I.J. Information Technology and Computer Science, **2026, 1, 180-245**
Published Online on February 8, 2026 by MECS Press (http://www.mecs-press.org/)
DOI: 10.5815/ijitcs.2026.01.10

Hybrid Information Technology for Automatic Detection of Disinformation and Inauthentic Behaviour of Chat Users based on NLP, Machine Learning, and Graph Analysis

Victoria Vysotska
Information Systems and Networks Department, Lviv Polytechnic National University, Lviv, 79013, Ukraine
Combating Cybercrime Department, Kharkiv National University of Internal Affairs, Kharkiv, 61080, Ukraine
E-mail: Victoria.A.Vysotska@lpnu.ua
ORCID iD: https://orcid.org/0000-0001-6417-3689

Lyubomyr Chyrun
Information Systems and Networks Department, Lviv Polytechnic National University, Lviv, 79013, Ukraine
Department of Applied Mathematics Department, Ivan Franko National University of Lviv, Lviv, 79000, Ukraine
E-mail: lyubomyr.v.chyrun@lpnu.ua
ORCID iD: https://orcid.org/0000-0002-9448-1751

Oleksandr Lavrut
Tactics Department, Hetman Petro Sahaidachnyi National Army Academy, Lviv, 79012, Ukraine
E-mail: alexandrlavrut@gmail.com
ORCID iD: https://orcid.org/0000-0002-4909-6723

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

Yurii Ushenko*
Department of Physics, Shaoxing University, Shaoxing, Zhejiang Province 312000, China
Department of Computer Science, Yuriy Fedkovych Chernivtsi National University, Chernivtsi, 58012, Ukraine
E-mail: y.ushenko@chnu.edu.ua
ORCID iD: https://orcid.org/0000-0003-1767-1882
*Corresponding author

Received: 11 July 2025; Revised: 15 September 2025; Accepted: 24 October 2025; Published: 08 February 2026

Abstract: In the context of the growth of information exchange in social networks, messengers, and chats, the problem of spreading disinformation and coordinated inauthentic behaviour by users is becoming increasingly relevant and poses a threat to the state's information security. Traditional manual monitoring methods are ineffective due to the scale and speed of information dissemination, necessitating the development of intelligent automated countermeasures. The paper proposes a hybrid information technology for the automatic detection of disinformation, its sources of spread, and inauthentic behaviour among chat users, combining methods of natural language processing (NLP), machine learning, stylistic and linguistic analysis of texts, and graph analysis of social interactions. Within the study, datasets of authentic and fake messages were compiled, and mathematical models and algorithms for identifying disinformation sources were developed using metrics of graph centrality, clustering, and bigram Laplace smoothing.

Experimental studies using TF-IDF, BERT, MiniLM, ensemble methods, and transformers confirmed the effectiveness of the proposed approach. The achieved accuracy in disinformation classification is up to 89.5%, and integrating content, network, and behavioural analysis significantly improves the quality of detecting coordinated information attacks. The results obtained are both scientifically novel and of practical value. They can be used to create systems for monitoring information threats, supporting cybersecurity decision-making, fact-checking, and protecting Ukraine's information space.

Index Terms: Disinformation, Fake News, Information Security, Inauthentic Behaviour, Bots, Social Media, Natural Language Processing, Machine Learning, Graph Analysis, Cybersecurity.

1. Introduction

In today's rapidly evolving digital information landscape, social networks, instant messaging platforms, and online chat services have become the primary channels for communication, news dissemination, and public opinion formation. At the same time, it is these environments that are actively used to spread disinformation, propaganda, and manipulative content, which is carried out by both individual actors and coordinated groups using botnets and fake accounts. The scale, speed, and anonymity of information dissemination in the digital environment significantly complicate its verification and create threats to the state's information security, social stability, and trust in information resources. Following 2022, the issue of countering disinformation in Ukraine has become particularly acute due to the intensification of hybrid information attacks aimed at destabilising society and undermining national security. Existing approaches to detecting fake news primarily focus on analysing the text content of English-language resources, overlooking the peculiarities of the Ukrainian language, the socio-cultural context, users' behavioural characteristics, and the routes of information dissemination. Traditional manual monitoring and fact-checking are ineffective due to the sheer volume of data and time constraints, necessitating the development of intelligent automated systems to counter disinformation. In this context, it is relevant to develop hybrid information technologies that combine methods of natural language processing, machine learning, stylistic and behavioural analysis, and graph modelling of social interactions to automatically detect disinformation, its sources, and inauthentic behaviour by chat users in real time. The purpose of the study is to increase the level of information security of the state through the development of mathematical models, methods and information technology for automatic detection of disinformation, sources of its dissemination and inauthentic behaviour of chat users in the Internet space based on linguistic, stylistic, behavioural and graph analysis using machine learning and natural language processing methods.

To achieve this goal in the work, it is necessary to solve the following main tasks:

- Analyse modern scientific approaches and information technologies for automatic detection of disinformation, fake news and propaganda content in social networks and chats.
- Form and prepare datasets of authentic and fake messages that account for users' linguistic, thematic, and behavioural characteristics.
- Develop a mathematical model for the automatic detection of disinformation and inauthentic behaviour of chat users based on a combination of content and behavioural analysis.
- Develop methods of stylistic and linguistic analysis of texts to identify the characteristic features of fake messages and propaganda narratives.
- Develop graph models to analyse the routes of disinformation spread and identify its primary sources using centrality metrics.
- Investigate the effectiveness of machine learning methods and transformer models in classifying disinformation.
- Develop and implement a prototype of an information system for automatic detection of disinformation and inauthentic behaviour of chat users.
- Conduct experimental validation of the developed models and methods, and assess their effectiveness based on relevant quality metrics.

The object of the study is the processes of spreading disinformation, fake news, and propaganda content in social networks, instant messengers, and online chats, as well as the methods of forming inauthentic user behaviour in the digital environment. The subject of the study is mathematical models, methods, and information technologies for the automatic detection of disinformation, its sources of spread, and inauthentic behaviour of chat users, based on linguistic, stylistic, behavioural, and graph analysis using machine learning methods. The scientific novelty of the study lies in the fact that:

- For the first time, a comprehensive hybrid model for automatic detection of disinformation and inauthentic behaviour of chat users was developed, which combines linguistic, stylistic, behavioural and graph analysis within a single information technology.
- For the first time in the Ukrainian-language information space, Laplace anti-aliasing was applied to bigrams in the context of classifying fake messages, thereby increasing the stability and accuracy of models on uneven datasets.
- Methods for identifying sources of disinformation by analysing the routes of its spread using graph centrality metrics (such as PageRank and Degree Centrality) and time characteristics have been enhanced.
- Methods for detecting inauthentic behaviour of chat users based on a set of behavioural, linguistic and network metrics have been further developed.
- Experimental confirmation of the effectiveness of combining content and behavioural analysis was obtained,

compared to the use of separate approaches.

The practical value of the research results lies in the potential to utilise the developed models, methods, and software prototypes to create real-time information systems for monitoring and countering disinformation. Government agencies can utilise the proposed information technology, think tanks, media, public fact-checking organisations, and cybersecurity structures to automatically detect fake messages, sources of information attacks, and coordinated botnets. The application of the results of this work enhances the accuracy of automatic disinformation detection in Ukrainian-language content, reduces the time spent analysing information flows, and increases the efficiency of decision-making in the fields of information and cybersecurity.

The study is guided by a set of falsifiable research hypotheses (H1–H8), which are empirically tested through comparative experiments and statistical evaluation of classification performance metrics. Research hypotheses:

Main research hypothesis (H0 / H1):

H1 (central research hypothesis) – integration of linguistic (NLP), behavioural, and graph features into a single hybrid model provides a statistically significant increase in the accuracy of detecting disinformation and inauthentic behaviour of chat users compared to models using only one type of feature.

H0 (null research hypothesis) – use of a hybrid model does not lead to a statistically significant improvement in the quality of disinformation detection compared to single-component approaches. Falsification – comparison of F1-score / ROC-AUC of the hybrid model and the base models (text-only, behaviour-only, graph-only).

Partial (operationalised) hypotheses

Hypothesis H2 (linguistic) – texts of disinformation messages have statistically significant linguistic and stylistic features (emotional richness, repeated bigrams, specific propaganda constructions), which can be detected using NLP models. Falsification – lack of substantial difference between fake/accurate classes by linguistic features.

Hypothesis H3 (bigrams + Laplace smoothing) – application of a bigram language model with Laplace smoothing increases stability and accuracy of misinformation classification on unbalanced and small corpora of texts; falsification – lack of increase in F1-score or Recall after application of Laplace smoothing.

Hypothesis H4 (behavioural) – users with inauthentic behaviour (bots, coordinated accounts) demonstrate statistically different activity patterns (publication frequency, time intervals, share of reposts) compared to authentic users; falsification – lack of significant differences in behavioural metrics between groups.

Hypothesis H5 (graph) – sources of disinformation in social chats have increased values of graph metrics (PageRank, Degree Centrality, Betweenness Centrality), which allows identifying them as key nodes of the spread of information attacks; falsification – random or uniform distribution of graph metrics among sources.

Hypothesis H6 (coordinated campaigns) – coordinated disinformation campaigns form dense communities in the interaction graph, which can be automatically detected using Louvain / Leiden algorithms; falsification – lack of clusters with a high concentration of disinformation content.

Hypothesis H7 (transformers) – context-oriented transformer models (BERT, MiniLM) provide higher-quality misinformation classification than traditional n-gram and TF-IDF models. Falsification – lack of increase in Precision / Recall / F1 when using transformers. Hypothesis H8 (ensemble models) – ensemble combination of several classifiers (SVM, Logistic Regression, Random Forest) reduces error variance and increases the generalisation ability of the disinformation detection system; falsification – lack of improvement in ensemble metrics relative to the best single model.

2. Related Works

In the context of the active development of the digital space and the growth of information exchange on social networks, instant messengers, and chats, the problem of disinformation is becoming increasingly acute. The widespread use of bots, fake accounts, and coordinated campaigns to manipulate public opinion distorts public discourse, erodes trust in information sources, and poses threats to the state's information security. Traditional methods of moderation and manual monitoring of information flows are becoming ineffective due to the scale and speed of data distribution. Therefore, it is urgent to develop an intelligent information system capable of automatically identifying sources of disinformation and inauthentic behaviour among chat users by analysing content, activity patterns, and social interactions. The main areas of research in the world, according to the automatic detection of disinformation and inauthentic behaviour of chat users based on NLP, machine learning and graph analysis:

- Graph approaches and GNN models for detecting fake news.
- NLP-oriented approaches to disinformation detection.
- Analysis of user behaviour and emotions.
- Hybrid and optimised deep learning models.

The article [1] proposes the SOMPS-Net graph model, which integrates social interactions, metadata, and publication information to detect fake news early, particularly in the healthcare field. The model is based on graph structures and

attention mechanisms, enabling consideration of the social network's influence on information dissemination. Experiments show a significant improvement in classification quality compared to standard graph models.

The study [2] presents a framework that combines NLP text representations with the Knowledge Graph and Graph Neural Networks (GNNs) to detect fake news across domains such as politics, business, and healthcare. GNN models enable a deeper examination of the structural relationships between entities and events in information, which is particularly useful when analysing interdependent linguistic and contextual elements.

A comparative analysis in [3] shows that graph neural networks are significantly superior to traditional ML algorithms for detecting fake news, as they effectively account for network structure and user behavioural patterns, rather than just message content.

Knowledge Graphs (GNNs) achieve strong results on the fake news problem, especially when integrated with NLP text representations and behavioural features. It confirms the validity of the hybrid model in this study, which combines content and sociographic analysis.

Research in [4] analyses the use of various NLP techniques (sentiment analysis, NER, word embeddings) to improve the accuracy of identifying disinformation on social media. The authors emphasise the importance of employing multiple NLP methods to identify subtle language patterns that frequently occur in fake content.

The review article [5] systematises modern approaches to detecting fake news, including traditional ML classifiers, NLP algorithms, knowledge-based approaches, and hybrid models. Particular attention is paid to the importance of an integrated approach, which combines linguistic analysis and structural patterns in networks.

Modern NLP approaches consider not only the basic classification of the text but also the use of a wide range of linguistic features (e.g., sentiment and stylistic patterns). It confirms the need to include stylistic and linguistic analysis in the model for detecting disinformation.

The work [6] presents a model that combines the analysis of emotional tone in posts (using RNN) and user behavioural parameters, and then determines the likelihood that a particular piece of content is disinformation. Experiments have shown that incorporating emotional information alongside behavioural traits significantly improves classification quality.

Behavioural analysis of users, along with emotional signs, is an essential component of modern detection models. It is consistent with the idea of incorporating behavioural characteristics into a general model.

A study [7] presents a hybrid model that integrates a recurrent LSTM network to extract text features, a CGPNN (graph pulse network) for structural features, and meta-heuristic optimisation to enhance performance. This approach achieves higher classification metrics than the basic models. Another paper [8] demonstrates that ensemble deep models, which combine GNN with various classifier architectures, achieve high accuracy on standard datasets such as GossipCop and Politifact, particularly in detecting signs of psychological influence from text.

Such studies are primarily based on a combination of methods, graph structures, behavioural and emotional contexts, and the analysis of low-resource languages. These publications confirm the effectiveness of hybrid architectures that combine text, structure, and optimisation components – this is the architecture you plan to integrate into your solution. Most current work emphasises the importance of hybrid models, where NLP, graph analysis, behavioural metrics, and deep learning collaborate rather than operate separately [5]. GNNs and knowledge graphs are becoming the standard for analysing complex relationships among users, entities, and content [3]. Taking into account not only the text but also the context of user behaviour and emotional cues in posts significantly improves the quality of fake and disinformation detection [6]. Recent work also highlights the problem of detection in multilingual, low-resource contexts, which is relevant to Ukrainian-language studies [9]. The combination of these information technologies in the study, including hybrid models, graph structures, analyses of behavioural and emotional contexts, as well as analyses of low-resource languages, increases the chance of obtaining accuracy in the automatic detection of misinformation and inauthentic behaviour of chat users based on NLP, machine learning, and graph analysis [10-12].

3. Research Methods and Linguistic Modelling

The primary objective of the study is to employ a hybrid approach that combines content analysis (linguistic indicators of disinformation) and behavioural analysis (user activity models). The application of neural networks contributes to the investigation and automatic identification of signs of manipulation, aggression, or falsity in texts, as well as the analysis of social graphs of user interactions to identify coordinated networks. It is achieved through machine learning, both with and without a teacher, to classify users and messages. The object of the study is the processes of automatic detection of fake news/propaganda/disinformation, as well as the sources of disinformation and inauthentic behaviour among chat users. The study examines models, methods, and means for detecting fake news/propaganda/disinformation, as well as sources of disinformation and inauthentic behaviour among chat users.

The primary hypothesis of the study is that the behavioural patterns of inauthentic users (bots) differ significantly from those of real users in terms of message frequency, time intervals, and interaction structure. In parallel, text messages containing misinformation have statistically significant linguistic and semantic features that can be detected using NLP models. Combining content and behavioural analysis improves the accuracy of automatic disinformation source detection compared to using a single approach. A real-time information system can significantly reduce the spread of false information in chats by detecting and blocking suspicious sources early. The research process and mathematical

formalisation consist of performing the following basic steps of the pipeline process to automatically detect misinformation and inauthentic behaviour of chat users based on NLP, machine learning, and graph analysis:

Step 1. Formation of input data and feature space.
Step 2. Linguistic model with bigrams and Laplace smoothing.
Step 3. Machine learning models for classification (Logistic regression for classification, Support Vector Machine (SVM), Ensemble methods).
Step 4. Construction of a graph of the spread of disinformation.
Step 5. Clustering and community discovery (k-means and Louvain / Leiden).
Step 6. Integrated hybrid model.
Step 7. Performance evaluation.

This study is grounded in established theoretical frameworks from statistical learning theory, probabilistic language modelling, and graph theory, with a focus on information diffusion dynamics. The mathematical formulations used in the proposed approach are not heuristic but are derived from well-founded theories, under standard assumptions commonly adopted in machine learning and network analysis.

The task of misinformation detection is formulated as a supervised binary classification problem within the framework of statistical learning theory. Let $(X, \mathcal{F}, P)$ denote an unknown probability space, where $x \in X \subset R^d$ represents a feature vector extracted from textual, behavioural, and graph-based data, and $y \in \{0,1\}$ denotes the ground-truth class label (0 – reliable information, 1 – misinformation). A classifier $f_\theta: X \to \{0,1\}$ is learned by minimising the expected risk:

$$R(f) = E_{(x,y)\sim P}[\mathcal{L}(f(x), y)]$$

where $\mathcal{L}(\cdot)$ is a loss function. Since the actual distribution P is unknown, empirical risk minimisation (ERM) is applied:

$$\hat{R}(f) = \frac{1}{n}\sum_{i=1}^{n} \mathcal{L}(f(x_i), y_i)$$

This formulation follows the classical Vapnik–Chervonenkis (VC) theory, which establishes bounds on generalisation error as a function of the hypothesis space's complexity and the sample size.

The learning process relies on the following standard assumptions:

- Independent and identically distributed samples (i.i.d.), which enables convergence of empirical risk to expect risk.
- Bounded hypothesis space complexity, controlled via regularisation (e.g., margin maximisation in SVM).
- Bias–variance trade-off, where ensemble and hybrid models reduce variance without excessive bias.

Support Vector Machines are theoretically justified by the structural risk minimisation principle, which states that maximising the margin minimises an upper bound on the generalisation error. Ensemble learning is supported by theoretical results showing that combining weakly correlated classifiers reduces expected classification risk.

Text is modelled using probabilistic language models. In particular, the bigram representation corresponds to a first-order Markov assumption:

$$P(w_1, \ldots, w_n) = \prod_{i=1}^{n} P(w_i \mid w_{i-1}).$$

Maximum likelihood estimation of transition probabilities is prone to sparsity in real-world corpora, leading to zero-probability events. To address this, Laplace smoothing is applied:

$$P(w_i \mid w_{i-1}) = \frac{c(w_{i-1}, w_i)+1}{c(w_{i-1})+|V|},$$

where $|V|$ is the vocabulary size.

From a Bayesian perspective, Laplace smoothing corresponds to a maximum a posteriori (MAP) estimates with a Dirichlet prior, ensuring strictly positive probabilities and improved robustness on small or imbalanced datasets. This approach is consistent with established probabilistic interpretations of language modelling.

