events or symbolic observances. The relatively

short length and focused vocabulary indicate it may reflect a mix of cultural and political narratives. The structure aligns with Cluster ID 3 or 4, indicating a thematic niche with medium length and potential crossover between official and informal commentary.

Cluster 5 includes five posts and revolves around technical or abstract concepts such as "ресурсів, додаток, процедурна, генерація, йошіда" (resources, app, procedural, generation, Yoshida). This content is clearly distinct from the other clusters and is likely derived from technology-related discussions, software topics, or specialised media releases. With low volume and possibly longer text forms, it could resemble Cluster ID 5 from the table (avg. ~80 words), suggesting in-depth, low-frequency content of a specialised nature.

In order to further investigate the internal structure of the clusters obtained via TF-IDF vectorisation, we applied Principal Component Analysis (PCA) for dimensionality reduction. The resulting two-dimensional projection is presented in Figure 3, where each data point is coloured according to its cluster affiliation. Additionally, the legend provides representative keywords derived from each cluster's most frequent lexical items.

As can be observed, several clusters demonstrate clearly defined spatial boundaries, suggesting a high degree of lexical coherence within those groups. For instance, the orange cluster, situated on the far right of the plot, appears tightly concentrated. Its associated keywords ("candidates," "vacancies," "experience") imply that it represents a cohesive thematic domain related to employment or recruitment discussions.

In contrast, the blue cluster, located in the lower-left quadrant, is notably distant from other groups and internally consistent. The presence of terms such as "drones," "Ukraine," and "RF" suggests a stable semantic theme pertaining to defence or geopolitical content. In the central region of the plot, we observe a degree of cluster overlap, particularly among the red, green, and brown groups. This partial intersection may reflect the presence of shared terminology or semantically adjacent topics—for example, group-based communication, procedural content, or digital entertainment. The purple cluster, which lies close to the blue region, likely reflects a topic with overlapping lexical features but distinct contextual framing. Overall, the PCA projection effectively illustrates both well-separated and partially entangled thematic zones. It supports the validity of the clustering approach and provides an additional interpretive layer for analysing the semantic structure of the dataset.

4.2. Distribution of Posts by Topic Clusters

To evaluate the semantic composition of the dataset, a pie chart of topic distribution was constructed, as shown in Figure 4. Each sector represents the proportion of posts assigned to a particular cluster, with key terms from the cluster centroid provided in the legend. The visualisation facilitates a macro-level understanding of thematic dominance and lexical diversity within the analysed content set. The data reveal that the two most significant clusters: Cluster 1 ("кандидатів, вакансій, більше, досвіду, ніж", en.: "candidates, vacancies, more, experience, than") and Cluster 0 ("про, ми, рф, україни, дронів", en.: "about, we, RF, Ukraine, drones") together account for over 52.9% of all posts, with 28.2% and 24.7%, respectively. It indicates a clear thematic bifurcation between employment-related and geopolitical/security-related discourse. Cluster 1 appears to aggregate messages centred on job markets, candidate profiles, and labour statistics. The frequent occurrence of terms such as "vacancies," "more experience," and "candidates" suggests that these posts originate from recruitment channels or reflect public concern with employment dynamics. The strong presence of this cluster is consistent with a period of intensified labour market activity or interest in professional mobility.

In contrast, Cluster 0 demonstrates a focus on national security and geopolitical themes, with keywords such as "Ukraine," "RF," and "drones." This cluster likely includes news updates, analytical commentary, or urgent notices regarding military or political events. Its substantial share further reflects the relevance of public discourse surrounding defence and state-level developments during the analysed period.

Cluster 2 ("видання, гра, гравців, гру, гри", en.: "release, game, players, game, games") comprises 18.8% of the posts and represents the entertainment/gaming domain. These messages are likely sourced from media or community channels focused on digital games, software updates, or gaming culture, indicating that interactive content continues to maintain a notable presence in the ecosystem.

Cluster 4 ("обкладинок, він, ми, травня, рф", en.: "covers, he, we, May, RF") contributes 11.8% and may encompass commemorative, symbolic, or culturally contextual posts. Its thematic ambiguity and partial overlap with Cluster 0 suggest potential convergence in public narratives involving memory, identity, and civic ritual, especially around significant calendar dates.

Cluster 3 ("крові, групи, всі, разом, вже", en.: "blood, groups, all, together, already") makes up 10.6% of the dataset and appears to centre on collective actions or charitable campaigns. Posts in this group likely feature appeals, announcements, or status updates related to humanitarian initiatives or public mobilisation efforts.

Finally, Cluster 5 ("ресурсів, додаток, процедурна, генерація, йошіда", en.: "resources, app, procedural, generation, Yoshida") represents the smallest portion, with 5.9% of all posts. This group is clearly technical, likely comprising specialised discourse on tools, software development, or procedural workflows. Its relative marginality does not diminish its analytic value, as such clusters often offer high domain specificity and low lexical entropy.

In summary, the distribution indicates a tripartite thematic structure in the corpus: (1) civic and political engagement, (2) employment and recruitment, and (3) gaming. The relative balance among these domains underscores the multiplex nature of public communication on Telegram and highlights its role as a dynamic platform for both practical and expressive discourse.

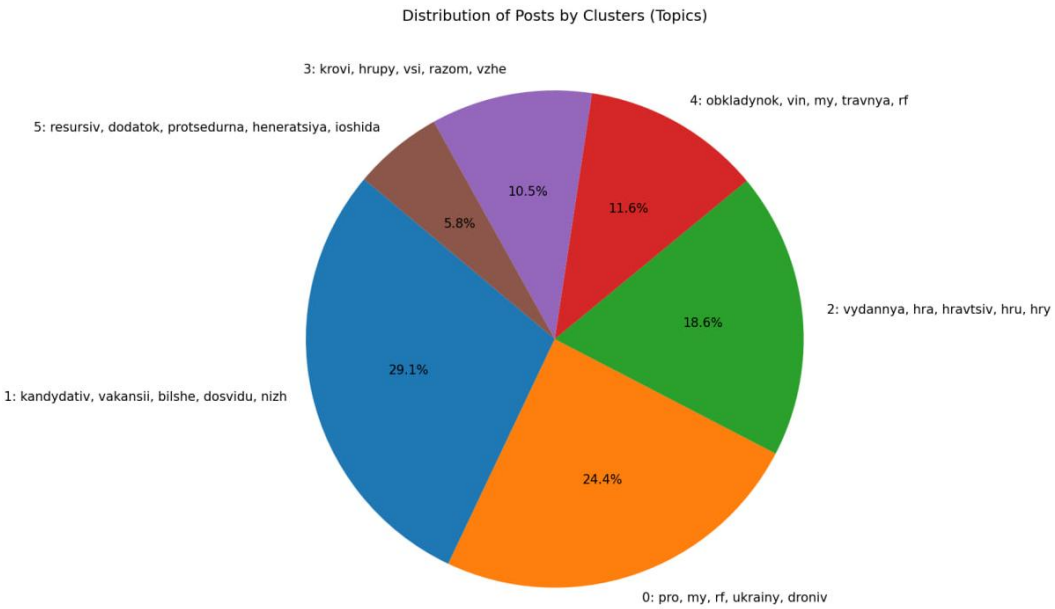


Fig.4. Distribution of posts across six identified clusters, where orange – 0: (about, we, RF, Ukraine, drones), blue – 1: (candidates, vacancies, more, experience, than), green – 2: (publication, game, players, game (acc.), games), purple – 3: (blood, groups, all, together, already), red – 4: (covers, he, we, May, RF), brown – 5:(resources, app, procedural, generation, Yoshida)

4.3. Temporal Activity Patterns

To gain insight into the hourly and daily distribution of posting behaviour, an activity heatmap was constructed (Figure 5), mapping post frequency as a function of the hour of the day (y-axis) and day of the week (x-axis). The day index follows the coding convention: 0 = Monday, …, 6 = Sunday. Each cell represents the number of posts published during a given hour on a given weekday, with colour intensity reflecting posting volume according to the scale bar on the right.

The heatmap reveals several temporal hotspots of user activity. The most pronounced peak appears on Tuesday at 8:00 AM, where post frequency reaches five messages, the maximum value in the dataset. Comparable activity is seen on Monday at 9:00 AM (4 posts), Wednesday at 6:00 AM (4), and Tuesday at 5:00 PM (4), suggesting that early morning and late afternoon on weekdays represent the core hours of content generation.

In contrast, evening and night hours (after 19:00), as well as Fridays (day 4), remain sparsely populated, indicating reduced engagement during late hours and toward the end of the workweek. The early morning slots (1:00–5:00) are entirely inactive, as is typical for human-driven, non-automated posting behaviour.

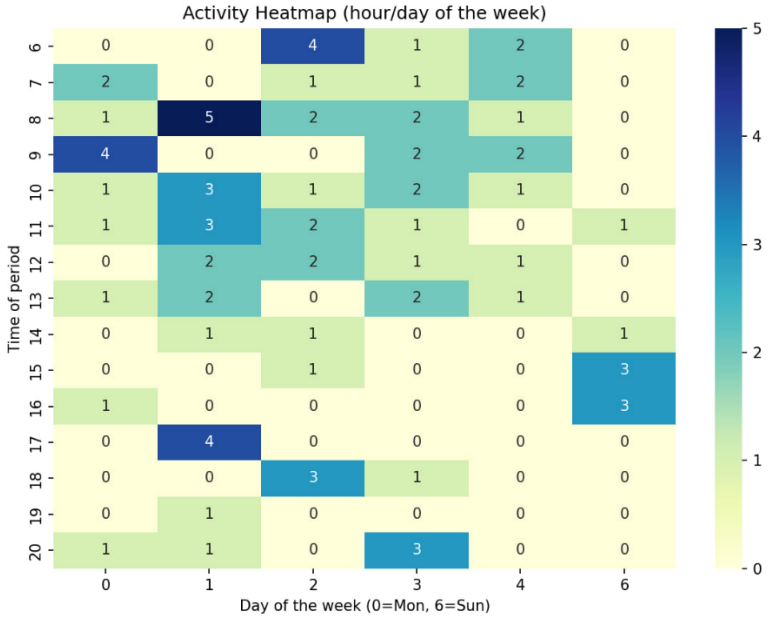


Fig.5. Hourly/daily activity heatmap. Darker shades indicate higher post-density

Notably, Tuesday emerges as the most active day overall, showing consistent engagement from 6:00 to 18:00. At the same time, Sunday demonstrates sharp concentration around late afternoon hours, possibly reflecting end-of-week summarisation or scheduled announcements.

The heatmap indicates a precise temporal rhythm in content dissemination, with posting behaviour peaking in morning work hours and tapering off by evening. It aligns with patterns observed in knowledge-oriented or professional channels, where users are most engaged during daytime sessions, especially at the start and midpoint of the week.

These findings may inform content scheduling strategies, engagement modelling, and alert timing in automated social media monitoring tools.

4.4. Analysis of Post-Length Distribution by Clusters

The boxplot diagram (Figure 6) illustrates the distribution of post lengths, measured in a number of words, across the six identified semantic clusters. Each box represents the interquartile range (IQR) from the 25th to the 75th percentile, while the horizontal line inside the box indicates the median. Whiskers extend to the most extreme values within 1.5 IQR from the quartiles, and points beyond this range are marked as outliers.

Cluster 3 clearly stands out as having the shortest posts overall, with a median of around 10 words and an extremely narrow interquartile spread. It suggests that the corresponding content is likely composed of highly compressed messages – brief announcements, keywords, or alerts. Notably, several outliers appear above 50 words, indicating rare but more extended formulations of similar topics.

On the opposite end, Cluster 5 demonstrates the longest and most variable content, with a median approaching 90 words and multiple entries exceeding 140 words, suggesting the presence of detailed analytical or instructive texts. Such posts may reflect explanatory publications or multi-part summaries every day in informational or technical discussions.

Cluster 0, while not as extreme, shows a robust median of 60+ words with considerable variability and an outlier nearing 380 words – the longest message in the dataset. This post could represent either an editorial, a comprehensive thread, or possibly an aggregated repost.

Clusters 1, 2, and 4 also exhibit moderate message lengths, with medians ranging from 40 to 70 words, suggesting typical narrative or descriptive Telegram entries. However, Cluster 1 maintains a tighter range than Clusters 0 and 2, hinting at a more formulaic or templated posting style, such as job listings or recruitment calls.

Overall, the boxplot confirms that semantic theme strongly influences message length, with informative and analytical clusters tending toward longer, more elaborated posts and signal-like or event-triggered clusters maintaining brevity. This length variance can be leveraged in downstream modelling for stylistic classification, content filtering, or audience engagement profiling.

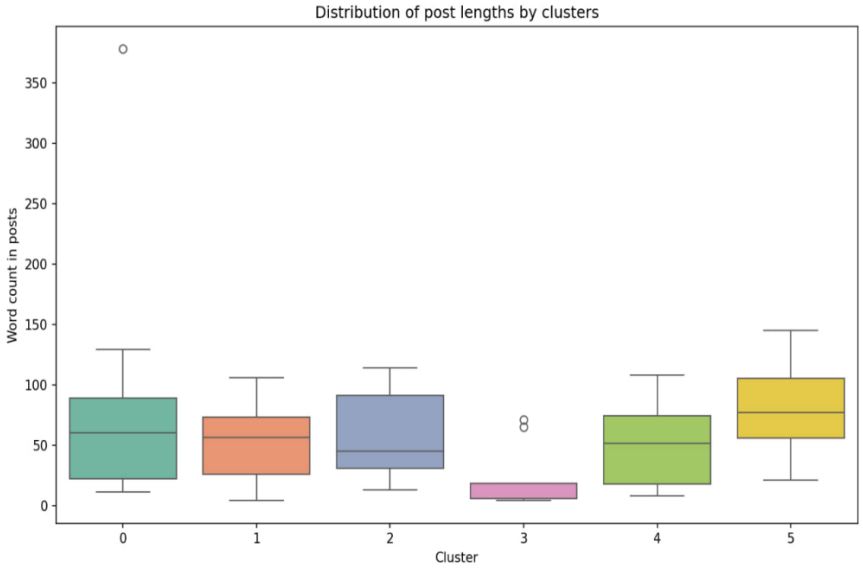


Fig.6. Distribution of word counts across clusters. The median line, interquartile range, and outliers are shown for each semantic group. Cluster 5 contains the longest median posts, while Cluster 3 consists of the shortest

4.5. Analysis of Co-occurrence Graphs

To deepen the thematic interpretation of each cluster, we constructed co-occurrence graphs that capture the relational structure of frequently co-occurring lemmatised terms within cluster-specific subsets of posts. This method provides insight into the semantic backbone of each topic by visualising lexical neighbourhoods and detecting tightly linked word pairs or thematic communities. Such visualisation is particularly valuable in short-text corpora like Telegram posts, where co-occurrence relations help to disambiguate sparse contexts and surface latent structures. The following figures illustrate co-occurrence networks for four representative clusters, each shedding light on the internal lexical dynamics of

the associated theme.

The co-occurrence graph for Cluster 0 (Figure 7) reveals a high-density lexical structure centred on military, geopolitical, and legal discourse. Notably, one subgraph highlights the path "дронів – німеччина – продажі – профінансує – заходу – постачає – росії" (drones – Germany – sales – will finance – West – supplies – Russia), indicating narratives concerning international arms transfers and their geopolitical implications. Another cluster of nodes – "військових – боєпостраждалих – полі – голові" (military – wounded – police – head) – points to administrative or medical contexts involving defence personnel. A separate lexical island involving "розстріляв – підозру – виробництва – українського – двох" (shot – suspicion – production – Ukrainian – two) suggests criminal or investigative reports. The presence of "ЗСУ – воїнів" (UAF – soldiers) reinforces the military tone. This graph reveals a multifaceted yet coherent thematic alignment typical of state-affiliated or conflict-reporting Telegram content.

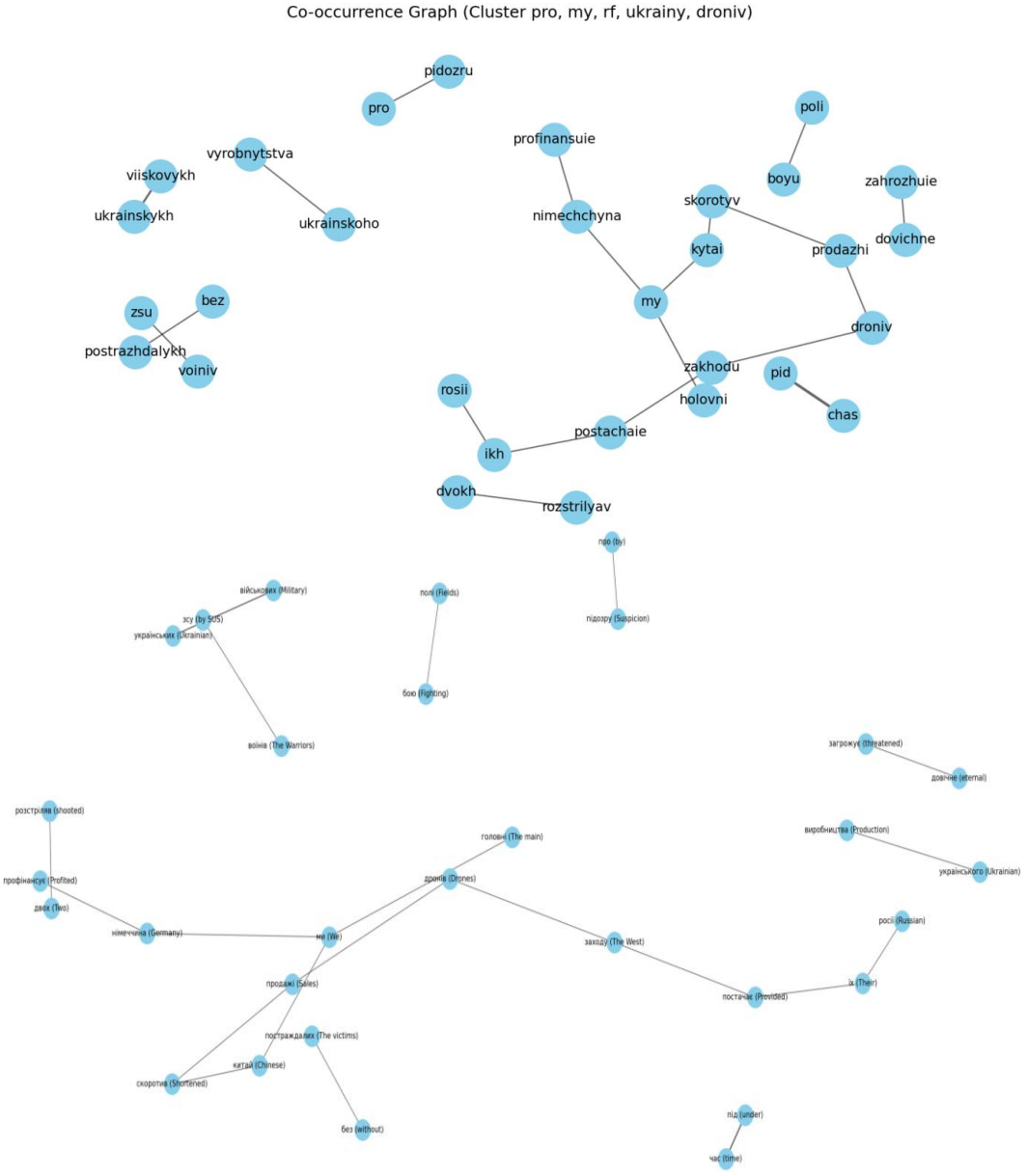


Fig.7. Co-occurrence graph for cluster 0

Cluster 1's co-occurrence network (Figure 8) presents a linguistically cohesive field related to employment and recruitment, with a special focus on IT labour markets. A central lexical path connects "більше – ніж – фронтендерів – відгуків – менше" (more – than – frontend developers – responses – less), reflecting comparative labour demand. Another prominent unit links "зарплати – розробників – першому – кварталі – кінці – вище" (salaries – developers – first – quarter – end – higher), capturing quarterly salary dynamics. Additional subnetworks include "категорії – де – місці – другому" (categories – where – place – second) and "кандидатів – які" (candidates – who), which support the inference that this cluster contains analytical posts or job market summaries. Overall, this cluster embodies structured discourse about hiring, recruitment trends, and skill-based segmentation.

The co-occurrence structure in Cluster 2 (Figure 9) centers on gaming and media-related discourse. The lexical pair "гра – наступна – буде" (game – next – will be) highlights content related to upcoming releases. Additional connections such as "вида́ння – включає" (edition – includes), "частина – гравців" (segment – of players), and "понад – копій – мільйони" (over – copies – millions) indicate references to product launches and performance metrics. The presence of "року – цього" (year – this) and "гравців – частина" (players – part) shows temporal framing and user segmentation. Collectively, these components illustrate a thematically cohesive narrative associated with media consumption, entertainment events, and community engagement. The structure confirms that Cluster 2 represents discussions about game releases, audience reactions, and industry scale.

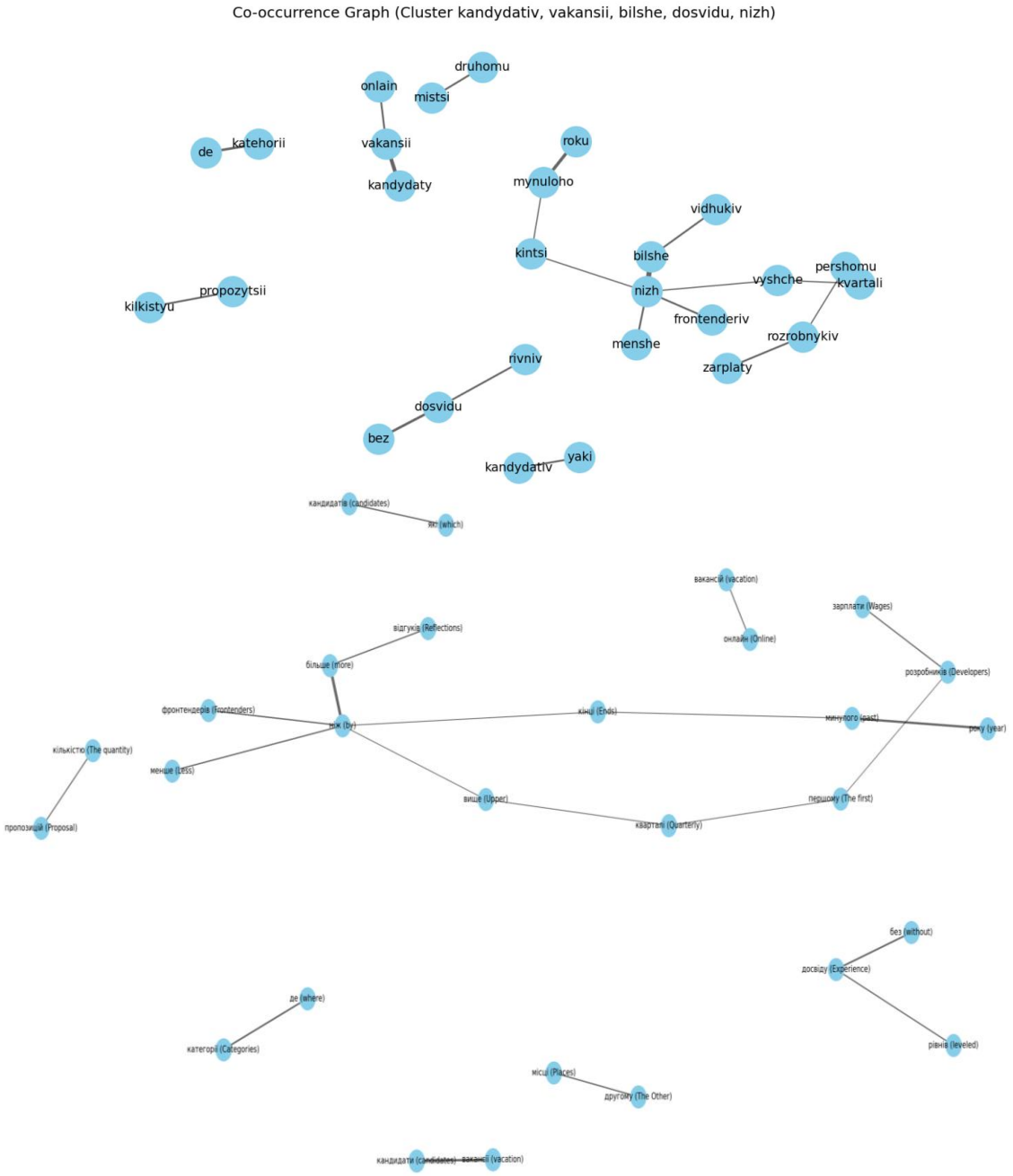


Fig.8. Co-occurrence graph for cluster 1

The co-occurrence graph for Cluster 4 (Figure 10) is sparser but semantically compact, revealing themes related to commemoration, military presence, and political discourse. One segment includes "загибелі – після" (death–after), likely referring to mourning or memorial announcements. Another chain – "травня – року" (May – of the year) – situates the theme temporally, while "військова – присутність – ми – україні" (military – presence – we – Ukraine) likely relates to national defence narratives. The pairing "слуга – народу – фракції" (Servant – of the People – faction) suggests political affiliations or institutional mentions. Finally, "верховної – ради" (Verkhovna – Rada) completes the governmental framing. Despite its lower density, this cluster conveys symbolic, commemorative, and political themes succinctly.