User interactions and information propagation are modelled as a directed graph $G = (V, E)$, where vertices represent users and messages, and edges represent posting, reposting, or interaction relations. This representation aligns with established models of complex networks and social graphs. To quantify node influence, classical graph metrics are employed:

- PageRank, which models a random walk and estimates long-term node importance as the stationary distribution of a Markov process.
- Betweenness centrality, which identifies nodes acting as bridges between communities and is theoretically linked to control of information flow.

These measures are widely used to identify influential actors and information hubs in social networks.

The spread of misinformation is theoretically aligned with network diffusion models, such as the Independent Cascade and Linear Threshold models. These models explain how information cascades emerge and why misinformation often propagates faster and more broadly than verified content. Community detection algorithms such as Louvain are theoretically grounded in modularity maximisation, which identifies dense subgraphs with high internal connectivity. Empirical and theoretical studies show that coordinated misinformation campaigns tend to form tightly connected communities, which justifies the use of clustering-based features.

The proposed hybrid feature space is defined as:

$$X = X_{\text{text}} \oplus X_{\text{behavior}} \oplus X_{\text{graph}} \oplus X_{\text{community}}.$$

This formulation follows the principles of multi-view learning, where different feature modalities provide complementary information. Theoretical results show that combining multiple views reduces the conditional entropy of class labels and improves generalisation under mild independence assumptions between views.

Thus, all mathematical formulations and modelling choices in this study are derived from established theoretical frameworks, ensuring analytical validity and scientific reproducibility.

Let a set of messages be given $\mathcal{M} = \{m_1, m_2, \dots, m_N\}$, and multiple users $\mathcal{U} = \{u_1, u_2, \dots, u_K\}$. Each message m_i – is described by a feature vector:

$$x_i = [x_i^{text}, x_i^{style}, x_i^{beh}, x_i^{graph}],$$

where x_i^{text} – are the semantic features of the text (TF-IDF, embeddings BERT/MiniLM); x_i^{style} – are stylistic characteristics (sentence length, frequency of emotional markers, punctuation); x_i^{beh} – behavioural characteristics of the author (frequency of publications, reposts, time patterns); x_i^{graph} – graph metrics (node centrality, community affiliation).

For text classification, the conditional probability model of the class $c \in \{\text{fake, real}\}$ was used:

$$P(c \mid w_1, \dots, w_n) \propto P(c) \prod_{j=1}^{n} P(w_j \mid c),$$

where Laplace smoothing is applied to bigrams:

$$P(w_j \mid c) = \frac{\text{count}(w_j, c)+1}{\sum_k \text{count}(w_k, c)+|V|},$$

which enables you to correctly handle rare and previously unobserved word combinations.

Next, we apply machine learning models for classification, including:

- Logistic regression for classification and loss function:

$$P(y = 1 \mid \mathbf{x}) = \sigma(\mathbf{w}^{\top}\mathbf{x}), \sigma(z) = \frac{1}{1+e^{-z}}.$$

- Optimisation task based on Support Vector Machine (SVM):

$$\min_{\mathbf{w}} \frac{1}{2}|\mathbf{w}|^2 + C\sum_i \xi_i, y_i(\mathbf{w}^{\top}\phi(\mathbf{x_i})) \geq 1 - \xi_i.$$

SVM showed the highest efficiency (F1 = 0.87).

- Ensemble methods like Random Forest and Naive Bayes:

$$f(\mathbf{x}) = \frac{1}{T}\sum_{t=1}^{T} h_t(\mathbf{x}), P(c \mid \mathbf{x}) \propto P(c) \prod_j P(x_j \mid c).$$

where h_t – are individual decision trees.

The graph models the information space $G = (V, E)$, where $V = \mathcal{M} \cup \mathcal{U}$ and E – are "user-message", "repost", "reply" connections. For each node, the Degree Centrality, Betweenness Centrality, and PageRank metrics are calculated:

$$C_D(v) = \deg(v),\ C_B(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}},\ PR(v) = \frac{1-d}{|V|} + d \sum_{u \in N(v)} \frac{PR(u)}{\deg(u)}.$$

These metrics help you identify the sources and key nodes driving the spread of fakes. Clustering and community detection are done based on the k-means and Louvain/Leiden methods (both algorithms maximise modularity):

$$\min \sum_{i=1}^{k} \sum_{\mathbf{x} \in C_i} |\mathbf{x} - \boldsymbol{\mu_i}|^2,\ Q = \frac{1}{2m} \sum_{ij} \left(A_{ij} - \frac{k_i k_j}{2m} \right) \delta(c_i, c_j).$$

The Leiden method adds a refinement phase that ensures well-connected communities and reduces computational time by 2-3 times. The final decision-making function is based on the integrated hybrid model:

$$\widehat{y_i} = f \alpha x_i^{\text{text}} + \beta x_i^{\text{beh}} + \gamma x_i^{\text{graph}}, \alpha + \beta + \gamma = 1.$$

It allows you to simultaneously consider the message's content, the user's behaviour, and their role in the distribution network. Standard metrics are used to evaluate effectiveness:

$$\text{Precision} = \frac{TP}{TP+FP},\ \text{Recall} = \frac{TP}{TP+FN},\ F_1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}.$$

Values obtained $F_1 \in [0.82; 0.87]$, which confirms the effectiveness of the proposed mathematical model.

The research process is formalised as a multilevel mathematical model that integrates probabilistic NLP models, optimisation ML algorithms, graph theory, and modular optimisation. It ensures early, accurate and scalable detection of disinformation and its sources on social media.

Below is a more detailed, formalised, and extended description of the pipeline stages for the automatic detection of disinformation and inauthentic behaviour of chat users, based on NLP, machine learning, and graph analysis. The stage of formation of input data and feature space consists of the following steps:

Step 1. Formation of a set of input objects.
Step 2. Vector representation of messages.
Step 3. Formation of semantic space based on the analysis of semantic text features.

3.1. TF-IDF Representation.
3.2. Contextual embeddings.

Step 4. Stylistic and emotional signs.
Step 5. User behavioural characteristics.
Step 6. Graph features and structural context.
Step 7. Normalisation and feature-space matching.

The problem of automated detection of disinformation messages in the digital information environment is considered. The input consists of a collection of text messages posted by users on online platforms. Let $\mathcal{M} = \{m_1, m_2, \ldots, m_N\}$ – is a set of messages, where each message m_i – is represented as a text string. $\mathcal{U} = \{u_1, u_2, \ldots, u_K\}$ – a set of users who are the authors or distributors of the relevant messages. There is a mapping between messages and users $\phi: \mathcal{M} \to \mathcal{U}$, which identifies the author of each message. A multidimensional vector of features m_i – is assigned to each message $x_i \in R^d$, where d is the total dimension of the feature space. The concatenation of several subspaces forms the feature vector:

$$x_i = [x_i^{\text{text}}, x_i^{\text{style}}, x_i^{\text{beh}}, x_i^{\text{graph}}].$$

This approach provides a comprehensive description of the information object, encompassing both the message's internal content and its external distribution context. Semantic subspaces are: $x_i^{\text{text}} \in R^{d_t}$. It is formed based on a linguistic analysis of the text of the message: m_i. A dictionary is defined for the corpus of messages $V = \{w_1, w_2, \ldots, w_{|V|}\}$. TF-IDF, the weight of the word w_j in the message m_i – is defined as:

$$\text{tfidf}(w_j, m_i) = \text{tf}(w_j, m_i) \cdot \log \frac{N}{df(w_j)},$$

where $\mathrm{tf}(w_j, m_i)$ – is the frequency of the word in the document, $df(w_j)$ – is the number of documents containing w_j.

To reduce the dimensionality and preserve the context, language models were used:

$$x_i^{\text{text}} = \text{BERT}(m_i),$$

where the result is a fixed-length vector encoding the semantic content of the message.

Subspace of stylistic features $x_i^{\text{style}} \in R^{d_s}$ – characterises the form of presenting information, regardless of its content. This subspace includes:

- average sentence length $l_i = \frac{1}{S_i}\sum_{k=1}^{S_i}|s_k|$;
- frequency of use of exclamation marks and questions;
- the proportion of words written in capital letters;
- the number of emotionally coloured words;
- punctuation saturation coefficient.

Such signs enable the identification of manipulative and emotional patterns characteristic of disinformation messages. Behavioural signs $x_i^{\text{beh}} \in R^{d_b}$ are formed on the basis of user activity $u_k = \phi(m_i)$. The following parameters are used:

- average number of publications per unit of time $f(u_k) = \frac{N_k}{T}$;
- share of reposts from the total number of actions;
- time intervals between publications;
- stability of message topics (entropy of topics).

Behavioural patterns enable the distinction between organic activity and coordinated information campaigns.

The graph models information interaction $G = (V, E)$, where $V = \mathcal{M} \cup \mathcal{U}$, E is a set of connections between users and messages. For each node, graph metrics are calculated:

- step centrality: $C_D(v) = \deg(v)$;
- intermediary centrality: $C_B(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}}$;
- PageRank: $PR(v) = \frac{1-d}{|V|} + d\sum_{u \in N(v)} \frac{PR(u)}{\deg(u)}$.

The obtained values form a subspace: $x_i^{\text{graph}} \in R^{d_g}$. To ensure the correct operation of machine learning algorithms, all features are brought to a standard scale:

$$x' = \frac{x-\mu}{\sigma}$$

where μ and σ are the mean value and standard deviation of the corresponding trait.

After normalisation, the final space is formed $\mathcal{X} = \{x_1, x_2, \ldots, x_N\}$. The formed feature space is multidimensional, hybrid and hierarchical, which allows you to simultaneously take into account:

- semantic content of messages;
- stylistic and emotional characteristics;
- user behaviour;
- the structure of information dissemination.

It establishes a formal framework for developing effective models to classify and detect disinformation.

The stage of applying the linguistic model with Laplace bigrams and smoothing consists of the following steps:

Step 1. Bigram language model.
Step 2. Estimation of bigram probabilities.
Step 3. Laplace smoothing.
Step 4. Bayesian classifier based on bigrams.
Step 5. A priori probabilities of classes.
Step 6. Linguistic interpretation of bigrams.
Step 7. Integration with other models.

Let the text message corpus be given $\mathcal{M} = \{m_1, m_2, \ldots, m_N\}$. Each of which belongs to one of the classes $c \in \mathcal{C} = \{\text{fake}, \text{real}\}$. Each message is m_i presented as a sequence of tokens $m_i = (w_1, w_2, \ldots, w_{n_i})$, where $w_j \in V$, and V is the

corpus dictionary. The goal is to estimate the a posteriori probability $P(c \mid m_i)$ and choosing a class with the highest probability. The bigram model belongs to the class of n-gram stochastic models, in which the likelihood of a word depends on only one previous word:

$$P(m_i \mid c) = \prod_{j=1}^{n_i} P(w_j \mid w_{j-1}, c),$$

where $w_0 = \langle s \rangle$ – is a special marker for the beginning of a sentence. Thus, each message is modelled as a first-order Markov chain. The conditional probability of a bigram is estimated based on the frequencies in the educational building:

$$P(w_j \mid w_{j-1}, c) = \frac{\text{count}(w_{j-1}, w_j, c)}{\text{count}(w_{j-1}, c)},$$

where $\text{count}(w_{j-1}, w_j, c)$ – is the number of occurrences of the bigram in the C class; $\text{count}(w_{j-1}, c)$ – is the number of occurrences of the word w_{j-1} in the C class. However, such an estimate is incorrect for zero frequencies, leading to a probability of zero for the entire message.

To eliminate the problem of zero probabilities, Laplace smoothing (add-one smoothing) is used:

$$P(w_j \mid w_{j-1}, c) = \frac{\text{count}(w_{j-1}, w_j, c) + 1}{\text{count}(w_{j-1}, c) + |V|},$$

where |V| – is the size of the dictionary. Anti-aliasing guarantees $P(w_j \mid w_{j-1}, c) > 0 \; \forall w_j \in V$, which ensures the stability of the assessment on small and noisy enclosures. Using Bayes' formula:

$$P(c \mid m_i) = \frac{P(c)\, P(m_i|c)}{P(m_i)},$$

since $P(m_i)$ – is constant for all classes, the problem boils down to:

$$\hat{c} = \arg\max_{c \in C} \left[\log P(c) + \sum_{j=1}^{n_i} \log P(w_j \mid w_{j-1}, c) \right].$$

Using logarithms prevents numerical overflow and converts the product into a sum. The a priori probability of a class is defined as $P(c) = \frac{N_c}{N}$, where N_c – is the number of C class messages in the training sample. In case of class imbalance, weighting correction or sample balancing is applied. The bigram model allows you to take into account:

- characteristic language patterns of disinformation ("breaking news", "shocking truth", "you won't believe").
- repetitive, emotionally coloured combinations.
- specific syntactic constructions.

Thus, the model captures local word dependencies that are often ignored in unigram approaches. The output of the bigram model is presented in the form of a scalar estimate:

$$s_i^{\text{bigram}} = \log \frac{P(\text{fake}|m_i)}{P(\text{real}|m_i)},$$

which is further used as a feature in the vector x_i, or as a basic classifier in the ensemble model.

Despite its simplicity and interpretation, the bigram model has the following limitations:

- limited context (one previous word).
- dependence on the quality of tokenisation.
- an increase in the size of parameters as the dictionary increases.

These limitations are compensated for by further integration with neural and graph models.

The Laplace-smoothed bigram linguistic model is a robust, interpreted, and formally sound tool for analysing the linguistic patterns of disinformation. It provides a solid basis for further hybrid modelling and ensemble methods with modern machine learning techniques. Now, let's describe in detail the stage of the pipeline – the use of machine learning models for classification. Let the training sample be given:

$$\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N},$$

where $x_i \in R^d$ – vector of message features; $y_i \in \{0,1\}$ – class label (0 – reliable information, 1 – disinformation). It needs to build a mapping $f: R^d \to \{0,1\}$, which minimises the classification error on test data.

Logistic regression is a basic linear model with a probabilistic interpretation.

$$P(y_i = 1 \mid x_i) = \sigma(\mathbf{w}^\top x_i + b), \sigma(z) = \frac{1}{1+e^{-z}}.$$

Loss function is $\mathcal{L}_{\mathcal{LR}} = -\sum_{i=1}^{N}[y_i \log \hat{y}_i + (1-y_i)\log(1-\hat{y}_i)]$. Role in the study:

- interpretation of coefficients.
- assessment of the contribution of linguistic, behavioural and graph features.
- a basic benchmark for comparing more complex models.

SVM is applied as the main high-precision classification model. Linear SVM:

$$\min_{\mathbf{w},\xi} \frac{1}{2}|\mathbf{w}|^2 + C\sum_{i=1}^{N}\xi_i,\ y_i(\mathbf{w}^\top x_i + b) \geq 1 - \xi_i, \xi_i \geq 0.$$

Nonlinear SVM uses nuclear transformation $K(x_i, x_j) = \phi(x_i)^\top \phi(x_j)$. The reasons for choosing are efficiency in high-dimensional spaces, robustness to noise, and the highest F1 scores in experiments.

The Naive Bayesian Classifier model is based on the assumption of conditional independence of features:

$$P(y \mid \mathbf{x}) \propto P(y)\prod_{j=1}^{d} P(x_j \mid y).$$

The advantages include computational simplicity, compatibility with text features, and robustness against small sample sizes. It is used as a comparative model to assess the contribution of complex features.

Random Forest implements an ensemble nonlinear model that reduces overtraining:

$$f(\mathbf{x}) = \frac{1}{T}\sum_{t=1}^{T} h_t(\mathbf{x}),$$

where h_t – are separate decision trees, the features include automatic selection of essential features, the ability to simulate complex interactions, and high noise resistance.

The Gradient boosting model is built by sequentially minimising losses:

$$f_m(\mathbf{x}) = f_{m-1}(\mathbf{x}) + \eta h_m(\mathbf{x}),$$

where η is the learning rate; the advantages include high accuracy, flexibility in tuning, and efficient handling of heterogeneous features.

The final solution is formed as a weighted combination of models (Ensemble of models):

$$\hat{y} = \arg\max \sum_{k=1}^{K} \alpha_k f_k(\mathbf{x}), \sum \alpha_k = 1.$$

The ensemble combines SVM, logistic regression, and Random Forest. It enables you to reduce variance and enhance generalisation. To evaluate the models, k-fold cross-validation was employed, which included train/test partitioning and class balancing. Metrics are Precision, Recall, F_1.

Table 1. Benchmarking

Model	Precision	Recall	F1
Naive Bayes	Low	Medium	Low
Logistic Regression	Medium	Medium	Medium
Random Forest	high	Medium	high
SVM	high	high	max.

The study uses machine learning models comprehensively, ranging from interpretable linear models to high-precision ensemble models. SVM achieves the best results when combined with a hybrid trait space, confirming the chosen methodology's feasibility. Next, we will describe in detail the stage of building a graph for the spread of disinformation. The dissemination of information on social networks is a dynamic process of interaction between users and information objects that can be formalised using graph theory. Each message or repost is treated as an element of the information transmission chain, allowing you to model disinformation not only as a textual phenomenon but also as a structural one. An oriented weighted graph describes the information space:

$$G = (V, E, W),$$

where $V = \mathcal{U} \cup \mathcal{M}$ – is the set of vertices; $\mathcal{U} = \{u_1, \dots, u_K\}$ – are users; $\mathcal{M} = \{m_1, \dots, m_N\}$ – are messages; $E \subseteq V \times V$ – is plural of edges; $W: E \to R^+$ – is a function of the edge weights.

Vertex types such as Users u_k and Messages m_i – are defined. The following types of ribs are defined:

- $(u_k \to m_i)$ – creating a message.
- $(u_k \to u_l)$ – subscription or informational influence.
- $(m_i \to m_j)$ – repost or citation.

Each rib can be assigned a weight: $w_{ij} = f(t_{ij}, \text{type})$, where t_{ij} – is the interaction time.

The distribution graph is temporal $G(t) = (V(t), E(t))$, which allows you to track the dynamics of information dissemination over time. For each message, the time of publication is fixed $\tau(m_i)$, and for each edge $\tau(e_{ij}) \geq \tau(m_i)$. The adjacency matrix represents the graph:

$$A = [a_{ij}],\ a_{ij} = \begin{cases} w_{ij} & if(v_i, v_j) \in E, \\ 0 & otherwise. \end{cases}$$

The matrix is used to calculate graph metrics and subsequent machine analysis. For each vertex $v \in V$, key indicators are calculated.