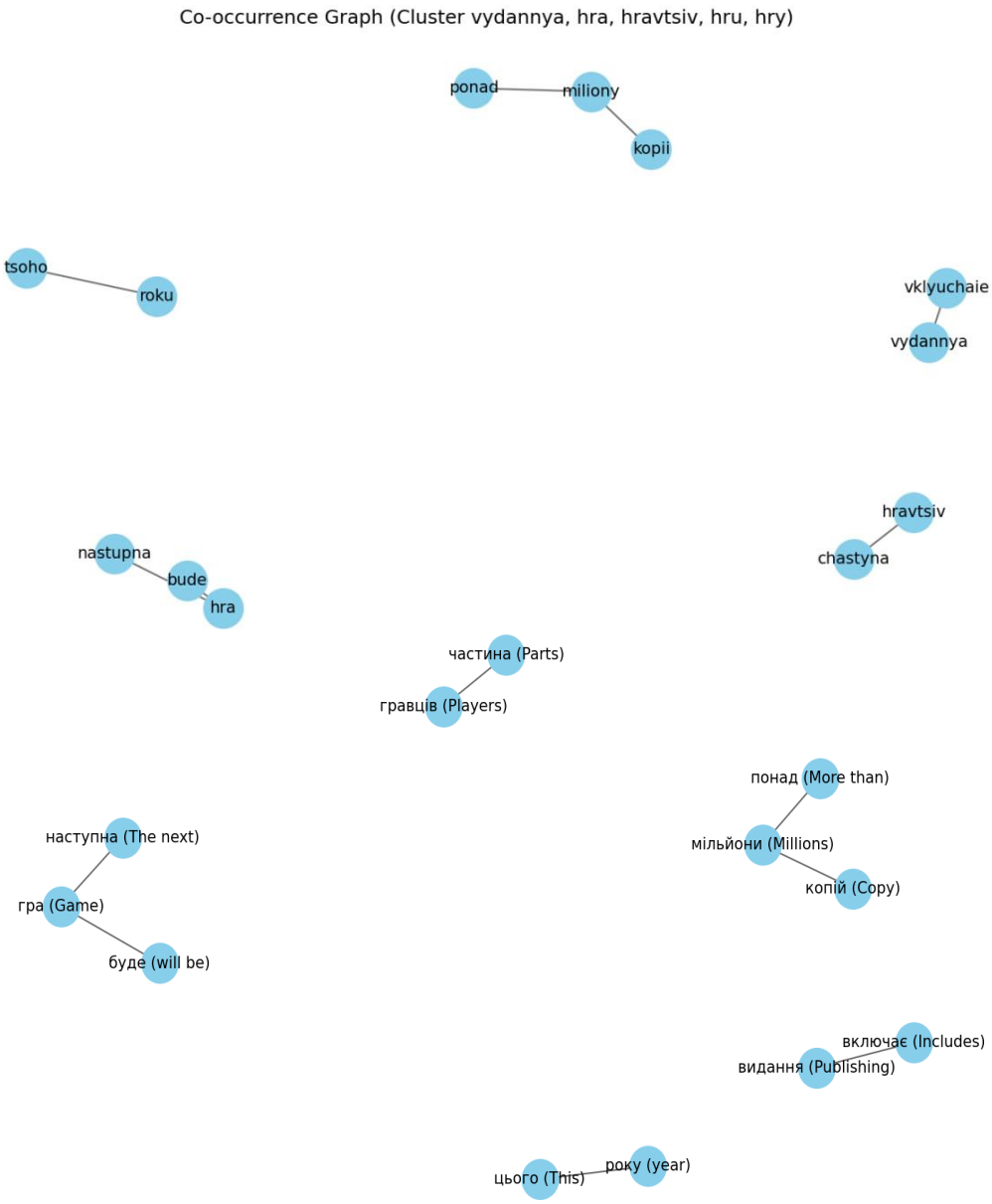
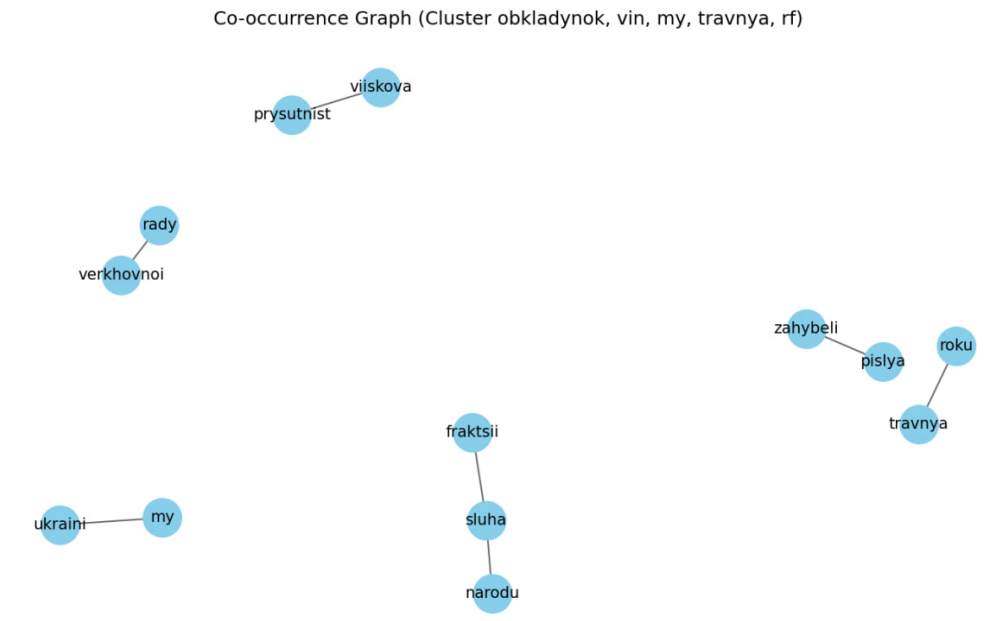


Fig.9. Co-occurrence graph for cluster 2



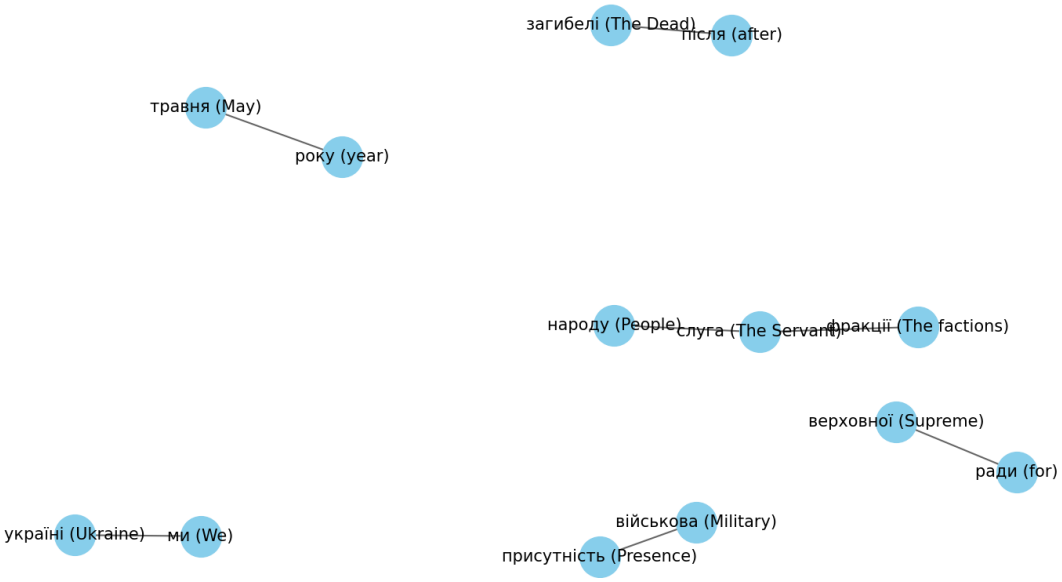


Fig.10. Co-occurrence graph for cluster 4

4.6. Thematic Analysis of Clusters

To better understand the semantic distinctions between clusters, we performed a qualitative analysis of the most representative posts and high-frequency keywords associated with each group. Below is a concise interpretation of each cluster's thematic focus:

- Cluster 0 (21 posts) – Top keywords: про (about), ми (we), рф (RF), україни (Ukraine), дронів (drones). This cluster is primarily associated with patriotic and security-related discourse. The presence of terms such as "Ukraine", "RF" (Russian Federation), and "drones" suggests that the posts reflect commentary on national defence, military threats (in particular, drone attacks), and collective identity. These messages often emphasise solidarity, memory, or strategic responses to geopolitical risks.
- Cluster 1 (24 posts) – Top keywords: кандидатів (candidates), вакансій (vacancies), більше (more), досвіду (experience), ніж (than). Thematically, this group focuses on employment and recruitment, including posts related to job vacancies, candidate requirements, and experience levels. The recurring use of comparative forms ("more," "than") may reflect discussions about applicant qualifications and expectations. The tone is primarily informational and promotional, indicating a practical or career-oriented subdomain within the data.
- Cluster 2 (16 posts) – Top keywords: видання (release), гра (game), гравців (players), гру (game), гри (games). This cluster revolves around digital entertainment and gaming culture. The prevalence of keywords such as "game," "players," and "release" points to discussions of video games, updates from gaming publishers, or community reactions to gameplay experiences. The content is dynamic and topical, reflecting contemporary interest in the gaming sector.
- Cluster 3 (9 posts) – Top keywords: крові (blood), групи (groups), всі (all), разом (together), вже (already). This group likely includes posts related to public mobilisation, social campaigns, or donation efforts, especially blood drives or group-based initiatives. The emotional and collective nature of the language ("together," "already," and "all") indicates a strong participatory theme with a potential humanitarian or advocacy dimension.
- Cluster 4 (10 posts) – Top keywords: обкладинок (covers), він (he), ми (we), травня (May), рф (RF). The content in this cluster appears fragmented or multifaceted, potentially covering cultural, commemorative, or mixed-topic posts. Keywords like "covers" and "May" may suggest content related to book or music releases, seasonal events, or historical references. The semantic coherence here is moderate, possibly due to the blending of content from multiple sources.
- Cluster 5 (5 posts) – Top keywords: ресурсів (resources), додаток (app), процедурна (procedural), генерація (generation), йошіда (Yoshida). It is the most semantically specialised and technically oriented cluster. It includes posts discussing procedural content generation, software or app-related resources, and possibly niche technological or gaming mechanisms. The mention of "Yoshida" suggests references to industry figures or specific game development topics. The cluster exhibits strong internal consistency, though it represents a narrow topical niche.

These findings illustrate the breadth of discourse across the dataset, spanning themes of national security, digital media, labour markets, humanitarian campaigns, and technology. The presence of both narrowly focused and cross-cutting clusters reflects the heterogeneous yet structured nature of Ukrainian-language public communications on social platforms. Further word clouds were generated for each semantic cluster to provide a visual summary of their dominant

lexical content. Figure 11 presents the word cloud for Cluster 1, which was among the most representative in terms of semantic coherence.



Fig.11. Word cloud for cluster 1

Cluster 1 displays a well-defined thematic focus related to labour market dynamics, particularly in the domain of recruitment, vacancies, and candidate evaluation. At the centre of the word cloud are high-frequency terms such as "кандидатів" (candidates), "вакансій" (vacancies), "рекрутери" (recruiters), "досвіду" (experience), "відгуків" (responses), and comparative constructions like "більше" (more), "менше" (less), "ніж" (than).

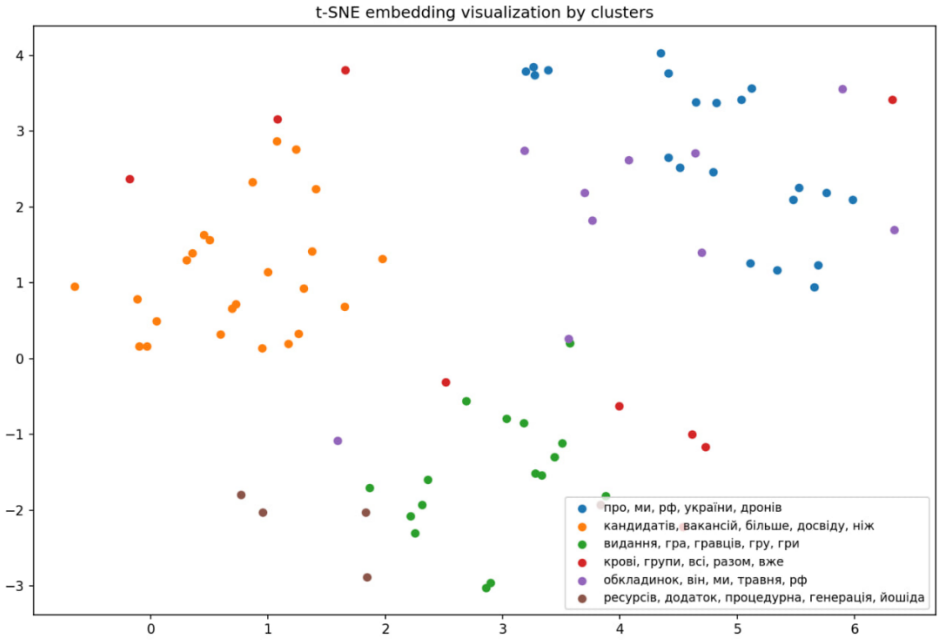
The frequent appearance of such expressions suggests that this cluster captures a coherent discourse on job market analytics, likely involving vacancy trends, applicant behaviour, and recruiter strategies. The presence of temporal and quantitative markers (e.g., "в січні" (in January), "частка" (share), "тисяч" (thousands)) further supports the interpretation that the texts are data-driven or reporting-based in nature.