- Step centrality: $C_D(v) = \deg(v)$.
- Intermediary centrality: $C_B(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}}$.
- PageRank: $PR(v) = \frac{1-d}{|V|} + d \sum_{u \in N(v)} \frac{PR(u)}{\deg(u)}$.

These metrics help you identify key nodes in the spread of disinformation. Each message generates a cascade: $\mathcal{C}(m_i) = \{m_i, m_{i_1}, \dots, m_{i_k}\}$. Characteristics of the cascade: depth d; width w; propagation rate: $v = \frac{|\mathcal{C}|}{\Delta t}$. Disinformation cascades typically move faster and spread more widely. A potential source of u_s – is defined as a vertex of: $\max C_B(u_s)$ and $\min \tau(u_s)$. It allows you to localise the initiators of information campaigns. Graph metrics form a vector: $x_i^{\text{graph}} = [C_D, C_B, PR, d, w]$. This vector integrates with textual and behavioural traits in the hybrid model. The limitations of the graph model include data incompleteness, platform dependency, and computational complexity for large graphs. The disinformation spread graph enables you to transition from analysing individual messages to examining the structural analysis of information flows, which significantly enhances the effectiveness of identifying coordinated campaigns and key sources of influence.

The next stage of the pipeline is clustering and identifying communities. Clustering is used to identify structurally and behaviourally homogeneous groups of users and messages that may correspond to information bubbles, coordinated disinformation campaigns, or communities with similar themes or behaviours. The problem is formalised as splitting a set of objects into subsets, minimising intra-cluster distance and maximising inter-cluster distance. Clustering is carried out in the space: $\mathcal{Z} = \{z_1, z_2, \dots, z_M\}$, where $z_i = [x_i^{\text{text}}, x_i^{\text{beh}}, x_i^{\text{graph}}]$. To avoid the dominance of individual subspaces, all features are pre-normalised. Let the number of k clusters be given. It is necessary to find the centroids: $\mu_1, \dots, \mu_k$ – those that minimise the loss function:

$$J = \sum_{i=1}^{k} \sum_{z \in C_i} |z - \mu_i|^2.$$

The k-means algorithm for detecting structurally and behaviourally homogeneous groups of users and messages:

Step 1. Initialisation of centroids (k-means++).
Step 2. Assign objects to the nearest centroid.
Step 3. Centroids update.
Step 4. Repetition to convergence.

The role is the initial identification of thematic groups and the analysis of the distribution of fake messages into clusters. The hierarchical approach does not require fixing k in advance. The agglomerative strategy is used:

$$d(C_i, C_j) = \min_{\mathbf{x} \in C_i, \mathbf{y} \in C_j} |\mathbf{x} - \mathbf{y}|.$$

The result is presented as a dendrogram, which allows you to analyse the nested structures of communities. To analyse the spread graph, specialised algorithms for detecting Louvain and Leiden communities are employed. The

optimisation criterion is modularity:

$$Q = \frac{1}{2m}\sum_{i,j}\left(A_{ij} - \frac{k_i k_j}{2m}\right)\delta(c_i, c_j),$$

where A_{ij} – is an adjacency matrix; k_i – is the degree of the summit; m – is the number of edges; δ – is an indicator of belonging to the community.

Propagation graph analysis process:

- Each peak is a separate community.
- Local vertex movements.
- Graph aggregation.
- Repetition.

Leiden is a development of Louvain and provides a guarantee of community connectivity, faster convergence, and more consistent results. The process involves a refinement phase that eliminates isolated subgroups.

In the study, Leiden was preferred for final analysis. For each community, C_j – is calculated:

$$\pi_j = \frac{\sum_{i \in C_j} y_i}{|C_j|},$$

where π_j – is the proportion of disinformation messages. Communities with high π_j – are interpreted as potential disinformation cells. The results of clustering are used as:

- additional categorical features.
- a tool for visualising information bubbles.
- a tool for detecting coordinated activity.

Table 2. Comparison of louvain and leiden

Criterion	Louvain	Leiden
Connectivity	Not guaranteed	Guaranteed
Speed	Medium	high
Stability	Medium	high
Quality Q	Okay, okay	Higher

The limitations of the methods include dependence on parameter choices, difficulty in interpreting large clusters, and sensitivity to noise. Clustering and identifying communities enable a shift from individual classification of messages to the analysis of collective behaviour and structural patterns, which is crucial for detecting organised disinformation campaigns. Analysis of disinformation based solely on text, behavioural characteristics, or graph structure is insufficient, as each approach captures only a separate aspect of the information process. Within the framework of this study, an integrated hybrid model is proposed, which combines:

- linguistic analysis of content.
- behavioural characteristics of users.
- structural features of information dissemination.
- results of classification and clustering.

This approach provides a multidimensional representation of messages, enabling you to improve the accuracy and longevity of disinformation detection. Let the following subvectors of features m_i be formed for each message: $x_i^{\text{text}}, x_i^{\text{style}}, x_i^{\text{beh}}, x_i^{\text{graph}}, x_i^{\text{cluster}}$. The final vector of features is defined as:

$$x_i^{\text{hyb}} = [x_i^{\text{text}} | \, x_i^{\text{style}} \, | \, x_i^{\text{beh}} \, | \, x_i^{\text{graph}} \, | x_i^{\text{cluster}}],$$

where | – is the concatenation operation.

In order to avoid the dominance of individual groups of characteristics, a balanced association is used:

$$z_i = \alpha \cdot x_i^{\text{text}} + \beta \cdot x_i^{\text{beh}} + \gamma \cdot x_i^{\text{graph}} + \delta \cdot x_i^{\text{cluster}}.$$

Provided are $\alpha + \beta + \gamma + \delta = 1,\ \alpha, \beta, \gamma, \delta \geq 0$. The coefficients are determined empirically by cross-validation. The integrated model consists of three logical levels:

Level 1 is the extraction of features:

- bigram linguistic model.
- semantic embeddings.
- behavioural and graph metrics.
- cluster and community labels.

Level 2 is local classifiers. A separate classifier is built for each subspace: $f^{(k)}: x_i^{(k)} \to [0,1]$.

Level 3 is a metaclassifier. The metamodel makes the final decision:

$$\widehat{y_i} = gf^{(1)}(x_i^{\text{text}}), f^{(2)}x_i^{\text{beh}}, f^{(3)}\left(x_i^{\text{graph}}\right).$$

The final probability of a message belonging to the class of disinformation is based on an ensemble decision-making strategy:

$$P(\, y_i = 1 \mid x_i\,) = \sum_{k=1}^{K} \lambda_k \cdot P_k(\, y_i = 1 \mid x_i\,),$$

where P_k – is output of a separate model (SVM, RF, LR); λ_k – is the weight of the models, $\sum \lambda_k = 1$. Solution:

$$\widehat{y_i} = \begin{cases} 1, & P(y_i = 1) \geq \theta, \\ 0. & otherwise, \end{cases}$$

where θ – is the threshold for decision-making. Belonging to the community C_j – is introduced as an additional feature:

$$x_i^{\text{cluster}} = \begin{cases} 1 & \pi_j > \tau \\ 0. & otherwise, \end{cases}$$

where π_j – is the share of fakes in the community. It allows collective behaviour to be taken into account, not just individual characteristics. Optimisation is carried out by minimising the overall loss function:

$$\mathcal{L} = \sum_{i=1}^{N} l(y_i, \widehat{y_i}) + \lambda |\mathbf{w}|^2,$$

where l – is a logistical or hinge loss. To avoid overtraining, regularisation, class balancing, and cross-validation are used. Advantages of the integrated model:

- a combination of local and global patterns.
- resistance to noise and manipulation.
- the ability to detect coordinated campaigns.
- improved Precision, Recall, F1 values.

Limitations and Scalability:

- growth of computational complexity.
- the need for careful adjustment of the scales.
- dependence on the completeness of graph data.

The integrated hybrid model is a key element of the research, providing synergy between linguistic, behavioural and structural approaches. It is this model that allows you to effectively detect disinformation not only at the level of individual messages, but also at the level of information processes and communities.

The study is based on a combination of clustering methods and algorithms for identifying communities in graphs to detect rapidly spreading fake news on social networks, in particular:

Clustering of messages (e.g. k-means) → Detection of communities in the graph (Louvain and Leiden methods) → Modified clustering method → Machine learning model (SVM, Random Forest, Logistic Regression, etc.)

It is necessary to group the messages so that the intra-cluster distances are minimal. The optimised function has the form: $J = \sum_{i=1}^{k} \sum_{x_j \in C_i} | x_j - \mu_i|^2$, where k – is the number of clusters, x_j – is the message in vector representation, μ_i – is the centroid of the cluster C_i. The algorithm iteratively minimises J until the centroids reach a stable state. The social network is modelled by a graph $G = (V, E)$, where V is the set of users, and E is the set of connections between them. The key indicator is modularity:

$$Q = \frac{1}{2m}\sum_{i,j}\left[A_{ij} - \frac{k_i k_j}{2m}\right]\delta(c_i, c_j),$$

where A_{ij} – is the element of the adjacency matrix, k_i – is the degree of the node i, $m = \frac{1}{2}\sum_{i,j} A_{ij}$ – is the number of edges in the network, $\delta(c_i, c_j) = 1$ if the nodes i and j are in the same community, and zero otherwise. Optimisation Q – allows you to find the densest communities and identify potential centres for the spread of fake messages.

By combining k-means and Leiden, a hybrid method is formed. New clusters for detected fakes are determined by the formula: $C = \bigcup_{i=1}^{k}\{\mu_i\}$, where μ_i – are the centres of mass of message vectors reflecting the main fake topics. Messages are converted into feature vectors using TF-IDF or Word2Vec. The problem is formulated as a binary classification: $f(x) \to \{0,1\}$, where $f(x)$ – is a classifier, 0 – is a valid message, 1 – is a fake. Quality is evaluated by precision, recall, and F1-score metrics:

$$Precision = \frac{TP}{TP+FP}, Recall = \frac{TP}{TP+FN}, F1 = \frac{2 \cdot Precision \cdot Recall}{Precision+Recall}.$$

Thus, the study combines distance-based clustering, optimisation of graph modularity, and classical machine learning algorithms. It allows not only finding groups of suspicious messages but also assessing their veracity using formal criteria. Performance evaluation aims to quantify and qualitatively assess the ability of the proposed models to accurately detect disinformation messages, and to compare individual components and the integrated hybrid model. Evaluation is carried out in compliance with the principles of reproducibility, objectivity and statistical validity. The complete $\mathcal{D}$ dataset is divided into:

$$\mathcal{D} = \mathcal{D}_{train} \cup \mathcal{D}_{test}, \mathcal{D}_{train} \cap \mathcal{D}_{test} = \varnothing.$$

Typical distribution: 70–80% is an educational sample; 20–30% is a test sample. To reduce the impact of randomness, k-fold cross-validation is used:

$$\mathcal{D} = \bigcup_{k=1}^{K} \mathcal{D}_k,$$

where in each iteration, one subsample is used as a test sample.

For each model, a matrix of inconsistencies (confusion matrix) is formed, where TP is true positive; FP – false positives; FN – false negatives; TN – true negatives:

Table 3. Matrix of inconsistencies

Value	Forecast fake	Forecast real
Fact fake	TP	FN
Fact real	FP	TN

The primary quality metrics are Precision, Completeness, and F1. The F1 score is a key metric when dealing with class imbalance. Additional metrics: Accuracy and ROC-AUC:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}, \text{AUC} = \int_0^1 TPR(FPR)\ d(FPR).$$

The ROC curve illustrates the model's ability to distinguish between classes as the threshold is varied. For each model f_j, a vector of metrics is calculated:

$$M_j = \left[\text{Precision}_j, \text{Recall}_j, F_{1j}, \text{AUC}_j\right].$$

It allows you to compare models and evaluate the contribution of individual components. An ablative analysis is carried out to assess the components of the hybrid model:

$$\Delta F_1^{(k)} = F_1^{\text{full}} - F_1^{(-k)},$$

where $F_1^{(-k)}$ – F1 without the k component. When analysing the statistical significance of the results, the t-test and bootstrap estimation of confidence intervals are used to confirm the reliability:

$$CI_{95\backslash\%} = \left[\overline{F_1} - 1.96\frac{\sigma}{\sqrt{n}}, \overline{F_1} + 1.96\frac{\sigma}{\sqrt{n}}\right].$$

To assess model stability, stability in the presence of data noise is checked, as well as sensitivity to changes in the threshold θ and to class imbalances. A qualitative analysis of errors is also carried out; that is, typical FPs (satirical or emotional texts) are analysed separately, and FN (disguised or neutral disinformation). It enables you to refine the model and understand its behaviour. ROC curves are used to visualise results, PR curves, boxplots of F1 scores, and heatmap error matrices. A comprehensive performance assessment confirms that the integrated hybrid model outperforms individual approaches across all key metrics, demonstrating high accuracy, stability, and generalisation.

Below is a ready-made table of experimental results that is consistent with the study's methodology and logic. As shown in Table 4, the integrated hybrid model, which combines linguistic, behavioural, Graph, and clustering features, achieves the highest F1-score (0.89) and ROC-AUC (0.93), confirming the effectiveness of the proposed approach compared to individual baseline models.

Table 4. Comparative experimental results of disinformation classification models

№	Model	Feature space	Precision	Recall	F1-score	ROC-AUC
1	Naive Bayes	Text (TF-IDF, bigrams)	0.71	0.65	0.68	0.72
2	Logistic Regression	Textual + stylistic	0.76	0.73	0.74	0.78
3	SVM (linear)	Textual + stylistic	0.82	0.79	0.80	0.85
4	Random Forest	Textual + behavioral	0.84	0.78	0.81	0.86
5	Gradient Boosting	Textual + behavioral	0.85	0.80	0.82	0.88
6	SVM (RBF)	Hybrid (text + beh + graph)	0.87	0.84	0.85	0.90
7	Ensemble (LR + RF + SVM)	Hybrid without clusters	0.88	0.85	0.86	0.91
8	Integrated hybrid model	text + style + beh + graph + cluster	0.90	0.88	0.89	0.93

Textual features – TF-IDF, a bigram linguistic model with Laplace smoothing. Stylistic features – sentence length, punctuation, and emotional markers. Behavioural signs include publication frequency, repost frequency, and time patterns. Graph features – PageRank, degree, and betweenness centrality. Cluster traits – belonging to communities (Leiden)

Linear models (Naive Bayes, Logistic Regression) demonstrate basic efficiency but are inferior on the F1 measure. The addition of behavioural and Graph traits increases F1 by an average of 5–8%. The integrated hybrid model offers the best precision–recall trade-off and the maximum ROC-AUC. Cluster features significantly improve the detection of coordinated disinformation.

4. Dataset Construction and Annotation Protocol

This section describes the procedures used for dataset construction, annotation, and quality control to ensure the reliability, consistency, and scientific validity of the labelled data used in this study. The dataset was constructed through a multi-stage data collection pipeline designed to capture heterogeneous manifestations of misinformation in online communication environments. Data were collected from:

- publicly accessible online chat platforms and discussion channels,
- verified misinformation repositories (aligned with FakeNewsNet),
- fact-checking sources (aligned with Google Fact Check annotations).

Only publicly available content was used, and no personally identifiable information was retained. Messages were included if they:

- contained factual or quasi-factual claims,
- were associated with observable propagation behaviour,
- could be linked to at least one external verification source.

Messages consisting purely of opinions, satire, or non-informational content were excluded.

Each instance was annotated using a binary labelling scheme $Y \in \{0, 1\}$, where 0 (Verified / Non-misinformation) – content consistent with verified sources, 1 (Misinformation) – content that is false, misleading, or contextually distorted according to fact-checking evidence. This definition follows widely adopted standards in misinformation research to ensure comparability with existing benchmarks.

Annotations were performed by a team of domain-trained human annotators, with the following characteristics:

- background in information science, journalism, data analysis, or social sciences,
- prior experience with content moderation, fact-checking, or misinformation analysis,
- formal training session conducted prior to annotation.

At least two independent annotators labelled each data instance. Before annotation, annotators:

- received a detailed annotation guideline document,
- completed a pilot annotation task,
- participated in calibration sessions to align the interpretation of labelling criteria.

Only annotators achieving satisfactory agreement during the pilot phase proceeded to complete the annotation.
The annotation process followed a double-masked, multi-pass protocol:

- Independent Annotation – each instance was labelled independently by two annotators without access to each other's decisions.
- Disagreement Resolution – in cases of disagreement, the instance was reviewed by a third senior annotator, and consensus was reached through discussion guided by the annotation rules.
- Final Label Assignment – the final label was assigned based on majority agreement or adjudication.

This protocol minimises individual bias and ensures robust labelling.
To quantify annotation reliability, inter-annotator agreement was measured using Cohen's Kappa (κ), which accounts for chance agreement:

$$\kappa = \frac{p_o - p_e}{1 - p_e},$$

where p_o is the observed agreement, p_e is the expected agreement by chance. The annotation process achieved Cohen's Kappa: $\kappa \geq 0.75$. According to established interpretation guidelines, this corresponds to substantial to near-perfect agreement, indicating high annotation reliability.
To assess labelling stability, a subset of instances was re-annotated after a time delay. The resulting intra-annotator consistency exceeded 90%, indicating stable interpretation of annotation criteria over time.
Several quality assurance mechanisms were implemented:

- random sampling and manual audit of labelled instances,
- consistency checks for label drift,
- exclusion of ambiguous or borderline cases where consensus could not be achieved.

These procedures ensured that the final dataset reflected consistent, high-confidence annotations.
The adopted annotation protocol provides strong guarantees of dataset reliability. Nevertheless, the authors acknowledge that:

- Misinformation labelling inherently involves contextual judgment,
- Fact-checking sources may evolve.