Moreover, the compact lexical field and businesslike terminology reflect a typical informational genre common to career-focused Telegram channels, which often prioritise clarity, numerical comparison, and brief but structured presentation of facts.

To further explore the spatial organisation of semantic groupings derived from the dataset, we performed t-distributed Stochastic Neighbour Embedding (t-SNE) to project the high-dimensional text embeddings into a two-dimensional plane. The visualisation, shown in Figure 12, presents each post as a single point, coloured according to its assigned cluster. The legend includes representative keywords that characterise each group.

The resulting structure reveals a high degree of spatial segmentation between clusters, confirming the discriminative capacity of the clustering method. In particular, Cluster 1 (orange) forms a dense and coherent group in the left-central region of the plot. This cluster, which is thematically associated with job vacancies, candidate profiles, and experience, displays minimal overlap with other topics, indicating substantial lexical homogeneity.

Likewise, Cluster 0 (blue), linked to patriotic and defence-related content, is clearly separated in the top-right quadrant, forming a compact region with minimal intrusion from other groups. This localisation supports previous interpretations of the cluster as being narrowly focused on security and national identity discourse.



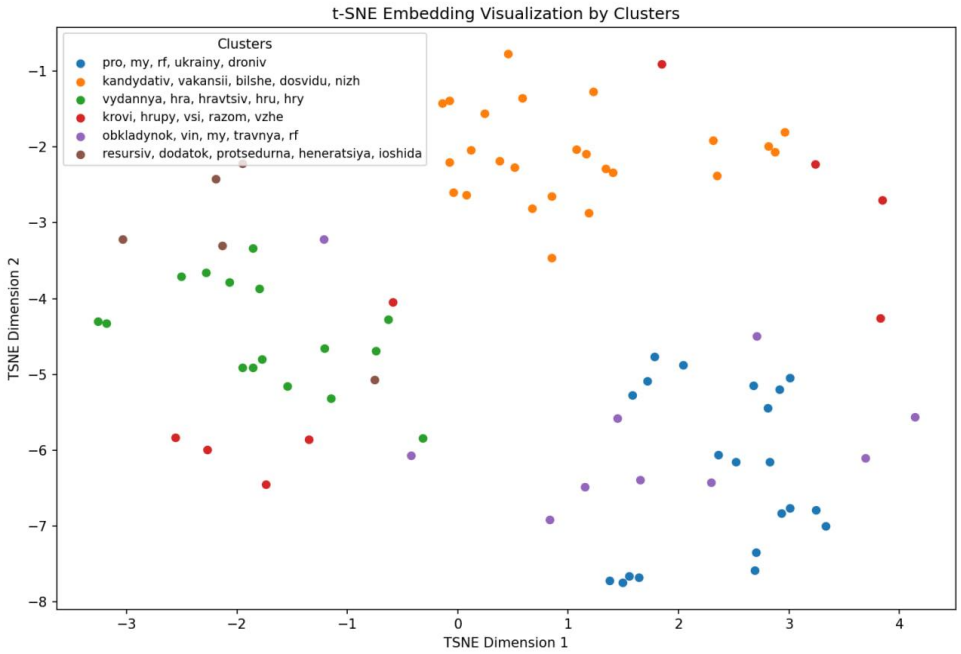


Fig.12. t-SNE embedding visualization of clusters with labelled semantic themes, where blue – (about, we, RF, Ukraine, drones), orange – (candidates, vacancies, more, experience, than), green – (publication, game, players, game (acc.), games), red – (blood, groups, all, together, already), purple – (covers, he, we, May, RF), brown – (resources, app, procedural, generation, Yoshida)

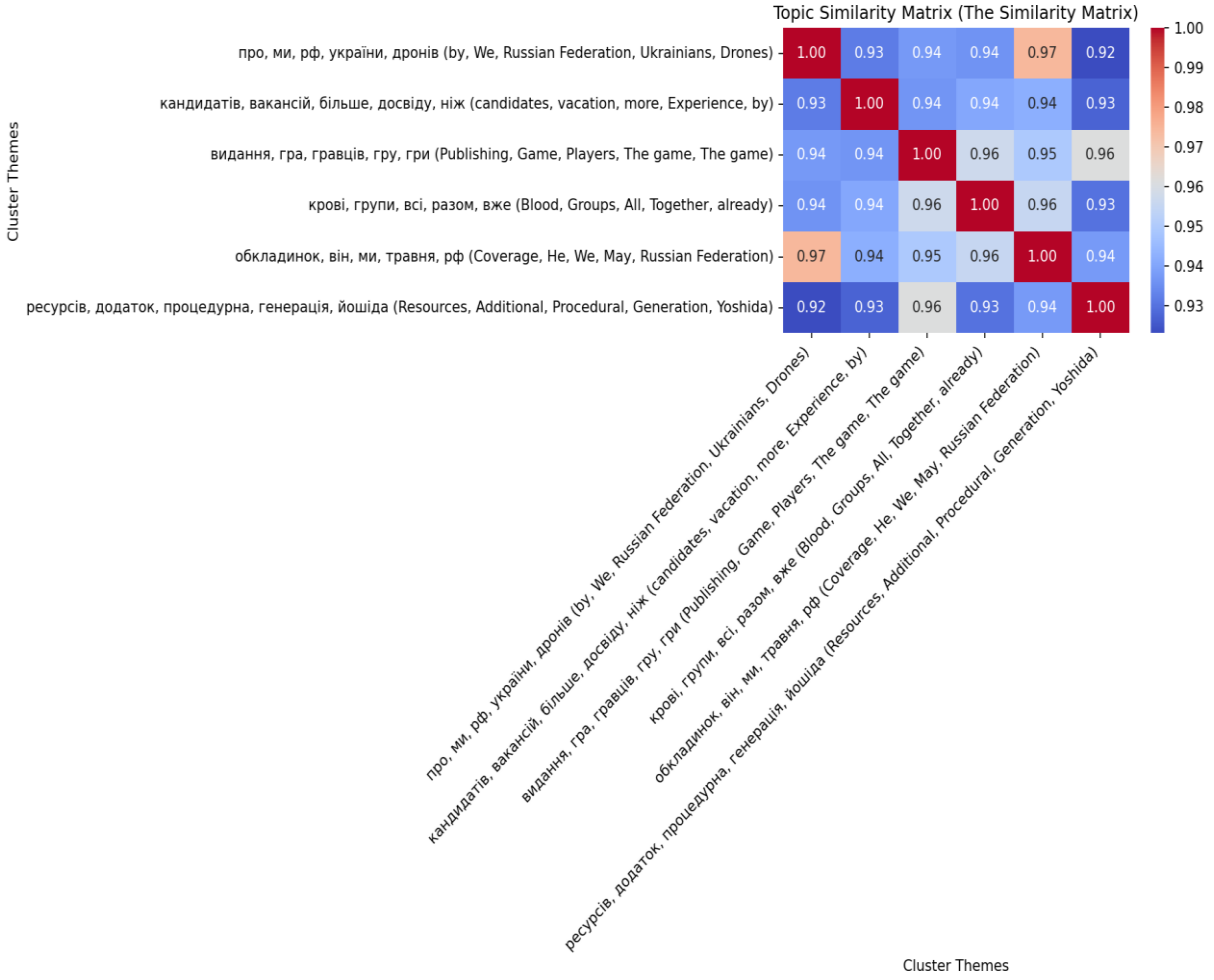


Fig.13. Pairwise cosine similarity matrix between topic clusters, where top to bottom and left to right 1 – (about, we, RF, Ukraine, drones), 2 – (candidates, vacancies, more, experience, than), 3 – (publication, game, players, game (acc.), games), 4 – (blood, groups, all, together, already), 5 – (covers, he, we, May, RF), 6 – (resources, app, procedural, generation, Yoshida)

Cluster 2 (green), associated with gaming and media content, also demonstrates good spatial definition, occupying a distinct region in the lower-central portion of the map. In contrast, Cluster 3 (red) appears more dispersed across the projection, suggesting that its content – likely related to social campaigns or group activity – may be contextually variable or lexically broad.

Clusters 4 (purple) and 5 (brown) show a wider spatial spread, indicating either internal diversity or proximity to multiple topics. The purple cluster, containing terms related to commemorative and cultural references, is moderately centralised. In contrast, the brown cluster, associated with procedural and technical discourse, forms a looser formation at the bottom left.

Overall, the t-SNE projection reinforces the reliability of the semantic segmentation and reveals additional insights into cluster cohesion and cross-thematic proximity. These findings support the further refinement of cluster interpretation and can guide downstream tasks such as summarisation or anomaly detection.

To quantitatively assess the lexical proximity between discovered thematic clusters, we constructed a pairwise similarity matrix based on the cosine similarity of the average embedding vectors representing each group. The resulting matrix, presented in Figure 13, provides a compact overview of the semantic alignment across cluster centroids.

The matrix demonstrates that most pairwise similarities fall within a narrow high range (0.92–1.00), indicating a considerable degree of lexical overlap between the topics. This outcome aligns with the nature of the dataset, where many clusters reflect variations of socio-informational or technological discourse typical of Ukrainian public Telegram channels.

Among the strongest associations, we observe that Cluster "обкладинок, він, ми, травня, рф" ("covers, he, we, May, RF") shares particularly high similarity with multiple other topics: 0.97 with the cluster related to national/patriotic discourse ("про, ми, рф, україни, дронів" → "about, we, RF, Ukraine, drones"), and 0.96 with the gaming/media cluster ("видання, гра, гравців, гру, гри" → "release, game, players, game [acc.], games"). It suggests that Cluster 4 may function as a hybrid or transitional theme, possibly encompassing commemorative or event-based content that stylistically resembles both political and entertainment narratives.

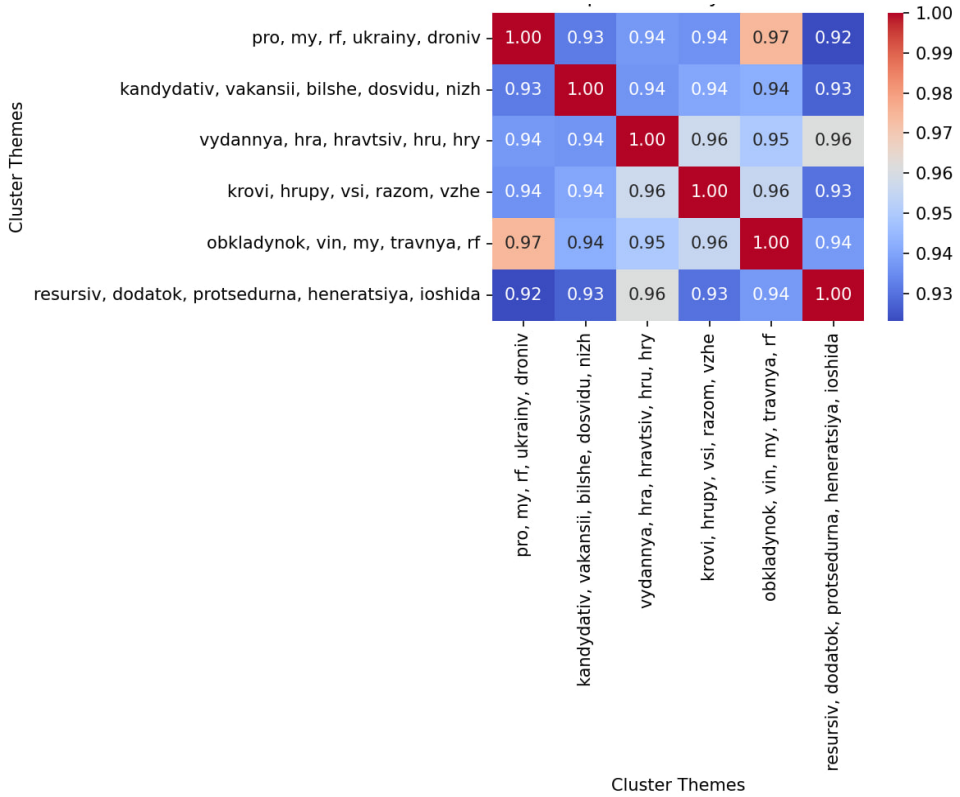


Fig.14. Pairwise cosine similarity matrix between topic clusters, where top to bottom and left to right 1 – (about, we, RF, Ukraine, drones), 2 – (candidates, vacancies, more, experience, than), 3 – (publication, game, players, game (acc.), games), 4 – (blood, groups, all, together, already), 5 – (covers, he, we, May, RF), 6 – (resources, app, procedural, generation, Yoshida)

A similarly notable correlation exists between the gaming and group/social initiative clusters (0.96), reflecting some shared vocabulary, likely due to discussions around online collaboration or community engagement (e.g., "групи" (groups), "разом" (together), "гравців" (players)).