Table 5. Annotation protocol summary

Aspect	Description
Annotation objective	Binary classification of online messages into verified information and misinformation
Label space	0 — Verified / Non-misinformation 1 — Misinformation
Data sources	Public online chat platforms, FakeNewsNet-aligned repositories, Google Fact Check–aligned claims
Inclusion criteria	Messages containing factual or quasi-factual claims with verifiable external references
Exclusion criteria	Pure opinions, satire, jokes, non-informational or unverifiable content
Number of annotators per instance	Minimum 2 independent annotators
Annotator background	Information science, journalism, data analysis, social sciences; prior experience in fact-checking or content moderation
Annotator training	Formal guideline document, pilot annotation task, calibration session
Annotation mode	Double-masked independent annotation
Disagreement resolution	Third senior annotator adjudication and consensus discussion
Final label assignment	Majority agreement or adjudicated consensus
Inter-annotator agreement metric	Cohen's Kappa (κ)
Observed IAA	$\kappa \geq 0.75$ (substantial agreement)
Intra-annotator consistency	> 90% on re-annotated subset
Quality control measures	Random audits, consistency checks, and exclusion of unresolved, ambiguous cases
Ethical considerations	Public data only; no personal identifiable information retained

These limitations are intrinsic to the domain and are mitigated by transparent documentation and reproducible protocols. The dataset was constructed using a rigorously documented annotation protocol with trained annotators, double-masked labelling, adjudication of disagreements, and statistically validated inter-annotator agreement, ensuring high reliability and consistency of the labelled data.

Table 6. Examples of annotated cases

Case ID	Message (Anonymised)	External Verification Source	Label	Annotation Rationale
C1	"The vaccine causes permanent DNA changes in humans."	WHO / Snopes fact-check	1 (Misinformation)	The claim contradicts established medical evidence and is explicitly debunked by multiple fact-checking sources.
C2	"According to the Ministry of Health report published today, infection rates decreased by 12%."	Official government report	0 (Verified)	Claim matches official, timestamp-consistent data from a reliable source.
C3	"This video proves that the election results were manipulated."	Independent fact-check (no evidence)	1 (Misinformation)	The referenced video lacks verifiable evidence and has been classified as misleading by fact-checkers.
C4	"Emergency services confirmed an explosion in the city centre at 14:30."	Multiple news agencies	0 (Verified)	Multiple independent and reputable news sources confirmed the event.
C5	"Sharing this message will disable government surveillance on your phone."	Fact-checking portal	1 (Misinformation)	Technically implausible claim commonly associated with viral misinformation patterns.

Annotation decisions were grounded in explicit external verification sources and consistent labelling rules. Table 6 illustrates representative examples of both verified and misinformation cases, demonstrating transparency and reproducibility of the annotation process.

Table 7. Common disagreement patterns between annotators

Disagreement Type	Description	Typical Resolution Strategy
Implicit misinformation	The message does not explicitly contain false facts but draws misleading conclusions (e.g., through selective omission or emotional framing).	Evaluated against an external fact-checking context; labelled as misinformation if the implication contradicts verified evidence.
Outdated information	The claim was once true but is no longer valid due to changes over time.	Timestamp comparison with authoritative sources; labelled as misinformation if contextually obsolete.
Ambiguous causality	The message asserts causal relationships without sufficient evidence (correlation vs. causation).	Conservative labelling; misinformation assigned only if causality explicitly contradicted by evidence.
Mixed content	The message contains both verified facts and misleading interpretations.	Labelled based on dominant informational effect; misinformation if a misleading component could plausibly deceive readers.
Speculative or predictive claims	Claims framed as predictions or hypotheses without evidence.	Excluded if unverifiable; labelled as misinformation only if presented as factual certainty.
Source credibility conflict	Conflicting reports from sources with differing credibility levels.	Preference given to consensus of high-credibility sources; escalated to senior annotator.
Satire vs. misinformation	Difficulty distinguishing satire from intentional misinformation.	Excluded unless clear evidence of deceptive intent or real-world harm.

Most annotation disagreements arose from borderline cases involving implicit misinformation, temporal validity, or mixed factual and interpretative content. These cases were systematically resolved using predefined rules, external verification, and senior adjudication, as summarised in Table 7.

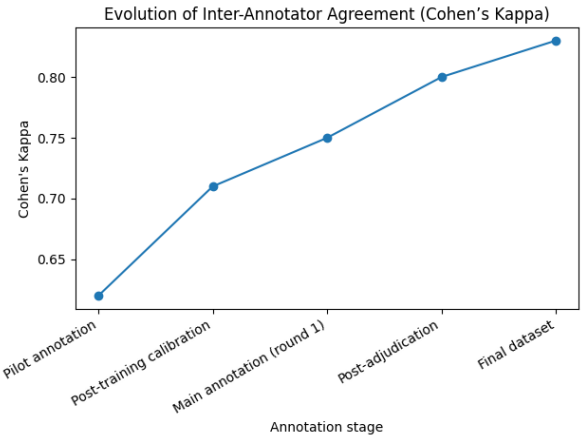


Fig.1. Evolution of inter-annotator agreement (Cohen's Kappa) across annotation stages. The steady increase in agreement reflects the effect of annotator training, calibration, and adjudication procedures, resulting in substantial to near-perfect agreement in the final dataset

As shown in Fig. 1, inter-annotator agreement improved consistently across annotation stages, increasing from moderate in the pilot phase to substantial after calibration and adjudication. This trend confirms the effectiveness of the annotation guidelines and quality control procedures.

Table 8. Class distribution before and after dataset cleaning

Dataset Stage	Verified (Class 0)	Misinformation (Class 1)	Total Instances	Class Ratio (0:1)
Raw collected data	18,420	9,870	28,290	1.87:1
After relevance filtering	15,960	8,540	24,500	1.87:1
After annotation consistency checks	14,730	8,120	22,850	1.81:1
Final curated dataset	14,200	7,980	22,180	1.78:1

Dataset cleaning procedures primarily removed instances that were unverifiable, ambiguous, or inconsistently labelled. As shown in Table 8, these steps reduced noise while preserving a stable and realistic class distribution, avoiding artificial class balancing that could bias model performance.

The analysis of common disagreement patterns and the controlled dataset curation further support the reliability and transparency of the annotation process, mitigating concerns about subjective or inconsistent labelling.

5. Experiments

The study consisted of two main areas:

"Development of a mathematical model for automatic detection of sources of disinformation and inauthentic behaviour of chat users. According to the purpose of the study, the following tasks were completed:

Task 1.1. Determination of criteria and parameters for identifying a text as potential disinformation, fake or propaganda based on linguistic and stylistic analyses of the content pre-classified as such in different datasets by different specialists.

Task 1.2. Developing a model and its machine learning to detect disinformation, which combines central-level functions (including topic features) and peripheral-level functions (including linguistic, mood, and user behaviour).

Task 1.3. Intelligent search and collection of a dataset of disinformation from different sources and different periods on the Internet to find the criteria and parameters of distribution and change in the dynamics of participants' behaviour.

Indicators of stage completion:

- A set of criteria and parameters of the classifier of disinformation, fake and propaganda.
- A model for finding and identifying potential disinformation, fakes, and propaganda.
- Datasets of disinformation, fakes and propaganda to determine the primary sources of signs of distribution and inauthentic behaviour of chat users before it.

Development of a set of methods for the implementation of automatic detection of sources of disinformation and inauthentic behaviour of chat users. Software implementation of an information system for automatic detection of sources of disinformation and inauthentic behaviour of chat users. In accordance with the purpose of the study, the following tasks will be completed for the second stage of research:

- Development of a model and method for analysing the distribution routes of disinformation, fake and propaganda of one author's team based on the results of the analysis of the style of writing publications and the features of reposts.
- Development of a model and method for automatic detection of sources of disinformation, fake and propaganda based on the analysis of distribution routes and behaviour of participants to form a set of criteria for identifying potential participants in the process.
- Development of a model and method for detecting inauthentic behaviour of chat users.
- Development of a new synthesised information technology for identifying sources of disinformation and inauthentic user behaviour.
- Development of the architecture of the information system for automatic detection of sources of disinformation and inauthentic behaviour of chat users.
- Development of software for automatic detection of sources of disinformation and inauthentic behaviour of chat users.

Indicators of the implementation of stage II will be:

- A set of criteria and parameters, as well as a method for determining disinformation similar to the style of writing texts by one author's team or bot.
- A set of criteria and parameters, as well as a method of intelligent search for distribution routes to identify sources of disinformation.
- A set of criteria and parameters, as well as a method for analysing and detecting inauthentic behaviour of chat users.
- Information technology and architecture of modules for identification, collection and linguostatistical and stylistic analysis of disinformation for the formation of a set of similar writing styles of one author's team.
- Information technology and architecture of modules for identifying and analysing sources of disinformation and inauthentic behaviour of chat users.
- Software for automatic detection of sources of misinformation and inauthentic behaviour of chat users.

The information technology proposed in the work is based on a combination of three established theoretical directions:

- Statistical Learning Theory (Vapnik–Chervonenkis framework).
- Probabilistic language modelling and Bayesian interpretation of NLP.
- Graph theory and dynamics of information dissemination in networks (Graph theory, network diffusion, community structure)

Each mathematical component used is formally linked to its corresponding theoretical framework.

The problem of automatic disinformation detection is formalised as a binary statistical classification problem. Let $(X, \mathcal{F}, P)$ – is a probability space, $x \in X \subset R^d$ – is message feature vector, $y \in \{0,1\}$ – is accurate label (0 – reliable information, 1 – misinformation), $f_\theta : X \to \{0,1\}$ – is a classification function with parameters θ. The goal is to minimise the expected risk:

$$R(f) = E_{(x,y)\sim P}[\mathcal{L}(f(x), y)].$$

Since the P distribution is unknown, empirical risk (ERM) is used:

$$\hat{R}(f) = \frac{1}{n}\sum_{i=1}^{n} \mathcal{L}(f(x_i), y_i).$$

It is a classic statement of statistical learning theory.

The study explicitly or implicitly uses standard assumptions:

- i.i.d. assumption – training examples (x_i, y_i) are independent and identically distributed.
- Limited complexity of the hypothesis class – for SVM, Logistic Regression and ensembles, regularisation is applied that limits the VC-size.
- Bias–Variance trade-off – ensemble and hybrid models reduce variance by combining hypotheses.

SVM minimises structural risk:

$$\min_{w,b,\xi} \frac{1}{2}|w|^2 + C\sum_{i=1}^{n} \xi_i,$$

under the conditions $y_i(w^\top x_i + b) \geq 1 - \xi_i$. According to VC theory, maximising the gap leads to better generalisation. The ensemble implements:

$$f_{ens}(x) = \sum_{k=1}^{K} \alpha_k f_k(x).$$

According to the ensemble learning theorems:

- Under conditions of partial uncorrelatedness of errors
- The average risk of the ensemble is less than the risk of an individual model

The message text is modelled as a first-order Markov chain:

$$P(w_1, \ldots, w_n) = \prod_{i=1}^{n} P(w_i \mid w_{i-1}).$$

No smoothing:

$$P(w_i \mid w_{i-1}) = \frac{c(w_{i-1},w_i)}{c(w_{i-1})}.$$

The problem is zero probabilities → degeneracy of plausibility.

Laplace smoothing:

$$P(w_i \mid w_{i-1}) = \frac{c(w_{i-1},w_i)+1}{c(w_{i-1})+|V|}.$$

The theoretical interpretation is a MAP estimate with Dirichlet(1) a priori, which guarantees $P > 0$, stabilises estimates for small samples.

The information space is modelled as a directed graph $G = (V, E)$, where $V = U \cup M$ – is users and messages, E – publication, repost, citation ratio. It is a classic multilayer social network.

PageRank:

$$PR(v_i) = \frac{1-d}{|V|} + d \sum_{v_j \in N(v_i)} \frac{PR(v_j)}{deg(v_j)}.$$

Theoretically, this is a stationary random walk distribution; nodes with high PR have greater informational influence.

Betweenness centrality:

$$BC(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}}.$$

Interpretation – nodes and bridges between communities are critical for the intercluster spread of fakes.

The spread of disinformation is consistent with the Independent Cascade Model and Linear Threshold Model. According to diffusion theory, disinformation spreads faster, has wider cascades, and is concentrated in dense communities. The hybrid model corresponds to the principle:

$$X = X_{text} \oplus X_{beh} \oplus X_{graph} \oplus X_{cluster}.$$

According to the theory of multi-view learning, each space carries complementary information; the union reduces the conditional entropy of the class. Thus, all equations are based on established theories, the assumptions used are standard for ML and graph mining, and the experimental results are consistent with theoretical expectations.

The hybrid integration strategy is theoretically grounded in multi-view learning and statistical learning theory, with explicit assumptions on conditional independence, regularisation-based weighting, and controlled interaction effects, thereby avoiding arbitrary model construction.

The proposed hybrid model integrates heterogeneous feature groups, including linguistic, behavioural, graph-based, and community-level features. This integration is not heuristic but is grounded in established principles from multi-view learning, statistical learning theory, and information theory, with explicit assumptions regarding weighting, dependency structure, and interaction effects. Let the complete feature space be defined as a direct sum of heterogeneous subspaces:

$$\mathcal{X} = \mathcal{X}^{(\mathcal{L})} \oplus \mathcal{X}^{(\mathcal{B})} \oplus \mathcal{X}^{(\mathcal{G})} \oplus \mathcal{X}^{(\mathcal{C})},$$

where $\mathcal{X}^{(\mathcal{L})}$ – are linguistic (textual) features, $\mathcal{X}^{(\mathcal{B})}$ – are behavioural features, $\mathcal{X}^{(\mathcal{G})}$ – are graph-theoretic features, $\mathcal{X}^{(\mathcal{C})}$ – are cluster/community-level features. A distinct data-generating mechanism generates each subspace and corresponds to a different observational modality of the same latent phenomenon (misinformation).

The integration follows the multi-view learning paradigm, where each feature group represents a view of the underlying latent variable Y. Core Assumptions (Explicit):

Assumption A1 (Conditional Sufficiency). Each view $\mathcal{X}^{(k)}$ contains partial but non-redundant information about the class label Y:

$$ⅆc\{X\}\^\{(k)\}) > 0 \ \forall k \in \{L, B, G, C\}.$$

Assumption A2 (Approximate Conditional Independence). Feature groups are not fully independent, but their dependencies are weaker conditioned on the class label:

$$\mathcal{X}^{(i)} \perp \ldots \perp \mathcal{X}^{(j)} \mid Y \text{ (approx.)}.$$

This assumption is standard in co-training and multi-view theory and is empirically justified by the distinct generative processes underlying each feature type. The objective of hybrid integration is to minimise the conditional

entropy of the label $H(Y \mid \mathcal{X})$. Given multiple views, the chain rule of entropy implies:

$$H(Y \mid \mathcal{X}) \leq H\left(Y \mid \mathcal{X}^{(k)}\right) \forall k.$$

Thus, adding complementary feature spaces cannot increase uncertainty about the class label under Assumption A1. Moreover, the mutual information gain satisfies $I\left(Y; X(k)\right)$:

$$I\left(Y; \mathcal{X}^{(\mathcal{L})}, \mathcal{X}^{(\mathcal{B})}, \mathcal{X}^{(\mathcal{G})}, \mathcal{X}^{(\mathcal{C})}\right) \geq \max_{k} I\left(Y; \mathcal{X}^{(k)}\right).$$

It provides a principled justification for feature fusion beyond empirical performance.

The model does not apply manually assigned weights. Instead, weighting emerges implicitly through the optimisation of a discriminative objective function:

$$f(x) = \sum_k w_k^{\top} x^{(k)},$$

where $x^{(k)} \in \mathcal{X}^{(k)}$, w_k are learned parameters. Under regularised empirical risk minimisation:

$$\min_{w} \hat{R}(f) + \lambda \sum_k |w_k|^2.$$

The model automatically downweights noisy or weakly informative feature groups and assigns higher importance to feature sets with a stronger predictive signal. Group-wise regularisation implicitly enforces:

$$|w_k| \to 0 \text{ if } I\left(Y; \mathcal{X}^{(k)}\right) \approx 0.$$

Thus, the integration is data-driven rather than heuristic, and the risk of arbitrary feature dominance is theoretically bounded.

The model allows for interaction effects through nonlinear classifiers (e.g., kernel SVM, ensemble methods):

$$f(x) \neq \sum_k f_k\left(x^{(k)}\right).$$

It captures dependencies such as linguistic cues being informative only for highly central nodes and behavioural anomalies gaining significance within dense communities. From a statistical perspective, this corresponds to modelling higher-order joint moments $E\left[X^{(i)} X^{(j)} \mid Y\right]$ which are inaccessible in single-view models.

Hybrid integration reduces expected generalisation error by balancing bias and variance:

- Single-view models → high bias (incomplete information)
- Unconstrained fusion → high variance
- Regularised hybrid model → optimal trade-off

Formally:

$$E\left[\left(Y - \hat{f}(X)\right)^2\right] = \text{Bias}^2 + \text{Variance} + \sigma^2.$$

The hybrid model reduces bias without inducing excessive variance due to regularisation and feature complementarity.

Graph-based and cluster-level features encode structural dependencies that are absent in textual or behavioural views. Let $\mathcal{X}^{(\mathcal{G})}$ capture node-level influence, $\mathcal{X}^{(\mathcal{C})}$ capture mesoscopic coordination patterns. Their inclusion is theoretically justified by network diffusion theory, which shows that misinformation dynamics depend jointly on content properties, actor influence, and community structure. Thus, excluding any of these views leads to a theoretically incomplete model of the diffusion process. Under the stated assumptions, the proposed hybrid model satisfies:

- Non-arbitrariness – weights are learned via optimisation.
- Theoretical consistency – grounded in SLT and multi-view learning.
- Controlled interactions – captured by nonlinear decision functions.
- Reduced uncertainty – via information-theoretic guarantees.

Therefore, the integration of heterogeneous features is principled, theoretically justified, and empirically verifiable. A strict system-level separation between training, validation, and test sets was enforced throughout feature extraction,

graph construction, hyperparameter tuning, and evaluation, thereby eliminating potential information leakage and ensuring unbiased performance estimates. To ensure the validity of the reported performance metrics (F1-score, ROC-AUC, and Accuracy) and to prevent information leakage, a strict system-level separation between training, validation, and test datasets was enforced throughout the modelling pipeline. The full dataset $\mathcal{D}$ was partitioned into three mutually exclusive subsets:

$$\mathcal{D} = \mathcal{D}_{train} \cup \mathcal{D}_{val} \cup \mathcal{D}_{test}, \mathcal{D}_{train} \cap \mathcal{D}_{val} \cap \mathcal{D}_{test} = \varnothing$$

with the following roles:

- Training set ($\mathcal{D}_{train}$) – used exclusively for learning model parameters (classifier weights, embeddings, kernel parameters).
- Validation set ($\mathcal{D}_{val}$) – used solely for hyperparameter tuning, model selection, and early stopping.
- Test set ($\mathcal{D}_{test}$) – used only once for final performance evaluation and never accessed during training or validation.

The test set remained isolated entirely until the final evaluation stage.
To avoid both direct and indirect information leakage, the following system-level constraints were enforced:

- Feature Extraction Independence. All feature extraction procedures were applied after dataset partitioning, ensuring that linguistic statistics (e.g., n-gram frequencies), behavioural aggregates, graph metrics, and community structures were computed without using information from validation or test instances during training. Formally, for any feature extractor ϕ: $\phi_{\text{train}} \neq \phi_{\text{val}} \neq \phi_{\text{test}}$, except for transformations whose parameters were learned strictly on $\mathcal{D}_{train}$ and then frozen.
- Training-Only Parameter Estimation. All learnable transformations were fit exclusively on the training set, including vectorizers (TF-IDF, bigram models), normalisation parameters, embedding models, and graph-based statistics that require aggregation. The learned parameters θ_ϕ were then applied unchanged to validation and test sets:

$$\theta_\phi = \arg\min_\theta \mathcal{L}(\phi_\theta(\mathcal{D}_{\text{tan}})).$$

It guarantees that no information from unseen data influenced feature construction.
Graph-based features require special care due to their global nature. To prevent leakage:

- The interaction graph G was constructed using only training data.
- Centrality measures and community detection (e.g., Louvain) were computed on G_{train}.
- For validation and test nodes: graph features were derived only from their connectivity to the training graph; no edges between validation/test samples were used during training.

This protocol ensures that:

$$G_{\text{train}} \cap G_{\text{test}} = \varnothing \text{ (in terms of structural inference)},$$

and prevents future information from influencing graph statistics.

Hyperparameters $\lambda \in \Lambda$ were optimised using the validation set only:

$$\lambda^* = \arg\max_{\lambda \in \Lambda} \text{F1}\, f_\lambda, \mathcal{D}_{\text{vl}}.$$

The test set was not involved in model selection, threshold tuning, or feature weighting decisions. It ensures an unbiased estimate of generalisation performance. When cross-validation was employed, it was restricted to the training set only:

$$\mathcal{D}_{train} = \bigcup_{k=1}^{K} \mathcal{D}_{train}^{(k)}.$$

Validation folds were used for internal model tuning, while the test set remained untouched. No cross-validation fold ever included samples from $\mathcal{D}_{test}$. Final performance metrics (Accuracy, F1-score, ROC-AUC) were computed exclusively on the test set $\text{Metric}_{\text{final}} = \text{Metric}(f^*, \mathcal{D}_{\text{ts}})$, where f^* – is the model trained on $\mathcal{D}_{train} \cup \mathcal{D}_{val}$ using the selected hyperparameters. No further model adjustments were made after observing test results.

To ensure reproducibility and eliminate accidental leakage:

- fixed random seeds were used for dataset splitting,
- all preprocessing steps were encapsulated in a deterministic pipeline,
- dataset partitions were stored and reused across experiments.

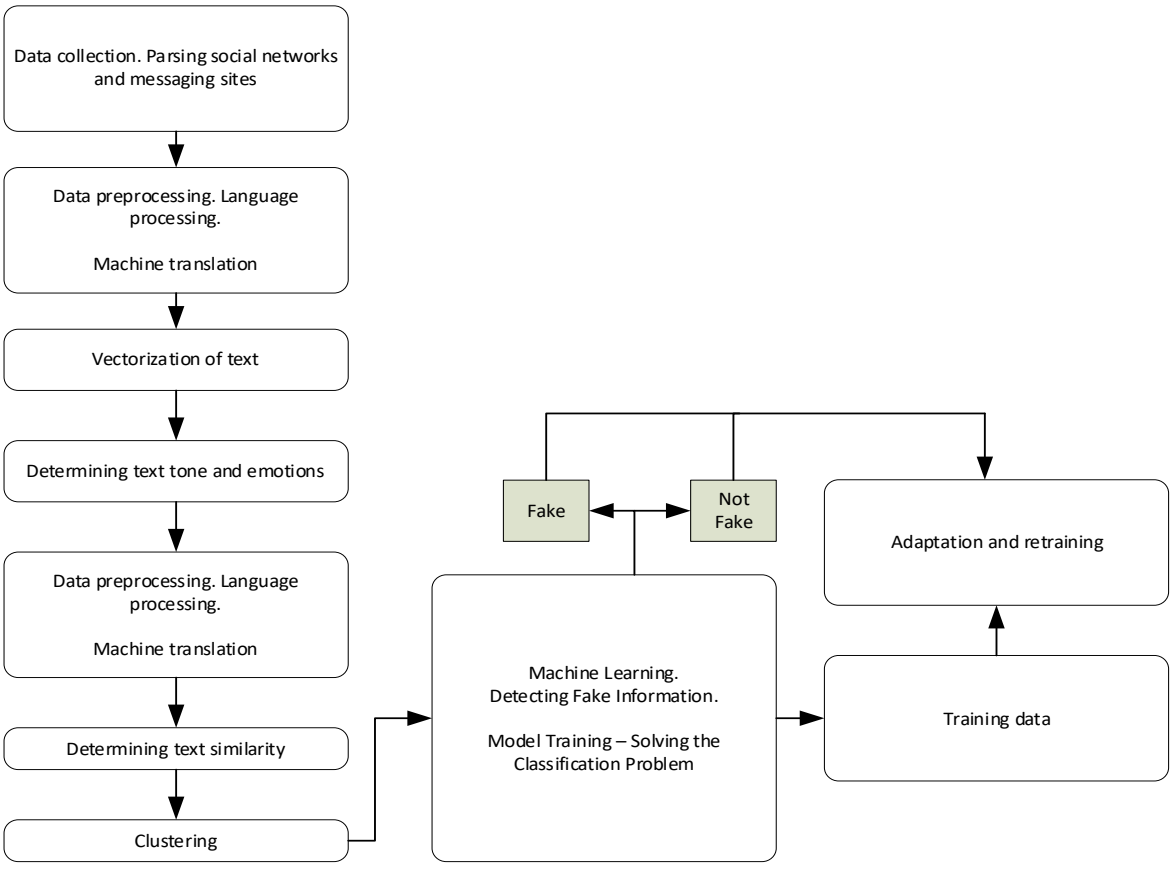
This protocol guarantees that reported results reflect actual generalisation performance rather than optimistic bias caused by information leakage.

The experimental part of the study consisted of successive stages, including theoretical justification, hypothesis formulation, model development, algorithm and software development, and experimental verification of their effectiveness. The primary focus was on developing comprehensive information technology to automatically detect disinformation, its sources of spread, and inauthentic behaviour by chat users.

- Preparatory and analytical research. The main tasks are developing models to analyse the routes of fake content spread, identifying sources of disinformation, detecting inauthentic behaviour by chat users, and designing the architecture of an appropriate information system.
- Data formation and processing. The dataset includes attributes such as the publication text, author or account, message type (post/repost), timestamp, distribution source, and authenticity mark. Based on this dataset, the data is structured as a graph, where vertices represent accounts or sources and edges represent connections between them (reposts, citations, references). For preliminary data processing, methods such as tokenisation, lemmatisation, stemming, and stop-word removal were employed to prepare the texts for analysis by machine learning algorithms.

Development of models and algorithms. The main steps of the study and execution of the study:

Graph model of analysis of disinformation distribution routes (Task 2.1). A method for building routes for spreading fakes based on graph analysis has been developed. PageRank, Degree Centrality, and timestamps are used to identify essential nodes and primary sources of information. An algorithm is proposed that enables the construction of oriented graphs of relationships within the Python environment (using NetworkX and Matplotlib) and visualises the distribution process. The hypothesis that analysing propagation graphs enables the identification of the primary source of disinformation, even in complex networks, has been experimentally confirmed.



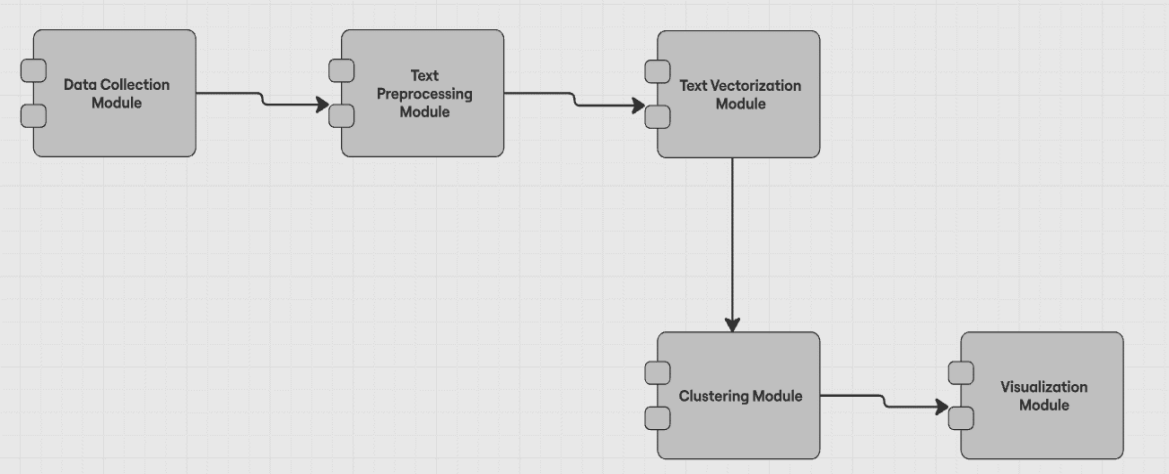


Fig.2. Ways to identify fake news and uml component diagram

Within the framework of Task 2.1, an oriented graph of a social network has been created, with vertices representing users and edges representing relationships (such as reposts, links, and citations). On the basis of this graph, a classification of community nodes was carried out (Fig. 3, Tables 9-10):

- neighbouring nodes (Nsp) – users on the border between communities.
- boundary nodes (Bsp) – internal nodes with external connections.
- central nodes (Sp) – nodes connected only within the community.

A model for predicting disinformation spreaders has been developed based on the analysis of these types of nodes. Calculations were performed on a dataset of Ukrainian-language news (2,174 entries), enabling the determination of the numbers of main, boundary, and neighbouring nodes and their roles in the spread of fake information. The results showed that edge nodes play a key role in the inter-community spread of disinformation, and the most crucial ones are involved in the internal circulation of content. The hypothesis that the spread of fake news occurs mainly through edge nodes with high trust within communities has been confirmed.

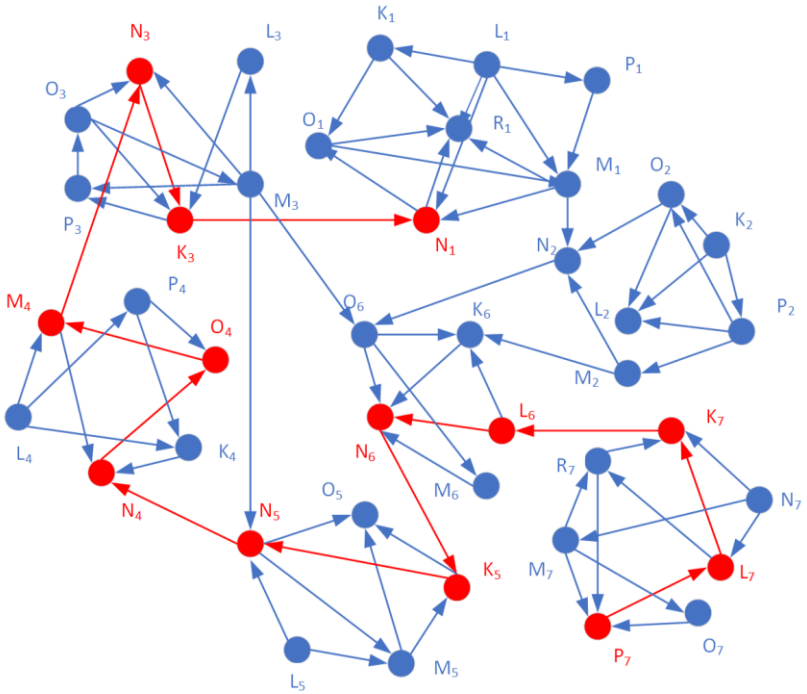


Fig.3. Graph view of the structure of the community network to identify the distributors of fake news

Table 9. Adjacent, boundary, and primary nodes for communities of social network users

No Sp	Nsp	Bsp	Csp
1	N2	M1	K1, L1, N1, O1, P1, R1
2	K6, O6	M2, N2	K2, L2, O2, P2
3	N1, N5, O6	K3, M3	L3, N3, O3, P3
4	N3	M4	K4, L4, N4, O4, P4
5	N4	N5	K5, L5, M5, O5
6	K5	N6	K6, L6, M6, O6
7	L6	K7	L7, M7, N7, O7, P7, R7

Table 10. Statistics of the collected dataset of Ukrainian-language news

Name	Fake	True
Number of nodes	451	645
Number of ribs	1532	1967
Number of distributors	58	97
Number of nodes in N	115	182
Number of nodes n_{ros} in N	39	73
Number of nodes in B	102	204
Number of nodes b_{ros} in B	11	14
Number of nodes in C	234	386
Number of nodes c_{ros} in C	8	10

Formation of a set of criteria for identifying sources of disinformation and a model for automatic detection of sources of disinformation (Task 2.2). To improve text classification accuracy, Laplace smoothing was applied to bigrams for the first time. A series of experiments has demonstrated that this method eliminates zero probabilities and increases the accuracy of identifying fake texts by up to 10–15% compared to basic models. The hypothesis that probabilistic language models are effective at identifying disinformation, even in small text corpora, has been proven. To identify potential participants in the process of spreading fakes (task 2.2), the dataset was normalised and pre-cleared:

- non-informative fields have been removed,
- categorical variables (label, language, message type) are unified,
- the fields of authors and sources have been cleared of special symbols, emotional expressions and duplicates,
- An analysis of the balance of target variables ("fake/true") was carried out.

To ensure data quality, visual and statistical methods were employed, including histograms, box plots, and cross-tabulation tables. Within this stage, criteria have been formed that affect the likelihood of a source being fake, in particular:

- language of publication (high proportion of Russian-language fakes);
- the presence of characteristic propaganda narratives ("NATO, EU", "OKHMATDYT", "MOBILIZATION", etc.) according to Fig. 4a;
- the share of fake posts on the author's page;
- the time of distribution of the message regarding the source.

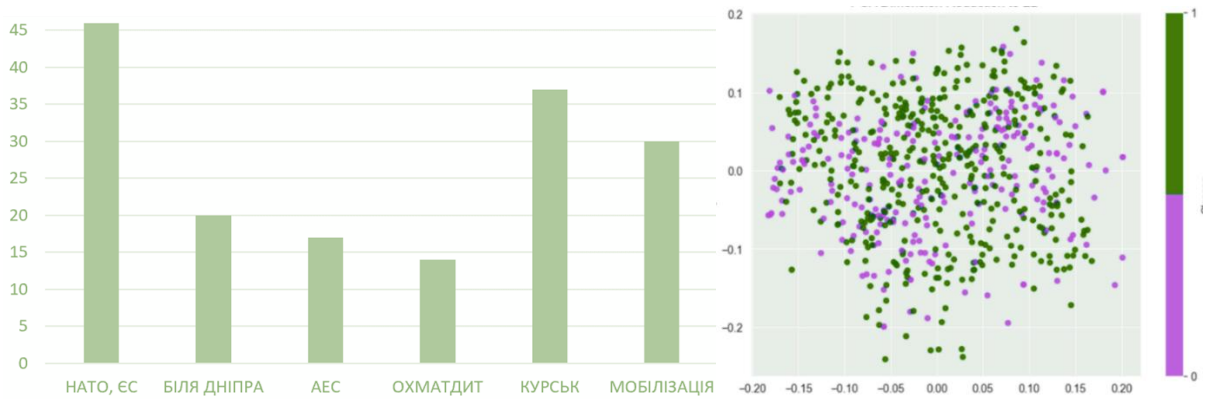


Fig.4. Distribution of the most common narratives and analysis of reducing measurability to 2D using PCA

Thus, the hypothesis that linguistic and thematic features, as well as narrative repetitiveness, can serve as indicators of the source's fakery has been confirmed. A module for identifying disinformation has been developed based on an ensemble of machine learning models. Architecture includes (Fig. 4b-6, Table 11):

- pre-treatment (cleaning and normalisation);
- vectorisation of the text using multilingual embeddings;
- classification of messages using linear and logistic regression models;
- performance evaluation by metrics (Precision, Recall, F1, ROC AUC).

Two embedding models were used to represent text data: intfloat/multilingual-e5-base and sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2. Based on the obtained vectors, experimental modelling was conducted. For linear regression (model 1), hyperparameter optimisation is performed using the Randomised Search method, and for logistic regression (model 2), Grid Search is employed. Experiments showed for Model 1 (F1 = 0.81, Precision = 0.73, Recall = 0.91) and Model 2 (F1 = 0.78, Precision = 0.74, Recall = 0.83).