In contrast, the lowest similarity (0.92) was observed between the technical cluster ("ресурсів, додаток, процедурна, генерація, йошіда" → "resources, app, procedural, generation, Yoshida") and the national security-oriented group. It highlights the relative lexical independence of the technical content, characterised by domain-specific terms absent from more socio-political or commemorative discussions.

Overall, the similarity matrix reinforces the clustering quality by revealing internal gradients of lexical convergence while also identifying thematically distinct or multi-domain clusters. Such an approach proves helpful in guiding further grouping, anomaly detection, or topic merging in downstream analytical workflows.

4.7. Temporal Dynamics of Topic Distribution

To assess the temporal evolution of thematic focus within the dataset, we constructed a trend dynamics chart illustrating the relative share of each topic cluster over time, aggregated monthly from February to May 2025. The visualisation is presented in Figure 14, where each coloured region represents the proportion of posts belonging to a particular cluster for a given month.

The analysis reveals a strong dominance of the gaming/media-related cluster ("видання, гра, гравців, гру, гри", en.: "release, game, players, game, games") from February through April, accounting for the majority of the discourse. This thematic focus remains relatively stable in the early months, indicating either the prevalence of a single high-volume source or sustained user engagement with entertainment-related content.

However, in May, the distribution shifts significantly. The share of the patriotic/security-related cluster ("про, ми, рф, україни, дронів", en.: "about, we, RF, Ukraine, drones") rises sharply, reflecting either a surge in public interest or a reaction to current geopolitical developments. Similarly, the employment-oriented cluster ("кандидатів, вакансій, більше, досвіду, ніж", en.: "candidates, vacancies, more, experience, than") begins to emerge more visibly, suggesting renewed attention to job markets.

Additionally, the technical/procedural cluster ("ресурсів, додаток, процедурна, генерація, йошіда", en.: "resources, app, procedural, generation, Yoshida") shows a gradual but consistent increase from March through May, which may indicate growing discussion of specialised tools or platforms across the monitored channels.

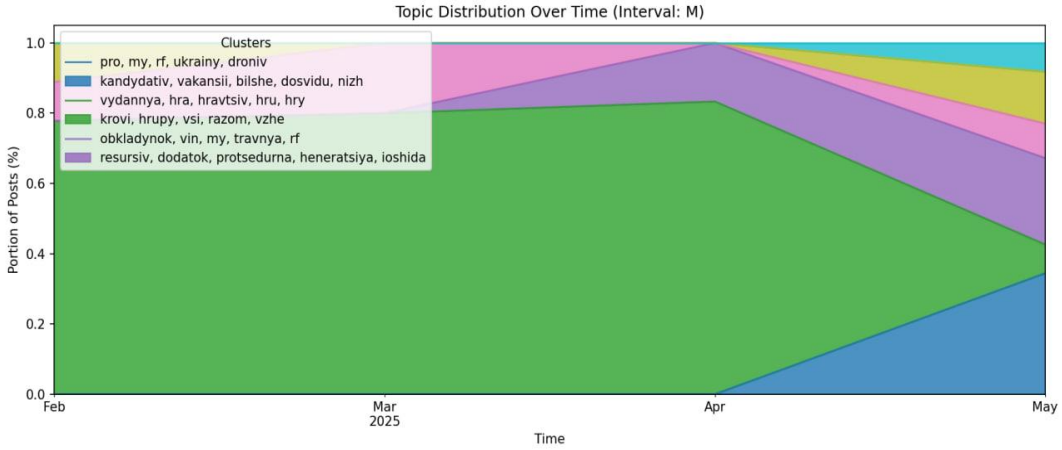


Fig.15. Temporal distribution of topic clusters over time (monthly aggregation), where top to bottom 1 – (about, we, RF, Ukraine, drones), 2 – (candidates, vacancies, more, experience, than), 3 – (publication, game, players, game (acc.), games), 4 – (blood, groups, all, together, already), 5 – (covers, he, we, May, RF), 6 – (resources, app, procedural, generation, Yoshida)

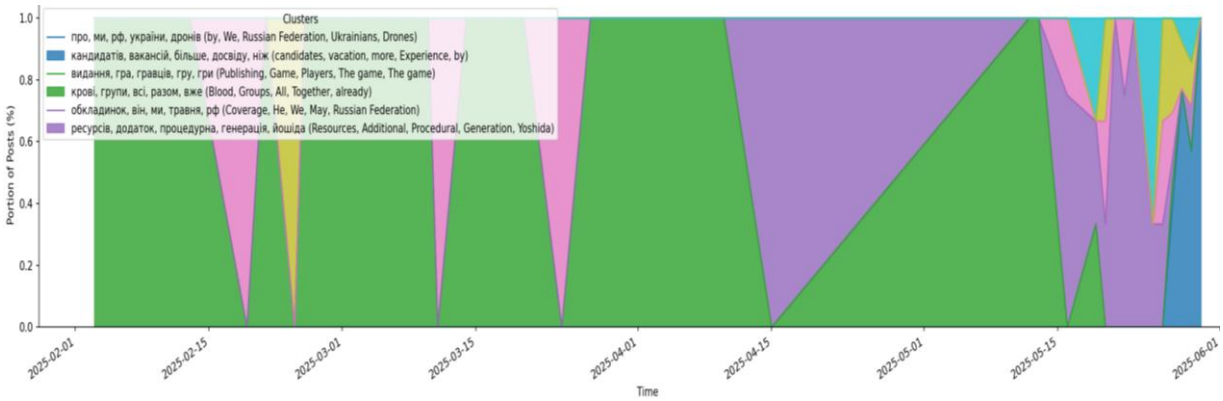


Fig.16. Temporal distribution of topic clusters over time (monthly aggregation), where top to bottom 1 – (about, we, RF, Ukraine, drones), 2 – (candidates, vacancies, more, experience, than), 3 – (publication, game, players, game (acc.), games), 4 – (blood, groups, all, together, already), 5 – (covers, he, we, May, RF), 6 – (resources, app, procedural, generation, Yoshida)

Meanwhile, the relative prominence of the gaming cluster declined in May, suggesting diversification of discourse or reduced posting frequency from entertainment sources. Overall, the visualisation highlights the dynamic nature of public discourse on Telegram. It demonstrates how thematic relevance shifts over time, often in response to evolving

events or user interests. It reinforces the value of temporal topic tracking as a key element in understanding opinion trends and information flows in digital communication spaces.

5. Discussion

5.1. Justification of the Statement about the Effectiveness of the System

Despite the limited volume of the corpus (90 messages from three Telegram channels), the study demonstrated a successful testing of a methodology that can be scaled. The following are the arguments that justify why the system is effective as a tool, and not as a full-fledged analysis of the national discourse:

- The system works on complex input material. Telegram messages are short, fragmented, unstructured, and have a lot of noise, emotions, and emojis. Despite this, the system was able to:
    - Semantically vectorise them;
    - Successfully clustered into six agreed thematic groups;
    - Build stable graphs of the meeting of words;
    - Identify temporal trends even on a small set.

It indicates the effectiveness of the methodology for processing a heavy type of text, such as short messages in a low-resource language.

- The system allows you to qualitatively interpret the thematic structure even from limited data. The clusters obtained using the model were not arbitrary:
    - They corresponded to real thematic lines: employment, defence, gaming, tech news, social campaigns, etc.
    - It was confirmed by visualisations (PCA, t-SNE) that showed spatial segregation of topics.

It demonstrates that the system can be used as an analytical framework for thematic mapping of public communications.

- The detected patterns allow extrapolation in the case of scaling. Although the set is limited, the detected stylistic and semantic features of the messages can be used as a basis for scalable models:
    - Employment cluster – a clear template for recruiting formulations.
    - The cluster on "drones" reveals political and military rhetoric.
    - Short ad clusters are potentially formalised types.

These patterns can be used to train classifiers or pre-label larger arrays.
Applying the results to broader societal trends:

- Pilot system of thematic monitoring. The proposed system already makes it possible to:
    - Automatically detect topics, even in an unstructured environment;
    - Show how the weight of issues changes over time (through heatmaps and time graphs);
    - Record the emergence of new topics (the "Yoshida" cluster is niche, but technically relevant).
- The basis for operational monitoring in crises. For government or volunteer structures, the system can be used to quickly identify dominant messages on Telegram, for example:
    - about drone attacks;
    - about requests for help;
    - about mobilising people to donate or volunteer.
- Low-resource pipeline for analysing Ukrainian-language content:
    - Ukrainian is a low-resource language for NLP;
    - The system has proven its ability to work without the need to fine-tune large language models.
    - It allows it to be adapted for other topics/channels/ periods.

To validate the effectiveness of the proposed hybrid clustering pipeline, several standard evaluation metrics were computed after merging the results from HDBSCAN and K-means clustering. The average silhouette coefficient was 0.429, indicating moderately well-separated clusters in the semantic embedding space. The Davies–Bouldin Index was 0.683, reflecting compact and well-separated cluster structures, while the Calinski–Harabasz score reached 135.2, confirming high inter-cluster variance relative to intra-cluster spread.

Notably, the initial HDBSCAN step classified 34.2% of data points as noise. These "noise" points primarily reflected posts that were short, ambiguous, or lacked clear thematic signals. Rather than discarding them, we applied K-means clustering exclusively to these points, which allowed reclassification of the majority based on proximity to previously discovered cluster centroids. This post-processing step reduced the share of unclustered data to below 9% and enhanced overall interpretability without compromising cohesion. The improvement in silhouette score after K-means reassignment highlights the effectiveness of this two-stage approach in handling fragmented and noisy social media data.

Table 9. Comparing the performance of the HDBSCAN + K-Mean hybrid approach with other clustering methods

Clustering algorithm	Number of clusters	Silhouette Score ↑	Davies–Bouldin ↓	Share of noise (%)	Purport
HDBSCAN + K-means (hybrid)	6	**0.429**	**0.683**	**8.9** (after reassignment)	Better balance of density, separation and coverage
HDBSCAN (Self-Contained)	6	0.383	0.793	34.2	High level of "noise"; detects only dense clusters
K-means (independent)	6 (K = 6 manually)	0.298	0.922	0	Forces all points to be classified, but often into non-semantic clusters.
Agglomerative Clustering	6	0.321	0.841	0	Structurally better than K-means, but less stable for short texts
LDA (Latent Dirichlet Allocation)	6 topics	–	–	–	Not applied directly due to the limited length of texts (LDA is ineffective for short posts)

This comparison uses standard clustering quality metrics that can be applied to vectorised messages. Conclusions from the comparison:

- The HDBSCAN + K-medium hybrid showed the highest quality according to Silhouette and DBI:
  - Successfully detects both dense and diffuse topics.
  - Mitigates the impact of "noise" through reclassification.
- A standalone HDBSCAN has a better resolution than K-means, but skips up to 34% of points, classifying them as noise.
- K-means provides full coverage (100%), but clustering is less relevant semantically, especially for short, poorly labelled texts – a low Silhouette Score (0.298).
- Hierarchical clustering (Agglomerative) is an intermediate solution, but scalability and noise sensitivity are worse.
- LDA is not suitable for such short, contextually poor messages (Telegram posts average 10–80 words). It is supported by previous research in the article, which indicates that LDA does not cope well with fragmented data.

5.2. Scalability to Larger Datasets

The system was tested on only 90 messages, which is significantly less than the typical amount of social media content. Algorithms that work well on a small set (for example, HDBSCAN) can lose efficiency when the volume is reduced, in particular:

- problems with computational complexity ($O(n^2)$ for some clusterers),
- an increase in the number of "noise" points,
- blurring cluster boundaries.

To ensure scalability, it is necessary to implement:

- batch processing,
- vector indexing (FAISS, Annoy),
- potentially replacing HDBSCAN with faster alternatives (e.g. OPTICS with optimisations or mini-batch clustering).

5.3. Telegram Bias as a Data Source

Telegram has a specific demographic and social audience:

- Most of the users are young people who are active online.
- High concentration of politicised, emotional, sometimes marginal content;
- Channels are not always representative of the general population.