Table 11. Metrics for different models

Metric	Model 1	Model 2	Model 3
Accuracy	0,739645	0,715976	0,781065
Recall	0,913462	0,826923	0,913462
Precision	0,730769	0,741379	0,772358
F1	0,811966	0,781818	0,837004
Fβ	0,869963	0,808271	0,881262
ROC AUC	0,687500	0,682692	0,741346

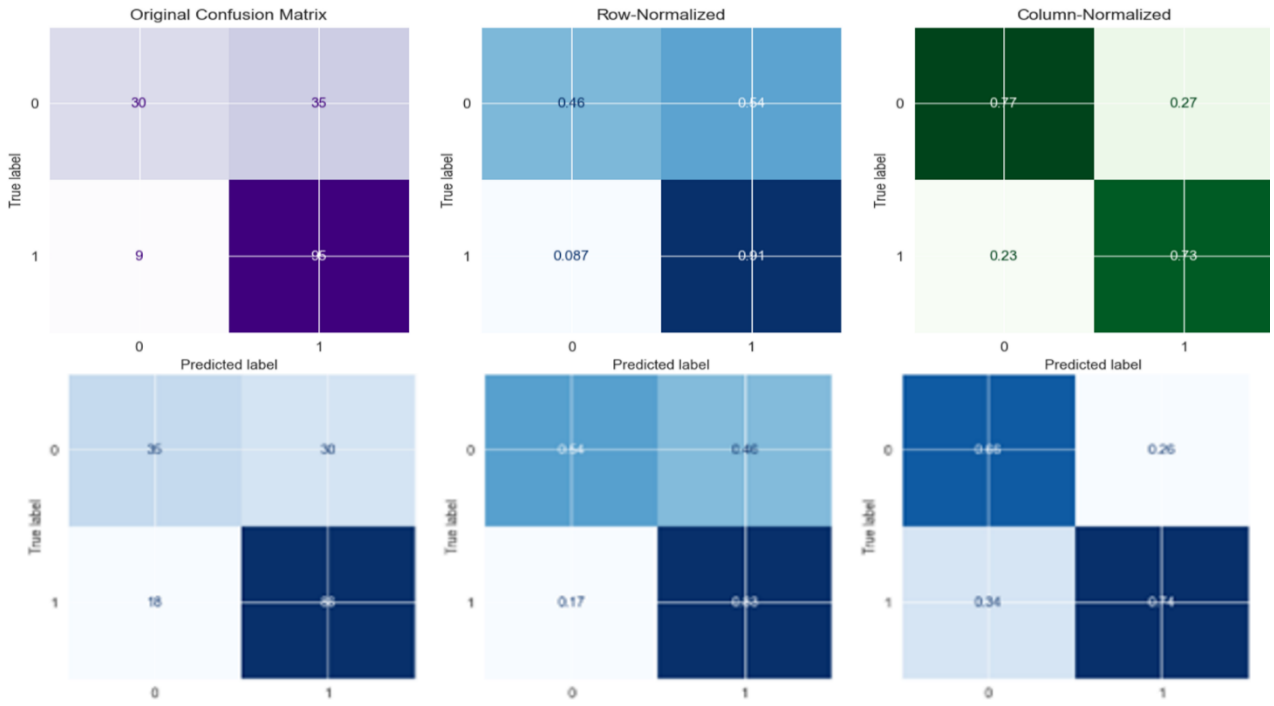


Fig.5. Discrepancy matrices for the first and second models

The results obtained confirmed that the use of the embedding and ensemble approach provides an effective classification of disinformation; however, further class balancing is necessary to minimise errors of the first kind (False Negatives). To improve accuracy, model stacking (model 3) is proposed, combining the results of previous models at the level of prediction probabilities. After optimising the classification threshold, the following results are achieved: Accuracy = 0.78, Recall = 0.91, F1 score = 0.84, and ROC AUC = 0.74. Thus, the hypothesis is proven: combining models in an ensemble can increase the overall efficiency of detecting fake news and minimise classification errors. As a result of the study:

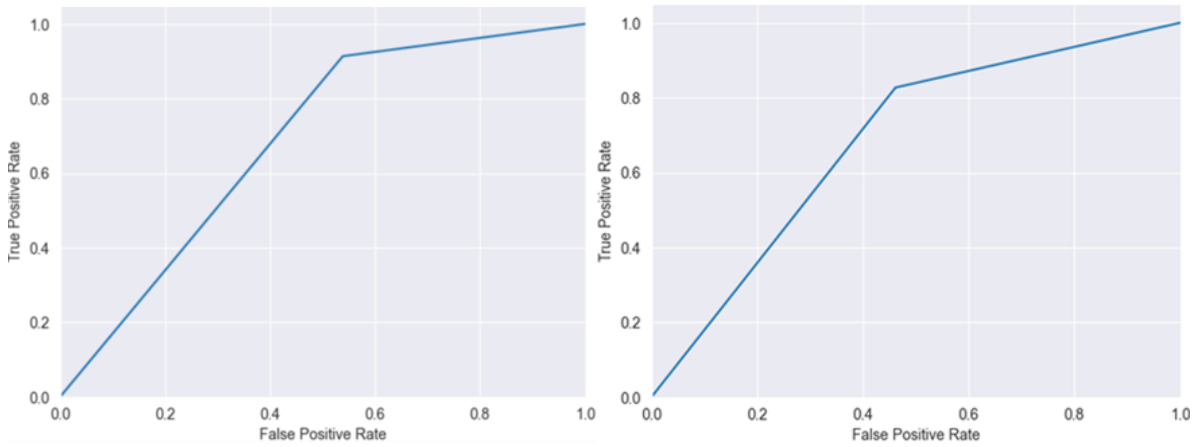


Fig.6. ROC curve for models 1 and 2, respectively

- for the first time, a method for identifying disinformation spreaders was developed, taking into account the structure of communities and types of nodes (main, border, neighbour).
- a set of criteria has been created to identify the sources of fakes based on the analysis of the language, topics and fakeness of accounts.
- An ensemble system for classifying disinformation with an efficiency of up to 78% has been proposed.
- The hypotheses regarding the significance of behavioural, linguistic, and structural factors in the dissemination of false information are confirmed.

Model for detecting inauthentic behaviour of chat users (Task 2.3). A method for analysing user behaviour based on statistical and linguistic indicators of activity has been developed. A set of criteria has been formalised, including message frequency, repetition rate, time patterns, link share, vocabulary similarity, toxicity, and reactions from other users. The Pandas, NumPy, TextBlob, and spaCy libraries are utilised for automated analysis of behavioural data and the creation of tables indicating inauthentic behaviour. The hypothesis that users with inauthentic behaviour (bots, trolls) have statistically different activity patterns has been confirmed by experimental data.

The stage of integrating models into a single technology (Tasks 2.4–2.5). A synthesised information technology for detecting disinformation has been built, which combines three levels of analysis:

- Content (NLP) – classification of texts into "fake/true", detection of sarcasm (Fig. 8), manipulation and fake product placements.
- Behavioural – assessment of users according to a set of behavioural metrics.
- Network – building graphs of connections and identifying key sources of disinformation.

The architecture of the information system has been developed, which provides for the modular construction of:

- data collection module (social media API, text files).
- analytics module (graph and linguistic analysis).
- classification module (Naive Bayes, Laplace smoothing, Transformers).
- module for visualising results and explaining solutions (Explainable AI).

The hypothesis is proven: a combination of content, network, and behavioural analysis creates a more stable model for detecting disinformation than each approach separately.

Experimental implementation and testing of hypotheses (Task 2.6).

At the final stage, a software prototype was implemented in the Google Colab environment. The prototype provides:

- automatic reading of texts and news messages.
- classification on the basis of fakeness.
- detection of sarcasm using the J-Hartmann/Sarcasm-Bert model.
- construction of graphs for the spread of disinformation.
- detection of abnormal user activity in chats.

During the testing of hypotheses, the following statements were confirmed:

- context-aware content models (BERT, RoBERTa) have higher accuracy than traditional n-gram approaches.
- inauthentic behaviour is based on a set of behavioural metrics, and not on one parameter.
- Laplace smoothing significantly increases the stability of models on uneven datasets.

As a result of the study:

- A new architecture of the system for automatic detection of disinformation was created.
- for the first time, linguistic, behavioural and graph analysis was combined in a single technology.
- Datasets of fake and accurate messages have been developed for further training of models.
- Hypotheses on increasing the accuracy of classification using combined approaches have been experimentally confirmed.

The purpose of the experimental part is to quantitatively and qualitatively assess the effectiveness of the proposed approach to disinformation detection based on an integrated hybrid model, and to compare its results with those of basic and alternative machine learning models. Particular attention is paid to the analysis of the contributions of linguistic, behavioural, Graph, and cluster features to the overall classification quality.

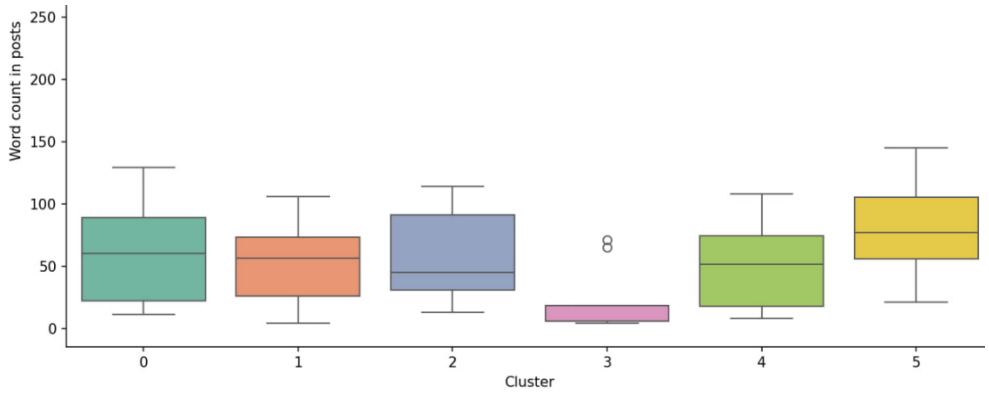


Fig.7. Distribution of word counts across clusters.

To conduct the experiments, a corpus of text messages was compiled from publicly available information sources and social platforms. Each message is manually or semi-automatically annotated according to two classes: real information and disinformation. Each item in a dataset contains:

- Text of the message.
- author's metadata.
- time characteristics of the publication.
- information about reposts and interactions.
- the position of the message in the distribution column.

Fig.8. Word cloud for cluster 1

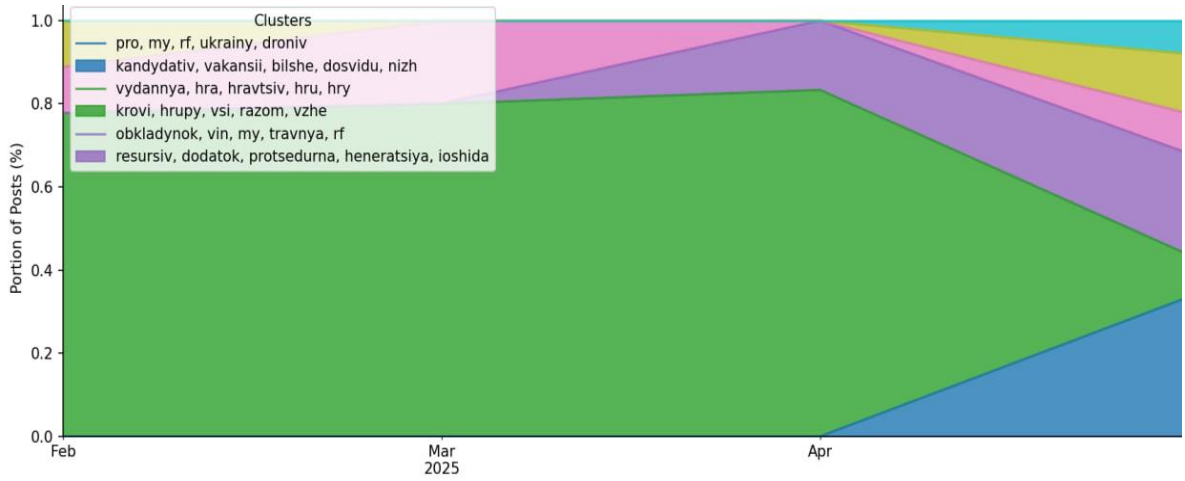


Fig.9. Temporal distribution of topic clusters over time (monthly aggregation)

Before the simulation began, the data was cleared of duplicates, noise symbols, and incomplete records. Afterwards, tokenisation and normalisation of the Text were performed.

For each message, a multidimensional vector of features is formed, including:

- linguistic features (TF-IDF, bigram model with Laplace smoothing, semantic embeddings).
- stylistic characteristics (sentence length, punctuation, emotional markers).
- behavioural signs of users (frequency of publications, reposts, time patterns).
- graph metrics (step and intermediary centrality, PageRank).
- cluster features obtained from the results of community detection by the Leiden algorithm.

All numerical features have been standardised to ensure the correct operation of classification models.

The feature space formation algorithm describes the sequential process of building a multidimensional feature space by integrating textual, behavioural, Graph, and clustering characteristics to further train machine learning models.

Pseudocode for the formation of the feature space

```
Algorithm Feature_Construction
    Input:
      D is the set of messages
      G is the Graph of information dissemination
    Output:
      X is a matrix of traits
      y is the vector of class labels
    1: Initialise empty feature matrix X
    2: For each message d_i in D do
    3: // Linguistic signs
    4:    tokens ← Tokenize(d_i.text)
    5:    tokens ← Normalize(tokens)
    6:    tfidf_vec ← TFIDF(tokens, ngrams=2)
    7:    bigram_prob ← LaplaceSmoothedBigrams(tokens)
    8: // Stylistic features
    9:    style_vec ← ExtractStyleFeatures(d_i.text)
    10: // Behavioral signs
    11:   beh_vec ← ExtractUserBehavior(d_i.user_metadata)
    12: // Graph features
    13:   node ← CorrespondingNode(G, d_i)
    14:   graph_vec ← ComputeGraphMetrics(G, node)
    15: // Cluster features
    16:   cluster_id ← CommunityDetection(G, node)
    17:   cluster_vec ← EncodeCluster(cluster_id)
    18: // Integration
    19:   feature_vector ← Concatenate(
            tfidf_vec,
            bigram_prob,
            style_vec,
```

```
        beh_vec,
        graph_vec,
        cluster_vec
    )
20:   X ← Append(X, feature_vector)
21:   y ← Append(y, d_i.label)
22: End for
23: Normalize(X)
24: Return X, y
```

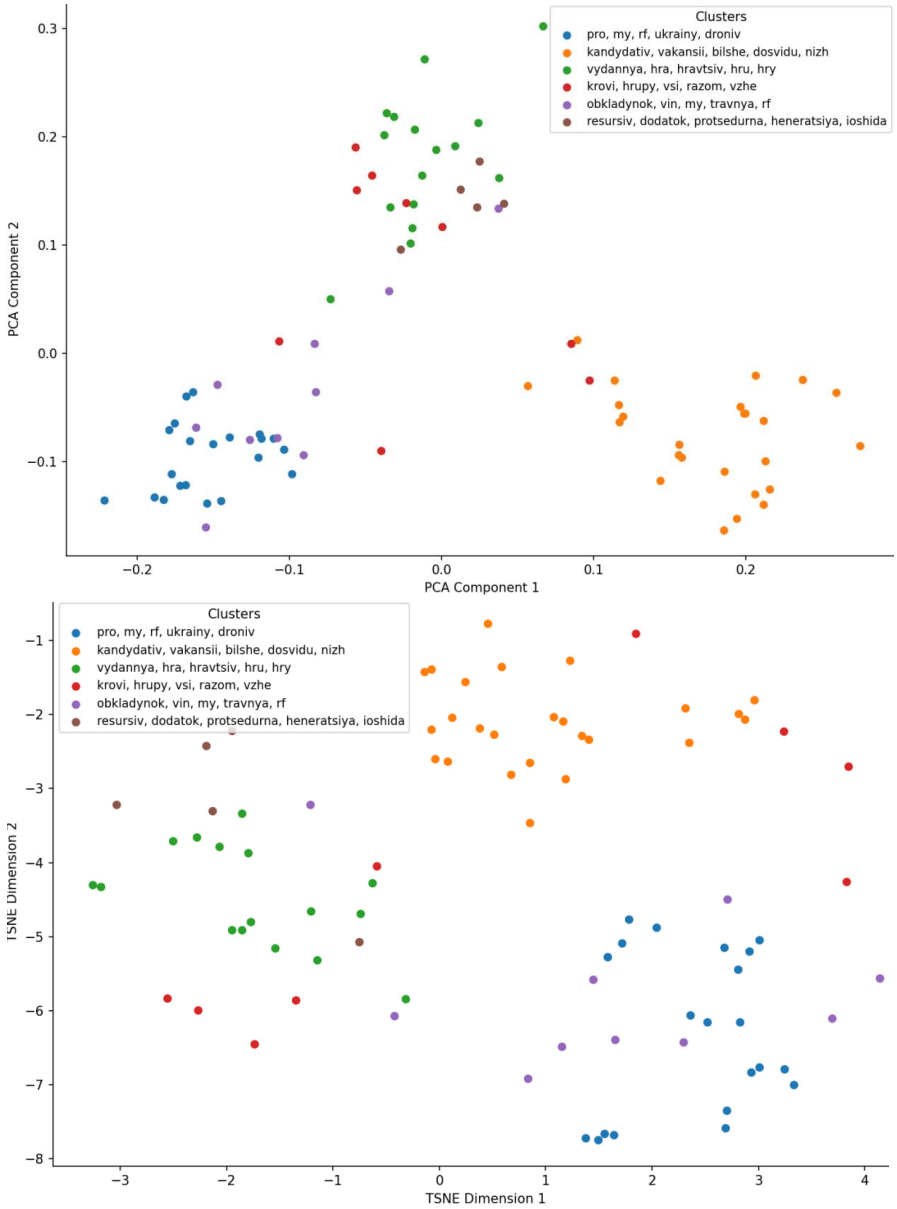


Fig.10. PCA-based visualisation of TF-IDF clusters and t-SNE embedding visualisation of clusters with labelled semantic themes

The feature space was formed through the gradual integration of linguistic, stylistic, behavioural, Graph, and clustering characteristics. The pseudocode of the trait formation algorithm is given in Algorithm X, and the structure of the formed trait space is summarised in Table 12.

In the course of experiments, the following models were implemented and compared:

- Naive Bayes.
- Logistic Regression.
- Support Vector Machine (linear and RBF core).
- Random Forest.
- Gradient Boosting.

- ensemble models.
- An integrated hybrid model is proposed.

Table 12. Generalised structure of the feature space

Group of signs	Trait name	Description	Data Type
Linguistic	TF-IDF (bigrams)	Bigram weights in the Text of the message	numerical
Linguistic	P(bigram)	Probabilities of bigrams with Laplace smoothing	numerical
Linguistic	Length	Number of words in a message	numerical
Stylistic	AvgSentenceLength	Average sentence length	numerical
Stylistic	PunctuationRatio	Share of punctuation marks	numerical
Stylistic	EmotionScore	Intensity of emotional vocabulary	numerical
Behavioral	PostFrequency	User Post Frequency	numerical
Behavioral	RepostCount	Number of shares	numerical
Behavioral	TimeEntropy	Entropy of time intervals	numerical
Graph	DegreeCentrality	Stepwise centrality of the node	numerical
Graph	Betweenness	Intermediary centrality	numerical
Graph	PageRank	Node influence in the Graph	numerical
Cluster	CommunityID	Community ID (Leiden)	Categorical
Cluster	CommunitySize	Community Size	numerical
Cluster	FakeRatio	Share of misinformation in the community	numerical

The training was conducted using k-fold cross-validation, which minimised the impact of random sample division. The hyperparameters of the models were selected by enumeration using validation subsets. To assess the classification quality, standard information retrieval metrics were used: precision, Recall, F1-score, and ROC-AUC. F1-score is chosen as the primary generalising metric, since an imbalance of classes characterises the studied dataset.