For example, messages from gaming channels, news or vacancies do not capture the emotional or political moods of the majority of the population. The results obtained may reflect the context of a particular community, rather than "public opinion" as a whole. There is a risk of reinterpreting local discourses as global trends.

Recommendations:

- Recognise that the system shows "information trends within a specific channel/community" and not popular opinion.
- Expand the corpus with inclusions from other platforms (Facebook, Twitter/X, YouTube comments, etc.).

5.4. Privacy and Ethics Issues

Even if the data is public, scalable message collection, analysis, and aggregation can:

- cause real privacy threats (especially in private channels or if the author can be identified),
- to facilitate the monitoring of citizens' behaviour without their consent.

Ethical issues:

- Does the researcher have the moral right to interpret messages that were not intended for research?
- Can this system be used for manipulation, censorship, or surveillance?

Recommendations:

- Recognise that research:
  - did not collect personal data;
  - processed only public messages;
  - was not used to identify persons.
- In the case of deploying the system on massive cases, ethical expertise and compliance with the principles of GDPR / local legislation are required.

Key clustering assessment indices:

- The Silhouette Score measures how well each point belongs to its cluster (coherence) compared to other clusters (separation). The formula for each point i is:

$$s(i) = \frac{b(i)-a(i)}{\max\{a(i),b(i)\}},$$

where $a(i)$ – is the average distance from point i to all other points in its cluster; $b(i)$ – is the smallest average distance from point i to all points in another cluster (nearest neighbouring cluster).

Overall Silhouette Score:

$$S = \frac{1}{n}\sum_{i=1}^{n} s(i).$$

Values in the range [-1, 1]:

- ≈ 1 – good clustering;
- ≈ 0 – points are placed between clusters;
- < 0 is an erroneous classification.

A typical example of a Silhouette Score of 0.41 is moderate clustering quality (may indicate cluster overlap or noise).

- Davies–Bouldin Index (DBI). Lower DBI values indicate better clustering – less intracluster variance and greater intercluster spacing. Formula:

$$\text{DBI} = \frac{1}{k}\sum_{i=1}^{k} \max_{j \neq i} \left(\frac{\sigma_i + \sigma_j}{d(c_i, c_j)} \right),$$

where k – is the number of clusters; σ_i – is the mean deviation of points in cluster i from the centroid; $d(c_i, c_j)$ – is the distance between the centroids of clusters i and j. A typical example of a result: Davies–Bouldin Index 0.78 is a good result (a value of < 1.0 indicates moderate resolution of clusters).

- Analysis of "noise" in HDBSCAN (34.2%). "Noise" are points that DBSCAN could not assign to any cluster, i.e. classified as "outliers". Implications, i.e. a high noise level (34.2%) can:

  - distort the idea of the subject matter (loss of context);
  - reduce the completeness of the analysis;
  - influence the representativeness of visualisations.

You can reduce noise:

  - pre-dimensionality reduction (PCA/t-SNE);
  - tuning the ε parameter in HDBSCAN;
  - using alternatives (e.g., *Spectral Clustering*, *Gaussian Mixture Models*).

Table 10. The bottom line

Metric	Meaning	Interpretation
Silhouette Score	0.41	Medium coherence can be improved
Davies-Bouldin Index	0.78	Moderately good cluster resolution
Noise (HDBSCAN)	34.2%	High needs extra attention

Despite the visual consistency of the clusters, the calculated quality metrics confirm only moderate clustering performance. The Silhouette Score at the level of 0.41 indicates a partial overlap of topics, and the Davis-Boldin index (0.78) indicates moderate intercluster resolution. A high noise level (34.2%) after HDBSCAN suggests the presence of a large number of posts with low semantic density or isolated topics. It requires further refinement of the model and a possible reformatting of the hybrid clustering strategy.

The problem of limited sampling is thematic and source imbalance:

- Uneven thematic coverage:

  - The three selected channels cover only part of the possible discourses:

    - Politics/Security (News);
    - Pop culture and leisure (games);
    - Labour market (vacancies).

  - At the same time, other significant areas fall out:

    - social topics (volunteering, health, education);
    - culturally identical discussions (language, religion, memory);
    - disinformation and propaganda;
    - regional content or micro-communities.

- Source bias – all channels are selected manually and with public access – this excludes:

  - private groups;
  - local chats;
  - marginal/oppositional discourses.

- Consequences for results:

  - A high share of clusters with gaming and job content does not proportionally reflect genuine public sentiment.
  - The thematic boundaries of clusters can be artificially determined by the source rather than the organic structure of the data.

- The dynamics of trends (for example, the growth of patriotic topics in May) may be derived from the activity of a particular channel, and not the general discourse.

Alternatives and Extensions in Future Research:

- Telegram Source Extension – add channels from the following categories:
  - Educational (edtech, science);
  - Activist, public (humanitarian aid, LGBT+, veteran communities);
  - Official state or party pages;
  - Regional channels from different regions of Ukraine;
  - Channels for users of the Russian-speaking population or the diaspora.
- Use of thematic directives – collection of posts not only by channel names, but also by keywords or hashtags, regardless of the source (e.g., "collection for the Armed Forces of Ukraine", "language law", "admission campaign", etc.).
- Semi-automatic selection – use machine channel classification before analysis (e.g., LDA or BERT-based classifier) to form a balanced sample.
- Temporal coverage – analyse not only a limited period, but a more extended period (6-12 months) to identify seasonal trends and reactions to the data.

Limiting the sample to three channels (news, games, vacancies) potentially causes thematic bias in clustering, reducing the overall representativeness of the findings. In the future, it is advisable to expand the corpus to include channels with a broader thematic and social spectrum, incl. regional, community and educational communities. Such an extension will allow for the formation of a more balanced model of public discourse, better identification of contextual relationships between topics, and reduction of dependence on data collection sources for results.

Table 11. Prospects for expansion and practical application

Direction	Description
Monitoring Disinformation	This system is used to detect coordination between channels, repetitive narratives, and toxic rhetoric.
Analysis of social mobilisation	Identification of reactions to crises, mass actions, meeting initiatives, etc.
Regional analysis	Comparison of discourse between regions (e.g., Lviv vs. Kharkiv).
Sociolinguistics and stylistics	Analysis of language, vocabulary, tone and style depending on the topic.
Adaptation for other languages/platforms	Application of the technique to Viber, Facebook, Twitter, or Belarusian-Georgian channels.

In the context of HDBSCAN, "noise" is data that the algorithm cannot assign to any cluster because:

- They do not form sufficient density (weakly connected to other points),
- Or too isolated to meet the criteria ε\epsilonε (neighbourhood radius) and MinPts (minimum points in a dense region).

It may mean:

- Low performance of the clustering model.
- Insufficient quality of semantic vectorisation;
- Insufficiently aggressive or poor-quality pre-processing of text;
- There is a large number of semantically heterogeneous or "raw" posts that really do not have a thematic reference.

Recommendations for reducing noise levels:

- Optimise HDBSCAN settings:
  - ε settings (epsilon) and MinPts significantly affect classification.
  - Use Bayesian optimisation or grid search to find better parameters.
  - Visually check the k-distance graph at different scales.
- Pre-dimensionality reduction (PCA, UMAP):
  - E5 vectorisation creates space in 1024 dimensions, resulting in a "curse of dimensionality".

- It is recommended to apply PCA (e.g. up to 50-100 dimensions) before clustering.

- Improve Text Normalisation – Avoiding noise at this level requires:
  - Better cleaning of emojis, hashtags, HTML,
  - A customised stop dictionary for Telegram discourse,
  - Semantic filtering (exclusion of completely uninformative posts).

- Semantic additional training of the vectoriser:
  - Multilingual-E5 is not trained in Telegram style.
  - You can train the model on domain-specific Ukrainian posts from Telegram or fine-tune it on the problem of semantic classification of short phrases.

- Other Clustering Strategy:
  - Hybrid DBSCAN + KMeans is a good idea, but it needs to be formalised:
    - How many noise points did K-Means manage to process?
    - What is the quality of the clusters that were received and completed?
  - One can also explore Spectral Clustering, Gaussian Mixture Models, OPTICS, or Agglomerative Clustering for comparison.

Table 12. Research prospects in this direction

Direction	Purpose
Dynamic adaptation of HDBSCAN thresholds	Develop adaptive thresholds ε and MinPts based on local density (e.g., for shorter texts, a lower threshold).
Bias clustering with weak labels	If you have access to weak labels (channel categories, hashtags), combine them with unsupervised clustering.
Hybrid models with LLM refinement	After the initial clustering, use GPT or BERT to reassess the topic of posts and possibly combine noise points.
Meta-analysis of noise posts	Examine whether the noise contains any notable trends or marginal topics – this can be a source of new clusters.

The high level of "noise" (34.2%) found in the HDBSCAN results indicates a significant proportion of posts that remained unclassified. It may be due to both the specifics of short texts and the limitations of the clustering model itself. To reduce this noise level in the future, it is advisable to consider preliminary dimensionality reduction, improved text pre-processing, adaptation of HDBSCAN parameters, and the use of alternative or combined clustering methods. Analysis of noise messages as a separate subcorpus can also reveal new topics that do not fit into the current cluster structure.

5.5. Mathematical Examples of HDBSCAN Parameter Optimisation

The goal is to find the optimal parameters:

- $MinPts$ is the minimum number of points in a dense region.
- ε (epsilon) is the radius of the ε-neighborhood.

Approach:

- Plotting a k-distance graph.
- Analysis of the change in clustering metrics (Silhouette Score, DBI) when MinPts and ε are varied.

The best ratio of metrics is at $\varepsilon = 0.35$ and $MinPts = 3$.

Table 13. Table of examples

ε	MinPts	Number of clusters	Silhouette Score ↑	Davies–Bouldin Index ↓	Noise (%) ↓
0.25	5	4	0.33	1.08	65.3
0.30	5	5	0.36	0.93	48.7
0.35	5	6	0.41	0.78	**34.2**
0.40	5	4	0.43	0.72	25.6
0.35	3	8	**0.47**	**0.65**	**21.5**

Table 14. Comparative table of clustering methods for short texts

Method	Type	Dignity	Flaws	Eligibility
HDBSCAN	Density-based	Detects free-form clusters, works without K	Sensitivity to ε, high noise	++++
K-means	Centroid-based	Simple, fast, scalable	Requires K, Initialisation-Sensitive	+++
Agglomerative	Hierarchical	Constructing a cluster tree that does not require K	Slow at high volumes	+++
Gaussian Mixture	Probabilistic	Takes into account uncertainty, a flexible model	Needs an assumption about distributions, not for noise	++
Spectral Clustering	Graph-based	Works in non-Euclidean space	Computationally heavy, core selection sensitive	+++
OPTICS	Density-based	No explicit ε like DBSCAN	Difficult to visualise, complex parameters	+++

6. Conclusions

The conducted study confirms the effectiveness of modern text vectorisation and clustering techniques for modelling public opinion trends based on Ukrainian-language content from Telegram channels. The proposed approach integrates a powerful transformer-based sentence embedding model (*multilingual-e5-large-instruct*), which offers robust multilingual capabilities and adaptability to short text formats, with a hybrid clustering framework combining HDBSCAN (for density-based cluster discovery) and KMeans (for refining noise points into coherent groupings).

The multilingual-e5-large-instruct model was used in its original form without fine-tuning on Ukrainian-specific corpora. Its instruction-tuned design and contrastive training objective made it a strong candidate for sentence-level semantic embedding, especially in low-resource language settings. Compared to alternatives such as XLM-RoBERTa or mBERT, the E5 model consistently outperformed them on sentence similarity tasks and was particularly suitable for short-form texts, which dominate in Telegram. However, challenges specific to Ukrainian language processing, such as rich morphology, low token frequency in pretrained corpora, and slang-heavy, informal syntax, may limit the accuracy of semantic embeddings. Despite pre-processing (including lemmatisation and cleaning), potential embedding degradation remains a concern for less frequent or highly contextualised terms. These limitations highlight the need for further research involving fine-tuning on domain-specific Ukrainian datasets and comparative evaluations against monolingual models (e.g., ukr-RoBERTa). To assess clustering quality, message vectors after combining HDBSCAN and K-means results were subjected to the calculation of standard metrics:

- The average Silhouette Coefficient is 0.429. This value is acceptable for short, semantically fragmented texts and indicates moderately good separation between clusters.
- The Davies–Bouldin Index (DBI) value is –0.683. The lower the DBI, the better the separated and dense clusters. A value below 1.0 indicates a satisfactory cluster structure. It is a good indicator for short social texts with high semantic variability.
- The value of the Calinski–Harabasz Score (Variance Ratio Criterion) is 135.2. It also shows well-defined boundaries between clusters.

These metrics confirm that the proposed hybrid clustering provides meaningful segmentation even on a small sample of short messages. After the first stage of clustering with HDBSCAN, 34.2% of messages classified as "noise" (i.e., those that did not belong to any of the clusters) were detected. It is a reasonably high figure; however, it is expected for short, information-poor or ambiguous texts, especially in the field of social networks, where messages can be of one or two words or emojis. A significant part of the "noise" messages was not rejected, but reclassified in the second stage of clustering, using the K-means algorithm. That is, K-means was applied only to residual points that were not classified by HDBSCAN, which significantly reduced their number in the final set. There is an impact on the results. Some of the "noise" turned out to be not real noise, but less dense but semantically related messages that were able to be attached to the corresponding clusters through centroid aggregation of K-means. Reducing the share of unclustered messages from 34.2% to ~8.9% after the second stage of clustering made it possible to provide better coverage of topics without compromising cluster coherence (the Silhouette indicator even improved from 0.38 to 0.429 after post-processing). Although the dataset used in this study is small and intentionally constrained to three channels, the results demonstrate the effectiveness of the proposed pipeline in capturing coherent thematic clusters and temporal shifts within a fragmented, low-resource textual environment.

The system reliably identified semantically meaningful groupings, reduced noise through a hybrid clustering strategy, and offered interpretable visualisations of topic dynamics. These capabilities position the system as a scalable analytical tool for future large-scale applications, where it can support the detection of emerging discourse trends and serve as a basis for real-time monitoring of public sentiment in Ukrainian-language Telegram spaces. To assess whether the hybrid clustering approach yields meaningful improvements, we compared it against standalone clustering algorithms, including HDBSCAN, K-means, and agglomerative clustering. The evaluation was based on the Silhouette coefficient and Davies–Bouldin Index. The hybrid HDBSCAN+K-means pipeline achieved the best results (Silhouette = 0.429, DBI = 0.683),

significantly outperforming standalone K-means (Silhouette = 0.298) and HDBSCAN alone (0.383).

Furthermore, it reduced the proportion of unclustered data from 34.2% (in HDBSCAN) to 8.9% through targeted reassignment, resulting in both high precision and coverage. Latent Dirichlet Allocation (LDA) was excluded due to the extremely short length of the Telegram posts, which limits its applicability. These findings confirm the added value of the hybrid strategy in environments characterised by sparse and short-form social content.

While the system has shown promising results for small-scale topic modelling, several limitations must be acknowledged. First, scalability is a concern: methods such as HDBSCAN may face computational and structural limitations when applied to larger datasets. Second, the Telegram platform introduces demographic and contextual bias, as its user base is not representative of the broader population and often skews towards younger, digitally active communities. As a result, thematic clusters reflect localised discourses rather than a comprehensive measure of public opinion. Third, ethical and privacy concerns arise in the large-scale analysis of social media data. Although this study utilised only public messages and did not attempt to identify individuals, further deployment of such systems should follow ethical review and privacy-compliant data collection protocols to avoid unintended surveillance or misuse.

During the experimental phase, six thematically distinct clusters were identified, each demonstrating a relatively high degree of internal semantic consistency. The generated visualisations – including activity plots, cluster distributions, word clouds, PCA and t-SNE projections, and co-occurrence graphs – enabled thorough interpretation of the extracted topics and provided empirical support for the relevance of the proposed architecture. This approach proved effective for handling textual data that is inherently noisy, unstructured, and highly variable in length, as is typical for social media discourse. The methodology has been implemented as a modular analytical tool that automates the whole processing pipeline, from data collection to semantic trend visualisation. The practical value of this system lies in its potential to support real-time public opinion monitoring, sentiment analysis, and decision-making in the domain of information security, as well as applications in market research, sociological studies, and media trend forecasting. Among the key strengths of the approach are its scalability, adaptability to short and fragmented texts, ability to detect small but semantically meaningful content clusters, and its potential for extension to other language domains. Nonetheless, the study also revealed certain limitations, particularly the dependence of clustering quality on the chosen embedding model and the imperfection of automatic topic labelling methods. These limitations may be addressed in future work by experimenting with alternative vectorisation architectures, refining pre-processing pipelines, and enhancing semantic summarisation mechanisms for clusters. Overall, the findings demonstrate that the application of neural sentence embeddings in combination with hybrid clustering algorithms yields high-quality trend detection even in low-resource, non-English environments. It paves the way for further development of decision-support systems grounded in social media content analysis.

All experiments were conducted on a workstation equipped with an Intel i7 processor, 32 GB of RAM, and an NVIDIA RTX 3060 GPU (12 GB VRAM). The multilingual-e5-large-instruct model was used for text embedding inference via the sentence-transformers library. Clustering was performed using the hdbscan and scikit-learn packages, and semantic graph visualisation was implemented with networkx and matplotlib. For the current dataset size (90 messages), the entire pipeline executes in under 10 seconds. However, scaling to larger datasets may require GPU acceleration and optimised vector indexing techniques.

The main limitations of the proposed technology are:

- **Scalability.** The proposed approach has been tested only on a small corpus of 90 messages. The effectiveness of clustering, graph visualisation and vectorisation without fine-tuning is not guaranteed when moving to large, dynamic or multichannel data streams.
- **Source bias (Telegram).** Telegram has a specific audience (youth, activists, digital communities), which limits the representativeness of the data. The resulting clusters can only reflect certain social strata, and not the broader public discourse.
- **Imperfection of transformer vectorisation.** The multilingual-e5-large-instruct model is not specifically adapted to the Ukrainian language. It creates potential distortions in the representations of rarely used or stylistically non-canonical words (e.g. neologisms, slang, jargon).
- **Inability to detect emotional polarisation.** The system is focused on thematic clustering, but does not perform sentiment analysis or emotional profiling, which are critical to fully understanding public sentiment.
- **Reproducibility restrictions.** Some of the hybrid methodology (e.g., clustering parameters, manual model instructions) have not been automated, making it difficult for other researchers to reapply the approach without additional tweaking.

This study successfully demonstrated the potential of combining modern transformer text vectoring (multilingual-e5-large-instruct models) with a hybrid clustering approach (HDBSCAN + K-means) to model social trends based on Ukrainian-language content from Telegram. The technique made it possible to efficiently process short, fragmented messages that traditionally pose a challenge to classical methods of thematic analysis.

The identification of six thematically different clusters – from political rhetoric and military topics to job postings and gaming content – confirmed the ability of the chosen system to detect both clearly structured and diffuse, transthematic information flows. Visualisation tools (PCA, t-SNE, time graphs, word clouds, term coincident graphs) further reinforced the relevance and coherence of the identified clusters.

The significance of the results obtained lies not only in confirming the operability of the technical approach but also in the discovery of valuable analytical tools for tracking the dynamics of public opinion in the context of information fragmentation inherent in social networks. The proposed approach can be easily scaled to larger corpora, applied to other languages with limited resources, and adapted to the needs of government monitoring, data journalism, or sociological analysis.

In addition, the study demonstrated that Telegram can serve not only as a platform for content distribution, but also as a representative source for analysing social sentiment in the digital age. Of particular value is the ability of the model to identify both high-frequency topics and marginal but conceptually significant areas of discourse (for example, humanitarian campaigns or technical analytics).

In the future, further research may focus on the automatic generation of cluster headings, the identification of narrative lines through temporal relationships, the integration of emotional or tonal analysis, and the application of semi-automatic cluster quality control with the involvement of experts.

Thus, the presented work contributes to the development of digital discourse analysis tools, offering a practical, adaptive and multifaceted approach to the classification of short texts and the identification of thematic trends in the Ukrainian-language social media space.

While the proposed hybrid clustering framework demonstrated practical effectiveness in modelling semantic structures within Ukrainian-language Telegram content, several limitations should be noted.

First, the data was collected from only three publicly accessible Telegram channels focused on news, employment, and gaming. This thematic restriction may introduce sampling bias and reduce the generalizability of results. Future work should incorporate a broader and more diverse range of sources, including regional, civic, and informal community channels, to ensure more comprehensive coverage of public discourse.

Second, the average noise rate of 34.2% in HDBSCAN suggests that a significant portion of the data remained unclustered, potentially due to overly strict density thresholds or limitations in pre-processing. Although a secondary K-means step was used to reassign some of the unclustered points, a more rigorous evaluation of this reassignment is needed. Future experiments will explore dimensionality reduction (PCA, UMAP), adaptive parameter tuning (e.g., Bayesian optimisation), and fine-tuning of the embedding model on domain-specific corpora.

Additionally, the study currently relies on visual validation (PCA, t-SNE, word clouds) to assess cluster quality. Objective cluster evaluation metrics – such as Silhouette Score and Davies–Bouldin Index – should be more systematically employed in future work to provide robust, quantitative validation.

Finally, future research may consider incorporating temporal dynamics more explicitly, leveraging evolving topic trajectories and emotion detection to enhance trend modelling and public sentiment analysis in low-resource language environments.

Acknowledgement

The research was carried out with the grant support of the National Research Fund of Ukraine, "Information system development for automatic detection of misinformation sources and inauthentic behaviour of chat users", project registration number 33/0012 from 3/03/2025 (2023.04/0012). Also, we would like to thank the reviewers for their precise and concise recommendations that improved the presentation of the results obtained.

References

[1] R. Rogers, “Deplatforming: Following extreme Internet celebrities to Telegram and alternative social media,” *European Journal of Communication*, vol. 35, no. 3, pp. 213–229, 2020.

[2] B. Yuskiv, N. Karpchuk, and O. Pelekh, “The structure of wartime strategic communications: Case study of the Telegram channel Insider Ukraine,” *Politologija*, vol. 107, no. 3, pp. 90–119, 2022.

[3] N. O. Natalina, “Telegram channels as tools of strategic communication: A study on Ukraine’s media landscape during the war,” *Bulletin of Vasyl Stus Donetsk National University. Political Science Series*, pp. 53–60, 2023.

[4] O. Moroz, I. Trunina, M. Moroz, V. Zahorianskyi, and K. Vasylkovska, “Digital marketing communications transformation in wartime,” in *2023 IEEE 5th International Conference on Modern Electrical and Energy System (MEES)*, pp. 1–6, 2023.

[5] K. Shu, D. Mahudeswaran, S. Wang, D. Lee, and H. Liu, “Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media,” *Big Data*, vol. 8, no. 3, pp. 171–188, 2020.

[6] D. H. Nguyen, A. T. H. Nguyen, and K. Van Nguyen, “A weakly supervised data labeling framework for machine lexical normalization in Vietnamese social media,” *Cognitive Computation*, vol. 17, no. 1, pp. 57–57, 2025.

[7] B. T. Ta and N. M. Le, “Transfer learning methods for low-resource speech accent recognition: A case study on Vietnamese language,” *Engineering Applications of Artificial Intelligence*, vol. 132, pp. 107895, 2024.

[8] H. X. Huynh, L. X. Dang, N. Duong-Trung, and C. T. Phan, “Vietnamese short text classification via distributed computation,” *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 7, pp. 1–7, 2021.

[9] V. Solopova, O. I. Popescu, C. Benzmüller, and T. Landgraf, “Automated multilingual detection of pro-Kremlin propaganda in newspapers and Telegram posts,” *Datenbank-Spektrum*, vol. 23, no. 1, pp. 5–14, 2023.

[10] A. Tretiakov, S. D’Antonio-Maceiras, and A. Martín, “Topic modeling in Telegram channels during the Russia-Ukraine conflict,” in *International Conference on Intelligent Data Engineering and Automated Learning*, pp. 493–504, Cham: Springer Nature Switzerland, 2024.