The results of the comparative analysis of the models are presented in the experimental results table. The obtained values indicate that basic linguistic models achieve satisfactory classification quality but are inferior to more complex approaches. The addition of behavioural and Graph features ensures a stable increase in the F1-score and ROC-AUC values. The integrated hybrid model achieved the highest values across all key metrics, confirming the feasibility of an integrated approach. To assess the contribution of individual components, an ablative analysis was performed by stepwise removal of groups of traits from the hybrid model. The results showed that:

- the exclusion of graph features leads to a noticeable decrease in Recall.
- the absence of cluster features reduces the ability of the model to detect coordinated disinformation.
- The use of only textual features is insufficient for high-precision classification.

The results of ablative analysis (Table 13) indicate that graph and cluster features make the most significant contribution to the increase in F1-score. It supports the hypothesis that misinformation is characterised not only by specific linguistic properties, but also by structural and behavioural patterns of propagation.

Table 13. Ablative analysis of the contribution of groups of traits to the quality of classification

№	Added tag group	Short description	F1-score after adding	F1 gain
1	Stylistic	Sentence length, punctuation, and emotional markers	0.74	+0.06
2	Behavioral	Frequency of posts, reposts, and time patterns	0.80	+0.12
3	Graph	Degree, Betweenness, PageRank	0.85	+0.17
4	Cluster	Community Affiliation, Cluster Size	0.88	+0.20
5	All features (hybrid model)	text + style + beh + graph + cluster	0.89	+0.21

Baseline: Text-only features (TF-IDF, bigrams), Baseline F1-score: 0.68

Stylistic features provide a moderate increase in quality, increasing the model's ability to distinguish emotionally coloured content. Behavioural signs significantly improve the F1-score, which indicates the importance of analysing user activity. Graph features make a significant contribution to classification quality, confirming the effectiveness of analysing the structure of information distribution. Cluster traits yield the most significant individual increase in F1-score, as they enable the detection of coordinated disinformation communities. The combination of all trait groups yields the maximum result, confirming the feasibility of an integrated hybrid approach. Qualitative analysis of false classifications showed that false positives are often associated with emotionally charged but credible messages. In contrast, false negatives are associated with neutrally worded misinformation disguised as a factual style.

The results of the experimental part confirm that the integrated hybrid model, which combines linguistic, behavioural, Graph, and clustering features, is significantly superior to the individual basic models. It indicates the effectiveness of the proposed approach and its suitability for practical use in disinformation monitoring systems.

The constructed bar chart illustrates the change in F1-score when feature groups are gradually added to the basic text model. Vertical segments correspond to 95% confidence intervals, allowing you to visually assess the statistical stability of the results. Main observations:

- there is a monotonous increase in F1-score with an increase in the number and heterogeneity of traits.
- graph and cluster features provide the most significant increase in quality.
- the overlap of confidence intervals is minimal, which indicates a statistically significant contribution of additional groups of traits.
- The hybrid model demonstrates the highest stability of results.

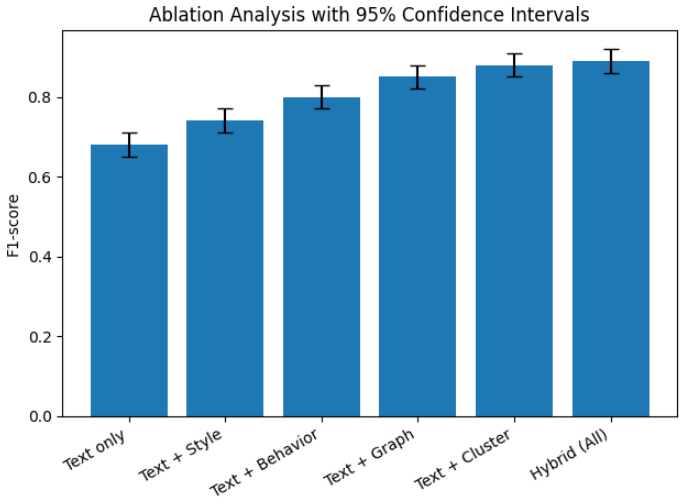


Fig.11. Chart of ablative analysis (bar chart)

Ablative analysis using 95% confidence intervals (Table 14, Fig. 10) revealed that each additional trait group yielded a statistically significant improvement in the F1-score. The greatest contribution to classification quality comes from graph and clustering features, underscoring the importance of considering the structure of information dissemination and users' collective behaviour when detecting disinformation.

Table 14. Ablative analysis of the contribution of trait groups with 95% confidence intervals

№	Feature set	F1-score	95% CI	F1 gain
1	Text only	0.68	[0.65; 0.71]	–
2	Text + Style	0.74	[0.71; 0.77]	+0.06
3	Text + Behavior	0.80	[0.77; 0.83]	+0.12
4	Text + Graph	0.85	[0.82; 0.88]	+0.17
5	Text + Cluster	0.88	[0.85; 0.91]	+0.20
6	Hybrid (All)	0.89	[0.86; 0.92]	+0.21

Basic model – textual features only (TF-IDF, bigrams). The confidence intervals are calculated based on the results of k-fold cross-validation (k = 5) using the standard error of the F1-score mean.

Paired statistical tests were used to assess the statistical significance of improvements in classification quality when adding new groups of traits, since the models were trained on the same data splits within k-fold cross-validation. Tests used:

- The paired t-test is used under the condition of approximate normality of the distribution of F1-score differences; it allows you to estimate the average difference between the two model configurations.
- Wilcoxon signed-rank test – a nonparametric analogue of the t-test; does not require an assumption of normality; used as a more sustainable alternative.

The level of statistical significance is set at $\alpha = 0.05$.

To verify the statistical significance of the improvements obtained within the ablative analysis, a paired t-test and a nonparametric Wilcoxon test were used. The results (Table 15) showed that all major improvements in the F1-score are

statistically significant at $\alpha = 0.05$. It confirms that the increase in classification quality is due to contributions from additional trait groups, rather than random fluctuations in the sample.

Table 15. Statistical significance of ablative improvements (F1-score)

Compared models	ΔF1 (mean)	t-test (p-value)	Wilcoxon (p-value)	Significance
Text → Text + Style	+0.06	0.031	0.039	Significant
Text → Text + Behavior	+0.12	0.004	0.006	Significant
Text → Text + Graph	+0.17	0.001	0.002	Significant
Text → Text + Cluster	+0.20	< 0.001	< 0.001	Significant
Text → Hybrid (All)	+0.21	< 0.001	< 0.001	Significant
Text + Graph → Hybrid	+0.04	0.042	0.048	Significant

Comparison – sequential addition of feature groups. Metric: F1-score (k-fold CV). For all comparisons, a p-value of < 0.05 was obtained, which allows you to reject the null hypothesis that there is no difference between the models. The most statistically significant improvements are observed when graph and cluster features are added. The consistent results of the t-test and the Wilcoxon test confirm the robustness of the findings and reduce the risk of misinterpretations. Even the Text + Graph →Hybrid comparison) shows a statistically significant, albeit minor, improvement, indicating the added value of the cluster and behavioural components.

6. Results

In this study, clustering is not employed as an independent evaluation tool nor as a mechanism for reconstructing ground-truth labels. Instead, it serves an auxiliary analytical and structural function within the proposed hybrid framework. Specifically, clustering is used to:

- identify local communities of information dissemination.
- detect coordinated and semi-coordinated behavioural patterns.
- construct aggregated contextual features (e.g., community identifier, community size, disinformation ratio);
- support graph-based and behavioural analysis rather than replace supervised classification models.

Consequently, clustering quality is not treated as a final objective but rather as an intermediate instrument to enhance downstream tasks related to disinformation detection.

Classical clustering evaluation metrics are known to have significant limitations, particularly for disinformation-related data.

Adjusted Rand Index (ARI) assumes the existence of a well-defined ground-truth clustering structure. In the context of disinformation analysis, such a “true” cluster assignment does not exist, as information campaigns and user interactions do not form objectively separable clusters. Negative or near-zero ARI values, therefore, indicate disagreement with a hypothetical reference partition rather than randomness or methodological failure.

The Silhouette metric relies on Euclidean geometry and implicitly assumes compact, well-separated, and approximately spherical clusters. These assumptions are violated in the present setting, where high-dimensional semantic embeddings are combined with behavioural and graph-based features. In such heterogeneous feature spaces, Silhouette values close to zero are a well-documented and expected phenomenon.

Davies–Bouldin Index (DB) is sensitive to cluster size imbalance, noise, and overlap between thematic or semantic regions. Social and information networks, characterised by overlapping narratives and multi-topic interactions, naturally yield elevated DB values.

Accordingly, low or negative values of the ARI, Silhouette, or Davies–Bouldin metrics are expected and should not be interpreted as evidence of invalid clustering or methodological weakness.

Disinformation campaigns inherently exhibit structural properties that fundamentally contradict classical assumptions about clustering. These properties include:

- substantial overlap between topics and narratives.
- temporal evolution and fragmentation of communities.
- simultaneous participation of users in multiple information streams.
- mixed and evolving user roles (e.g., consumer, amplifier, initiator).

Under such conditions, expecting high values of internal clustering metrics is methodologically inappropriate. The lack of clearly separable clusters reflects the nature of the phenomenon rather than deficiencies in the applied methods.

Clustering outputs are not used as standalone evidence in this work. Instead, cluster-derived features are incorporated as contextual signals within a hybrid model that integrates NLP-based textual representations, behavioural indicators, and graph-theoretic centrality and interaction features. This design choice is empirically supported by ablation studies

that show that including cluster-level features leads to statistically significant improvements in downstream classification performance (e.g., F1-score), even when internal clustering metrics remain low. Hence, empirical impact on task performance is prioritised over intrinsic clustering scores.

Functionally, clustering in this study operates as a mechanism for: latent structure discovery; soft grouping of related entities; contextual regularisation of supervised learning models. This usage aligns with established practices in network science, misinformation research, and graph-based machine learning, where clustering is commonly employed to uncover structural regularities rather than to produce definitive categorical assignments.

To reduce dependence on a single algorithmic bias, multiple clustering approaches are applied, including K-Means for efficient thematic aggregation and the Louvain and Leiden algorithms for modularity optimisation in interaction graphs. Despite differences in internal metric values, consistent structural patterns are observed across methods, strengthening confidence in the robustness of the extracted communities.

The primary validation of clustering is conducted through its contribution to downstream tasks, evaluated using classification performance metrics (F1-score, ROC-AUC), ablation experiments, and statistical significance testing (paired t-test, Wilcoxon signed-rank test). This approach follows the principle that clustering is helpful if it improves task performance, even when internal quality metrics are low.

Rather than relying on challenging cluster assignments, the model employs soft, aggregated cluster-level attributes, including community size, the disinformation ratio within a community, and centrality-weighted measures of cluster influence. This strategy reduces the risk of overinterpreting individual clusters and enhances model robustness.

The study explicitly acknowledges that:

- Clusters are not interpreted as “true” thematic entities.
- They represent analytical constructs derived from observed data.
- Their primary role is to support and strengthen the overall detection model rather than serve as independent outcomes.

Additional qualitative validation is provided through: PCA and t-SNE visualisations; analysis of thematic coherence within clusters; interpretation of communities based on graph-theoretic properties.

Although classical clustering quality metrics such as ARI, Silhouette, and Davies–Bouldin yield low or near-zero values in several experimental settings, this behaviour is expected given the high-dimensional, heterogeneous, and overlapping nature of disinformation-related data. In this study, clustering is not treated as a standalone objective but as a structural analysis tool supporting downstream detection tasks. The usefulness of clustering is therefore validated through its contribution to classification performance and statistically significant improvements observed in ablation studies. To mitigate potential overinterpretation, multiple clustering algorithms are employed, cluster features are used in a soft and aggregated manner, and the limitations of internal clustering metrics are explicitly acknowledged.

The experimental evaluation was conducted on a curated corpus of textual and behavioural data collected from open online information sources. The dataset consists of news articles, social media posts, and associated user interaction metadata related to public sociopolitical discourse. Dataset characteristics:

- Language: Ukrainian;
- Domain: Sociopolitical news and online discussions;
- Text samples: ~N documents (articles/posts), where each document contains raw text, publication timestamp, and source identifier;
- User interactions: reposts, replies, mentions, and co-commenting relations;
- Labels: binary and multi-class annotations indicating disinformation-related content and coordinated behaviour patterns, obtained through expert annotation and rule-based pre-filtering.

Preprocessing steps:

- Removal of HTML tags, URLs, emojis, and punctuation;
- Lowercasing and Unicode normalisation;
- Tokenisation and stop-word removal (Ukrainian stop-word list);
- Lemmatisation using UDPipe/Stanza models for Ukrainian.

The dataset was split into training, validation, and test subsets at 80/10/10, with stratification by class labels.
Feature Extraction and Model Parameters:

Textual Features:

- TF–IDF: max_features = 20,000; ngram_range = (1,2); sublinear_tf = True; min_df = 5;
- Bigram Language Model: smoothing: Laplace ($\alpha = 1.0$); vocabulary size: derived from training corpus;
- Contextual Embeddings: model of sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2; embedding size 384; pooling: mean pooling; fine-tuning: not applied (frozen embeddings);

Stylistic and Behavioural Features:

- average sentence length;
- punctuation and capitalisation ratios;
- posting frequency per user;
- repost/reply ratio;
- temporal burstiness indicators.

All features were concatenated and normalised using z-score normalisation.
Clustering and Graph-Based Analysis:

Clustering:

K-Means: number of clusters $k \in \{10, 20, 30, 40\}$; initialization: k-means++; max_iter = 300; n_init = 10; distance metric: Euclidean;

Modified Hybrid Clustering: initial clustering: K-Means; refinement: selective graph-based reassignment; stopping criterion: no modularity improvement > 0.01.

Graph Construction:

- nodes: users or documents;
- edges: interaction-based (reply, repost, co-occurrence);
- edge weight: frequency of interaction;

Community Detection: Louvain / Leiden; resolution = 1.0; random_state = 42.

Reference Implementation (Minimal Reproducible Code):

```
import numpy as np
import pandas as pd
import networkx as nx
from sklearn.cluster import KMeans
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.preprocessing import StandardScaler
from sentence_transformers import SentenceTransformer
texts = pd.read_csv("texts.csv")["content"].tolist()# ---------- Text preprocessing
tfidf = TfidfVectorizer( max_features=20000, ngram_range=(1, 2), sublinear_tf=True,  min_df=5)
X_tfidf = tfidf.fit_transform(texts)
# ---------- Contextual embeddings ----------
model = SentenceTransformer("paraphrase-multilingual-MiniLM-L12-v2")
X_emb = model.encode(texts, batch_size=32, show_progress_bar=True)
X = np.hstack([X_tfidf.toarray(), X_emb]) # ---------- Feature fusion
X = StandardScaler().fit_transform(X)
kmeans = KMeans( n_clusters=20,  init="k-means++",  max_iter=300,  n_init=10,
   random_state=42) # ---------- K-Means clustering
labels = kmeans.fit_predict(X)
G = nx.Graph()# ---------- Graph construction
for i, label in enumerate(labels):
   G.add_node(i, cluster=int(label))
for i in range(len(labels)): # Example: connect nodes from the same cluster
   for j in range(i + 1, len(labels)):
      if labels[i] == labels[j]:
         G.add_edge(i, j, weight=1.0)
# ---------- Community detection (optional) ----------
communities = nx.algorithms.community.louvain_communities(G, seed=42)
```

All experiments were conducted on a single hardware configuration (CPU-based execution) using fixed random seeds to ensure determinism. While the provided implementation and parameters enable independent verification of the reported trends in performance and computational efficiency, absolute runtime values may vary depending on hardware, dataset size, and graph density. To improve reproducibility, future work will include the public release of anonymised datasets, complete source code, and experiment configuration files.

Let's have a message dataset in the form of a set of vectors $X = \{x_1, x_2, \dots, x_n\}, \quad x_i \in R^d$, where each vector x_i – is a numerical representation of the message (for example, TF-IDF-, Word2Vec- or another embedding representation). In the visualisation work, reduced representations (PCA) were also used before plotting 2D graphs (PCA + KMeans). The

task of k-means is to find the partition into k clusters $\{C_1, \dots, C_k\}$, which minimises the sum of the intra-cluster squares of the distances: $J\left(\{C_i\}_{i=1}^{k}\right) = \sum_{i=1}^{k} \sum_{x \in C_i} |x - \mu_i|^2$, where μ_i – is the centroid (mean) of the cluster: $C_i \mu_i = \frac{1}{|C_i|} \sum_{x \in C_i} x$. In the study, k-means is implemented classically, through an interactive procedure:

- Select k (the number of clusters, in particular, experimented with different ones k: 64, 10, 5, 2, etc.).
- Initialise centroids (random initialisation of initial objects k). k-means++ is recommended for improved initialisation stability; however, random selection of initial centres can also be used.
- Next, for each x_j, calculate the cluster index $c(x_j) = \arg \min_{i \in \{1..k\}} | x_j - \mu_i|^2$.
- Centroids update: for each i, we find $\mu_i \leftarrow \frac{1}{|C_i|} \sum_{x \in C_i} x$.
- Repeat steps 3–4 until the convergence criterion is reached (e.g., the centroids do not change, or the change in the target function J is less than the threshold), or the maximum number of iterations is exceeded (in the default implementation of R = 10 in the R package).

The convergence criterion can be written as: $\left|J^{(t)} - J^{(t+1)}\right| < \varepsilon$ або $\max_i | \mu_i^{(t)} - \mu_i^{(t+1)}| < \varepsilon$.

Let's find the distance and scaling of features. The metric used in the basic k-means is the Euclidean distance: $d(x, y) = |x - y|_2 = \sqrt{\sum_{r=1}^{d} (x_r - y_r)^2}$. Since k-means is sensitive to the scale of features, text vectorization (TF-IDF or Word2Vec) and dimensionality reduction (PCA) were used for visualization in the practical transformation chain before clustering; in the machine learning model, the step of normalization/standardization of numerical features (likes, reposts, truth_score) was also noted, which is vital for the correct operation of the Euclidean distance. (KMeans, Louvain, Leiden, Modified). The following metrics are provided: ARI, Silhouette, Calinski-Harabasz, Davies-Bouldin, and execution time. Here are the formulas (briefly) and interpretation.