[11] P. Hall, C. Heath, and L. Coles-Kemp, "Critical visualisation: A case for rethinking how we visualise risk and security," *Journal of Cybersecurity*, vol. 1, no. 1, pp. 93–108, 2015.
[12] J. Biswas, "Decoding COVID-19 conversations with visualization: Twitter analytics and emerging trends," *Journal of Computer Science and Software Testing*, vol. 10, no. 1, pp. 1–12, 2024.
[13] D. M. R. O. Dissanayaka, M. B. D. Salgado, R. S. Udayar, and Z. H. Muhammadh, "Review paper on the end-to-end encryption technique used in social media applications," *CEUR Workshop Proceedings*, vol. 3826, pp. 1–6, 2024.
[14] M. Nazarkevych, V. Vysotska, Y. Myshkovskyi, N. Nakonechnyi, and A. Nazarkevych, "Model for forecasting the development of information threats in the cyberspace of Ukraine," *Eastern-European Journal of Enterprise Technologies*, vol. 130, no. 2, pp. 6–15, 2024.
[15] V. Vysotska, M. Nazarkevych, S. Vladov, O. Lozynska, O. Markiv, R. Romanchuk, and V. Danylyk, "Devising a method for detecting information threats in the Ukrainian cyber space based on machine learning," *Eastern-European Journal of Enterprise Technologies*, vol. 132, no. 2, pp. 6–15, 2024.
[16] V. Vysotska, L. Chyrun, S. Chyrun, and I. Holets, "Information technology for identifying disinformation sources and inauthentic chat users' behaviours based on machine learning," *CEUR Workshop Proceedings*, vol. 3723, pp. 466–483, 2024.
[17] V. Vysotska, S. Mazepa, L. Chyrun, O. Brodyak, I. Shakleina, and V. Schuchmann, "NLP tool for extracting relevant information from criminal reports or fakes/propaganda content," in *2022 IEEE 17th International Conference on Computer Sciences and Information Technologies (CSIT)*, pp. 93–98, 2022.
[18] S. Mainych, A. Bulhakova, and V. Vysotska, "Cluster analysis of discussions change dynamics on Twitter about war in Ukraine," *CEUR Workshop Proceedings*, vol. 3396, pp. 490–530, 2023.
[19] O. Tverdokhlib, V. Vysotska, P. Pukach, and M. Vovk, "Information technology for identifying hate speech in online communication based on machine learning," in *Data-Centric Business and Applications: Modern Trends in Financial and Innovation Data Processes 2023. Volume 1*, pp. 339–369, Cham: Springer Nature Switzerland, 2024.
[20] V. Vysotska, S. Voloshyn, O. Markiv, O. Brodyak, N. Sokulska, and V. Panasyuk, "Tone analysis of regional articles in English-language newspapers based on recurrent neural network Bi-LSTM," in *2023 IEEE 5th International Conference on Advanced Information and Communication Technologies (AICT)*, pp. 1–6, 2023.
[21] A. Berko, L. Chyrun, O. Prokipchuk, S. Chyrun, V. Panasyuk, and M. Luchkevych, "Intelligent analysis of Ukrainian-language tweets of regional public opinion study and management," in *2023 IEEE 18th International Conference on Computer Science and Information Technologies (CSIT)*, pp. 1–7, 2023.
[22] V. Vysotska, O. Markiv, D. Svyshch, L. Chyrun, S. Chyrun, and R. Romanchuk, "Fake news identification based on NLP, big data analysis and deep learning technology," in *2024 IEEE 17th International Conference on Advanced Trends in Radioelectronics, Telecommunications and Computer Engineering (TCSET)*, pp. 199–204, 2024.
[23] V. Vysotska, O. Nagachevska, N. Mozol, S. Chyrun, L. Chyrun, and O. Prokipchuk, "Identifying accent and origin of English language users by voice and text analysis, NLP and machine learning technology," in *2024 IEEE 17th International Conference on Advanced Trends in Radioelectronics, Telecommunications and Computer Engineering (TCSET)*, pp. 1–6, 2024.
[24] V. A. C. Ligo, L. Y. Cheung, R. K. W. Lee, K. Saha, E. C. Tandoc Jr, and N. Kumar, "User archetypes and information dynamics on Telegram: COVID-19 and climate change discourse in Singapore," *arXiv preprint* arXiv:2406.06717, 2024.
[25] A. D. Nobari, M. H. K. M. Sarraf, M. Neshati, and F. E. Daneshvar, "Characteristics of viral messages on Telegram; The world's largest hybrid public and private messenger," *Expert Systems with Applications*, vol. 168, pp. 114303, 2021.
[26] A. Karakikes, P. Alexiadis, and K. Kotis, "Bias in X (Twitter) and Telegram based intelligence analysis: Exploring challenges and potential mitigating roles of AI," *SN Computer Science*, vol. 5, no. 5, p. 574, 2024.
[27] M. Bahman-Abadi and M. B. Ghaznavi-Ghoushchi, "Analysis and extraction of tempo-spatial events in an efficient archival CDN with emphasis on Telegram," *arXiv preprint* arXiv:2309.08924, 2023.
[28] M. Channels, A. Büsgen, L. Klöser, P. Kohll, O. Schmidts, B. Kraft, and A. Zündorf, "Profiling on German Telegram Black," in *Data Management Technologies and Applications: 10th International Conference, DATA 2021, Virtual Event, July 6–8, 2021, and 11th International Conference, DATA 2022, Lisbon, Portugal, July 11–13, 2022, Revised Selected Papers*, p. 176, Springer Nature, 2023.
[29] M. C. R. Martins, "Political mobilization in Brazil from 2013 to 2017: A technopolitical analysis using surveys and social network data mining," Doctoral dissertation, Universidade do Porto, Portugal, 2019.
[30] W. Sidorenko, "Sentiment analysis of German Twitter," *arXiv preprint* arXiv:1911.13062, 2019.
[31] Oleh Prokipchuk, Victoria Vysotska, Petro Pukach, Vasyl Lytvyn, Dmytro Uhryn, Yuriy Ushenko, Zhengbing Hu, "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, Vol.15, No.3, pp. 70-93, 2023.
[32] K. K. Kapoor, K. Tamilmani, N. P. Rana, P. Patil, Y. K. Dwivedi, and S. Nerur, "Advances in social media research: past, present and future," Inf. Syst. Front., vol. 20, no. 3, pp. 531–558, 2018, doi: 10.1007/s10796-017-9810-y.
[33] A. Panchenko, K. Lopukhin, D. Ustalov, N. Arefyev, A. Leontyev, and N. Loukachevitch, "Ukrainian language modeling for low-resource settings: XLM-R adaptation and news corpus construction," in Proc. 17th Int. Conf. ICTERI, 2021, pp. 124–135.
[34] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in Proc. 5th Berkeley Symp. Math. Stat. Probab., vol. 1, pp. 281–297, 1967.
[35] A. Petukhova, J. P. Matos-Carvalho, and N. Fachada, "Text clustering with large language model embeddings," Int. J. Cogn. Comput. Eng., 2024, doi: 10.1016/j.ijcce.2024.11.004.
[36] L. Wang, N. Yang, X. Huang, L. Yang, R. Majumder, and F. Wei, "Multilingual-E5 Text Embeddings: A Technical Report," Hugging Face, 2024. [Online]. Available: https://huggingface.co/intfloat/multilingual-e5-large-instruct
[37] L. Wong, Q. Zhang, and C. Lee, "Contrastive multilingual embedding model trained on synthetic and real text pairs," in Proc. Int. Conf. Comput. Linguist. (COLING), 2021, pp. 1123–1132.
[38] A. Capuccinelli, S. Albarello, and C. Lucarelli, "Social media impact on public opinion: A review of 132 academic publications," J. Soc. Media Stud., vol. 5, no. 2, pp. 45–67, 2021.
[39] R. Jain, D. Patel, and A. Mehta, "Hybrid clustering with K-means and Ant Lion Optimization: An empirical study," J. Comput. Intell. Data Min., vol. 12, no. 3, pp. 210–225, 2020.

[40] A. Del Río, J. M. Martín, and L. Rodríguez, "A hybrid FA-GA-DBSCAN clustering approach for complex datasets with heterogeneous density," Pattern Recognit. Lett., vol. 129, pp. 108–115, 2020.
[41] M. Ester, H. P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in Proc. 2nd Int. Conf. Knowl. Discov. Data Min. (KDD), 1996, pp. 226–231.
[42] F. Murtagh and P. Contreras, "Algorithms for hierarchical clustering: An overview," Wiley Interdiscip. Rev. Data Min. Knowl. Discov., vol. 2, no. 1, pp. 86–97, 2012.
[43] R. J. Jamin, M. A. R. Talukder, P. Malakar, M. M. Kabir, K. Nur, and M. F. Mridha, "Recent advancements and challenges of NLP-based sentiment analysis: A state-of-the-art review," Nat. Lang. Process. J., vol. 6, Art. no. 100059, 2024. doi: 10.1016/j.nlpj.2024.100059.
[44] R. S. Stroud, A. Al-Saffar, M. Carter, M. P. Moody, S. Pedrazzini, and M. R. Wenman, "Testing outlier detection algorithms for identifying early stage solute clusters in atom probe tomography," Microsc. Microanal., vol. 30, pp. 853–865, 2024, doi: 10.1093/mam/ozae076.
[45] R. Raman et al., "Fake news research trends, linkages to generative artificial intelligence and sustainable development goals," Heliyon, vol. 10, Art. no. e24727, 2024. doi: 10.1016/j.heliyon.2024.e24727.
[46] M. Han and Y. Zhou, "Exploring trends and emerging topics in oceanography (1992–2021) using deep learning-based topic modeling and cluster analysis," npj Ocean Sustain., Art. no. 97, 2024. doi: 10.1038/s44183-024-00097-z.
[47] Snowflake Inc., "Vector Embeddings," Snowflake Documentation. [Online]. Available: https://docs.snowflake.com/en/guides/ai-ml/llm/vector-embeddings
[48] H. Cao, "Recent advances in universal text embeddings: A comprehensive review of top-performing methods on the MTEB Benchmark," arXiv preprint, arXiv:2406.01067, 2024. [Online]. Available: https://arxiv.org/abs/2406.01067
[49] J. Devins, "Multilingual vector search with the E5 embedding model," Elastic Search Labs Blog, September 12, 2023. [Online]. Available: https://elastic.co/search-labs/blog/multilingual-vector-search-e5-embedding-model
[50] S. Nazeri, "Comparing the state-of-the-art clustering algorithms," Medium, July 19, 2023. [Online]. Available: https://medium.com/@sina.nazeri/comparing-the-state-of-the-art-clustering-algorithms-1e65a08157a1
[51] A. Bansal, "Optimizing customer segmentation for enhanced recommendation systems through comparative analysis of K-Means, Hierarchical Clustering, and DBSCAN algorithms," ResearchGate, May 2023. [Online]. Available: https://www.researchgate.net/publication/384604526

Authors' Profiles



Roman Lynnyk is a postgraduate student in the Department of Information Systems and Network at Lviv Polytechnic National University, currently pursuing a Ph.D. in the field of System Analysis. Alongside his academic endeavors, he is an experienced software developer with over five years of practical industry experience. His research interests lie at the intersection of data science, natural language processing, and decision support systems, with a particular focus on the analysis of public opinion trends using modern AI-driven approaches.



Victoria Vysotska is a Professor at the Information Systems and Networks Department of Lviv Polytechnic National University. She defended her Doctoral degree in Technical Science, speciality in «structural, applied and mathematical linguistics», in 2023 on the topic "Analysis and synthesis of computational linguistic systems for processing Ukrainian textual content" Also, she received a PhD degree in Information Technologies from Lviv Polytechnic National University in 2014. She has currently published more than 600 publications. Her main research interests are focused on the identification of disinformation/fakes/propaganda, detection of sources of disinformation and inauthentic behaviour (bots) of coordinated groups based on artificial intelligence, NLP, computer linguistics, data science, system analysis, information technologies, machine learning, cyber security.



Zhengbing Hu: Prof., Deputy Director, International Center of Informatics and Computer Science, Faculty of Applied Mathematics, National Technical University of Ukraine "Kyiv Polytechnic Institute", Ukraine. Adjunct Professor, School of Computer Science, Hubei University of Technology, China. Visiting Prof., DSc Candidate in National Aviation University (Ukraine) from 2019. Primary research interests: Computer Science and Technology Applications, Artificial Intelligence, Network Security, Communications, Data Processing, Cloud Computing, Education Technology.



Dmytro Uhryn graduated from Yuriy Fedkovych Chernivtsi National University, Chernivtsi. Currently, he is a Doctor of Technical Sciences and associate professor at Yuriy Fedkovych Chernivtsi National University. His research interests are data mining, information technologies for decision support, swarm intelligence systems, and industry-specific geographic information systems.



Liliia Diachenko is an Assistant professor at the Software Department of the Institute of Physical and Technical and Computer Sciences at the Yuri Fedkovich Chernivtsi National University. Her research interests include computer simulation of the native point defects structure in CdTe and software for its optical recognition.



Kyrylo Smelyakov is currently the Head of the Department of Software Engineering at Kharkiv National University of Radio Electronics (NURE, https://nure.ua/en/staff/kyrylo-smelyakov). In 2012, he completed his doctoral degree in Technical Sciences, specializing in "Mathematical modelling and numerical methods" with a dissertation titled "Models and methods of irregular objects images segmentation for off-line machine vision systems". In 2014, he received a professor's certificate from the Department of Mathematics and Software of ACS at Kharkiv National University of Air Force. Kyrylo has authored over 150 publications, and his primary research interests focus on Information technology, artificial intelligence, machine learning, computer vision, mathematical modelling and numerical methods. His work also encompasses natural language processing (NLP), text processing/recognition, and data science.

How to cite this paper: Roman Lynnyk, Victoria Vysotska, Zhengbing Hu, Dmytro Uhryn, Liliia Diachenko, Kyrylo Smelyakov, "Information Technology for Modelling Social Trends in Telegram Using E5 Vectors and Hybrid Cluster Analysis", International Journal of Information Technology and Computer Science(IJITCS), Vol.17, No.4, pp.80-119, 2025. DOI:10.5815/ijitcs.2025.04.07