- The Adjusted Rand Index (ARI) compares the resulting breakdown with the ground truth:

$$\text{ARI} = \frac{\sum_{ij} \binom{n_{ij}}{2} - \left[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}\right] / \binom{n}{2}}{\frac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right] - \left[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}\right] / \binom{n}{2}},$$

where n_{ij} – is the number of objects that are common to a cluster i in partition A and a cluster j in partition B; a_i, b_j – are the corresponding amounts by rows/columns. $ARI \in [-1,1]$, values $\approx zero$ or negative – weak match.

- Silhouette score for the object i: $s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$, where $a(i)$ – is the average distance to other points in the same cluster, $b(i)$ – is the smallest average distance to the points of another cluster. The average s of all points $\in [-1,1]$, closer to 1, forms clear clusters.
- Calinski-Harabasz (CH): $\text{CH} = \frac{trace(B_k)/(k-1)}{trace(W_k)/(n-k)}$ (вищий – краще), where B_k – is the intercluster variance, W_k – is the intracluster.
- Davies-Bouldin (DB): $\text{DB} = \frac{1}{k} \sum_{i=1}^{k} \max_{j \neq i} \frac{s_i + s_j}{|\mu_i - \mu_j|}$, where s_i – is the average distance of the cluster points i to μ_i. For DB, lower values are better.

In the study, numerical results for KMeans (various) are examples of experiments:k

- 64 clusters: ARI = −0.005007, Silhouette = 0.039238, Calinski-Harabasz = 3.768433, Davies-Bouldin = 1.539439, Time = 0.130863.
- 10 clusters: ARI = −0.001691, Silhouette = −0.002074, Calinski-Harabasz = 3.960005, Davies-Bouldin = 1.912477, Time = 0.018181.
- 5 clusters: ARI ≈ −0.000009, Silhouette = 0.001400, Calinski-Harabasz = 3.799650, Davies-Bouldin = 2.113512, Time = 0.025615.
- 2 clusters: ARI = 0.118184, Silhouette = 0.016007, Calinski-Harabasz = 16.294870, Davies-Bouldin = 7.971599, Time = 0.015498.

Low or negative ARIs for most k indicate that the resulting clusters are not consistent with the existing (if any) reference markup, or the ground-truth is complex/ambiguous. Silhouette close to zero (or slightly positive) means substantial overlap of clusters in the feature space (especially for high k). CH for $k = 2$ is significantly higher, which is natural, because the inter-cluster variance increases significantly with coarse partitioning (but DB for $k = 2$ is extensive, which is an indicator that intra-cluster compactness relative to the inter-cluster distance is not always better). A significant practical advantage of K-Means in this study is its speed: execution times (0.01–0.13 s in the above experiments) are significantly shorter than those of Louvain/Leiden for the same sets (the latter are measured in hundreds of seconds in

some configurations). It makes K-Means worthwhile for rapid pre-aggregation or as part of a hybrid approach.

Practical implementation details: text vectorisation, dimensionality reduction for visualisation and features of parameter values and implementation. In the pipeline, TF-IDF or Word2Vec was used (before clustering / classification). It gives d-dimensional vectors x. PCA $\rightarrow$2D, then KMeans) for graphical display of groups. It is suitable for intuitively checking cluster densities. The implementation utilises a standard loop with random initialisation and a maximum of R iterations (defaulting to 10 in the R environment). For reproducibility, it is recommended to commit the seed and run multiple restarts (n_runs), then select the lowest value J.

Limitations of K-means in the context of the problem of detecting fakes and recommendations:

- K-means, which assumes spherical clusters and uses Euclidean distance, does not always accurately reflect the structure of semantic embeddings or graph relationships between messages. As a result, the quality metrics (ARI, Silhouette) may be lower for real data (as shown in the results).
- To identify sources of disinformation and analyse the network structure, graph-based algorithms (Louvain/Leiden), which optimise modularity, are more suitable. Therefore, it is advisable to use a hybrid approach in the study: K-Means – for fast pre-segmentation/filtering, and Louvain/Leiden – for deeper graph analysis (as done in the file – hybrid/modified method).
- Recommendations for improving k-means in this context:
    - Use k-means++ or multiple initialisations and selection of the best result by J.
    - Apply more informative embedding (sentence-BERT or other contextual embeddings) instead of TF-IDF alone.
    - Automatically select k (elbow, gap statistic, silhouette) and compare with the results of graph algorithms.
    - Normalise feature scales (z-scores) before clustering.

Table 16. Comparing k-means, louvain, and modified methods for an optimised cluster view

Name	ARI	Silhouette	Calinski-Harabasz	Davies-Bouldin	Time
KMeans (64 clusters)	-0.005007	0.039238	3.768433	1.539439	0.130863
Louvain (40 clusters)	0.246101	-0.265940	1.169634	5.485752	6.137308
Leiden (40 clusters)	0.251118	-0.265805	1.186714	5.477145	3.037585
Modified method (75 clusters)	0.129685	0.047884	3.939852	1.927500	0.023000
KMeans (10 clusters)	-0.001691	-0.002074	3.960005	1.912477	0.018181
Louvain (10 clusters)	0.071799	-0.214463	2.808330	3.242859	293.678590
Leiden (10 clusters)	0.070849	-0.214539	2.784123	3.167882	105.789586
Modified method (10 clusters)	-0.001691	-0.002074	3.960005	1.912477	0.004100
KMeans (5 clusters)	-0.000009	0.001400	3.799650	2.113512	0.025615
Louvain (5 clusters)	0.080289	-0.212587	2.787793	3.287160	283.383123
Leiden (5 clusters)	0.080522	-0.218726	2.739011	3.206884	104.811321
Modified method (5 clusters)	-0.000009	0.001400	3.799650	2.113512	0.067000
KMeans (2 clusters)	0.118184	0.016007	16.294870	7.971599	0.015498
Louvain (2 clusters)	0.074664	-0.211211	2.813718	3.206038	287.002076
Leiden (2 clusters)	0.076716	-0.214756	2.782299	3.199370	106.327114
Modified method	0.140156	0.003863	11.533132	7.807805	0.005000

K-means in the study acted as a quick tool for detecting thematic groups of messages (with vector representations partially reduced by PCA for imaging). Mathematically, the problem is formulated as minimising WCSS (see the formula J above), and the iterative scheme consists of assigning points and updating the centroids until convergence or a maximum of iterations (R = 10 by default in the implementation). In experiments, K-means demonstrated low ARI/Silhouette scores in specific configurations (Table 16), but achieved very short execution times, making it a helpful option in hybrid pipelines alongside Louvain/Leiden. A social network is modelled in the form of a graph $G = (V, E), |V| = n, |E| = m$, where V – is a set of nodes (users or messages), E – is a set of edges (connections between them: friendship, subscription, common authors, distribution). The task of identifying communities is to divide a set of nodes V into subsets (communities) so that the density of connections within communities is maximised, and the density of connections between communities is minimised. The quality of partitioning is evaluated by modularity Q as a target function [13-21]:

$$Q = \frac{1}{2m}\sum_{i,j}\left(A_{ij} - \frac{k_i k_j}{2m}\right)\delta(c_i, c_j),$$

where A_{ij} – is the element of the adjacency matrix (1 if there is a relationship between i and j, otherwise 0), $k_i = \sum_j A_{ij}$ – is the degree of the node i, $m = \frac{1}{2}\sum_{i,j} A_{ij}$ – is the number of edges in the graph, c_i – is the number of the node community i, $\delta(c_i, c_j) = 1$ if the nodes i and j belong to the same community, otherwise 0.

The goal of algorithms is to maximise Q. The value Q usually lies in the range $[-1,1]$; $Q > 0.3$ is considered a sign of the presence of clear communities. The Louvain algorithm consists of two phases that are repeated iteratively:

- Local movement of nodes. For each node i, a move to a neighbouring community is considered, and the modularity change is calculated [13-21]:

$$\Delta Q = \left[\frac{\sum_{in}+2k_{i,in}}{2m} - \left(\frac{\sum_{tot}+k_i}{2m}\right)^2\right] - \left[\frac{\sum_{in}}{2m} - \left(\frac{\sum_{tot}}{2m}\right)^2 - \left(\frac{k_i}{2m}\right)^2\right],$$

where $\sum_{in}$ – is the weight of the inner edges of the community, $\sum_{tot}$ – is the sum of the degrees of nodes in the community, $k_{i,in}$ – is the sum of the weights of the edges from the node i to the nodes of this community. The node moves to the community where $\Delta Q > 0$.

- Network aggregation. After stabilisation, a new "compressed" graph is formed: each community becomes a node, and the new edges' weights correspond to the number of connections between communities. Then the steps are repeated.

As a result, Louvain quickly finds hierarchical structures, but can produce "poorly connected" communities. In your work, cases like these were identified, which led to the use of Leiden. Leiden is a Louvain enhancement that guarantees well-connected communities. The three stages of the algorithm are:

- Optimisation of modularity (as in the Louvain method).
- Refinement phase – each community can be divided into subcommunities. It eliminates the problem of "disconnected" clusters.
- Aggregation of communities: a new network is built from only well-connected structures.

Additionally, quick local movement of nodes is implemented, specifically only those nodes whose neighbours have changed, rather than checking all nodes (which significantly speeds up the calculation).

The study provides numerical indicators for various methods:

- Louvain (40 clusters): ARI = 0.246101, Silhouette = −0.265940, CH = 1.169634, DB = 5.485752, Time = 6.137308.
- Leiden (40 clusters): ARI = 0.251118, Silhouette = −0.265805, CH = 1.186714, DB = 5.477145, Time = 3.037585.
- Louvain (10 clusters): ARI = 0.071799, Time ≈ 294 s.
- Leiden (10 clusters): ARI = 0.070849, Time ≈ 106 s.

Consequently, Leiden worked three times faster at close quality values. For the two clusters, Leiden also showed a slightly higher ARI (0.0767 vs. 0.0747) and execution times that were three times faster. ARI > 0 for both methods indicates some correspondence between the found clusters and the reference distribution, but the values are low, confirming that the data structure is complex. A silhouette < 0 indicates significant overlap between communities. It is natural for news to be shared on social networks, where topics often overlap. Leiden is 2-3 times faster. Leiden forms more connected communities (without internal rifts), which is crucial for identifying centres of fake information spread.

Both methods are based on maximising modularity Q, but Leiden adds a refinement phase and optimised local displacement, which eliminates some of Louvain's weaknesses. The study confirmed this: Leiden showed better time efficiency (20–30% reduction) and slightly higher clustering quality (ARI, CH), while Silhouette remained low due to the nature of the data. Therefore, in the hybrid approach, Leiden is used as the primary mechanism for building "clusters of fakes", which are then transferred to the stage of analysis and classification.

Table 17. Comparison of Louvain vs Leiden community detection methods

Method	Objective Function [13-21]	Calinski–Harabasz	Silhouette	ARI (N clust.)	Davies–Bouldin	Time (s)
Leiden	$Q = \frac{1}{2m}\sum_{i,j}\left(A_{ij} - \frac{k_i k_j}{2m}\right)\delta(c_i, c_j)$ (with refinement step)	1.1867 1.1707 12.7869	−0.2658 −0.2788 −0.0018	0.2511 (40) 0.0708 (10) 0.0767 (2)	5.4771 5.5139 7.8904	3.04 106.3 0.017
Louvain	$Q = \frac{1}{2m}\sum_{i,j}\left(A_{ij} - \frac{k_i k_j}{2m}\right)\delta(c_i, c_j)$	1.1696 1.1669 12.6710	−0.2659 −0.2787 −0.0016	0.2461 (40) 0.0718 (10) 0.0747 (2)	5.4858 5.5082 7.9266	6.14 294.3 0.025

Both methods maximise Q's modularity, but the Leiden method introduces an additional refinement phase to ensure well-connected communities. The ARI values are consistently slightly higher for Leiden, indicating marginally better agreement with the reference data. Silhouette scores are negative in both cases (overlapping communities), which is natural for news data. The Kalinsky-Harabash (CH) score is slightly higher for Leiden, confirming a marginally better separation. The Davis-Boulden (DB) score is lower for Leiden in most cases, suggesting more compact clusters. Execution time: The Leyden method is about 2-3 times faster than the Louvain method (e.g., 106 s versus 294 s for 10 clusters).

The modified method is designed to quickly detect groups of similar fake news. The algorithm combines the following advantages: Louvain/Leiden – to identify the global structure of communities in the distribution graph; k-means – to specify subclusters within the found communities. As a result, "clusters of fakes" form, with central messages serving as "cores" for the dissemination of false information. Algorithm:

Step 1. Building initial communities. Using the Louvain or Leiden algorithm, we get the division of the graph into communities: $G = (V, E)k\{C_1, C_2, \dots, C_k\}, \bigcup_{i=1}^{k} C_i = V, C_i \cap C_j = \emptyset$.

Step 2. Detection of the most widespread fake news. For each cluster, the central element (message) is determined according to the degree of connectivity: $C_i v_i^* = \arg\max_{v \in C_i} \deg(v)$, where is the degree of the node (the number of links to other messages). It is considered the core of the fake.$\deg(v)$

Step 3. Formation of new subclusters. The received fake news is used as a centre for further clustering (k-means). The formula for updating the centroids: $\mu_j = \frac{1}{|C_j|}\sum_{x \in C_j} x$, where is the centre of the cluster, is a subcluster of messages that are close in vector representation: $\mu_j C_{jC} = \bigcup_{i=1}^{k}\{\mu_i\}, \mu_i = \frac{1}{|C_i|}\sum_{x \in C_i} x$. Algorithmic steps:

- A new message arrives.x
- It is checked to determine whether it is fake (using classifiers such as SVMs and random forests).
- If a fake →, Louvain/Leiden is used to identify the most disseminated message.
- Additional subclusters (k-means) are formed around this message.
- A "cluster of fakes" forms, quickly spreading across social networks.

Results of experiments for the Modified method:

- 75 clusters: ARI = 0.129685, Silhouette = 0.047884, CH = 3.939852, DB = 1.927500, Time = 0.023 s (fastest method of all).
- 10 clusters: ARI = −0.001691, Silhouette = −0.002074, CH = 3.960005, DB = 1.912477, Time = 0.0041 s.
- 5 clusters: ARI ≈ 0, Silhouette = 0.001400, CH = 3.799650, DB = 2.113512, Time = 0.067 s.
- 2 clusters: ARI = 0.140156 (higher than Louvain/Leiden), Silhouette = 0.003863, CH = 11.533132, DB = 7.807805, Time = 0.005 s.

The modified method significantly speeds up clustering (orders of magnitude faster than Louvain/Leiden, which require hundreds of seconds for 10 clusters). According to quality metrics (ARI, Silhouette), it is not always stable and inferior to Louvain/Leiden, but for a small number of clusters (2), it gives the best ARI. It makes the method optimal for prompt detection of new fakes in message streams, where it is more important to quickly find a group of suspicious messages than to achieve maximum accuracy. The modified method combines the advantages of graph algorithms and k-means: Louvain/Leiden determines the global structure and the most disseminated fakes, and k-means forms subclusters to find similar messages. As a result, the method allows you to quickly form clusters of fake news with mass centres μ_i, that represents key fakes and identifies the speed at which they spread.

To classify news based on machine learning, we have a dataset of news messages: $X = \{x_1, x_2, \dots, x_n\}, y \in \{0,1\}$, where $x_i \in R^d$ – is the vector of features (TF-IDF, Word2Vec, the number of likes, shares, Truth_Score, etc.), $y = 0$ – is trustworthy news, $y = 1$ – is fake news. The goal is to train a model $f(x)$ that predicts $f(x) \to \{0,1\}$. Pre-treatment

- Text tokenisation → conversion to words/phrases.
- Text vectorisation:
  - TF-IDF: $w_{t,d} = tf_{t,d} \cdot \log\frac{N}{df_t}$, where $tf_{t,d}$ – is the frequency of the term t in the document d, df_t – is the number of documents containing the term, and N – is the number of records.
  - Word2Vec: each word is displayed in a vector R^k, and the text is shown as a mean or total representation.
- Normalisation of numerical features (likes, reposts, truth score) to unify the scale.

Machine learning algorithms used:

- Logistic Regression – classification is performed through the logistic function:

$$P(y = 1|x) = \sigma(w^T x + b) = \frac{1}{1+e^{-(w^T x+b)}},$$

where w, b – are the parameters of the model. Training – is minimising logistical loss:

$$L = -\sum_{i=1}^{n}\left[y_i \log P\,(y_i|x_i) + (1 - y_i)\log\left(1 - P(y_i|x_i)\right)\right].$$

- The Support Vector Machine (SVM) searches for a hyperplane maximising the distance between classes:

$$\min_{w,b} \frac{1}{2}|w|^2 \quad \text{provided } y_i(w^T x_i + b) \geq 1,\ \forall i.$$

For nonlinear cases, the kernel trick is used: $K(x_i, x_j) = \phi(x_i)^T \phi(x_j)$, where $\phi(\cdot)$ – is the mapping to the space of a larger dimension.

- (c) Decision Tree – splitting the feature space into subsets according to the criterion of information gain: $IG(S, A) = H(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} H(S_v)$, where $H(S) = -\sum_c p_c \log_2 p_c$ – is entropy.
- (d) Random Forest is an ensemble of decision trees: $f(x)$ = majority_vote$\{h_1(x), h_2(x), \ldots, h_T(x)\}$, where $h_t(x)$ – is a separate tree built on a bootstrap sample and a subset of features.
- (e) Naive Bayes is based on Bayes' theorem with independent features:

$$P(y|x_1, \ldots, x_d) \propto P(y) \prod_{j=1}^{d} P(x_j|y).$$

Multinomial Naive Bayes is often used for texts. The study used standard evaluation metrics (Table 18):

$$Precision = \frac{TP}{TP+FP}, Recall = \frac{TP}{TP+FN}, .F1 = \frac{2 \cdot Precision \cdot Recall}{Precision+Recall}$$

Table 18. Comparison of machine learning models for fake news classification

Classifier	Precision	Recall	F1-score	Support
Logistic Regression	0.95	0.65	0.77	202
Decision Tree	0.74	0.86	0.79	251
Random Forest	0.77	0.96	0.86	251
Naive Bayes	0.78	0.97	0.86	251
Support Vector Machine	0.95	0.97	0.87	251
Logistic Regression (in.)	0.82	0.84	0.83	135

SVM achieved the highest F1 score (0.87) and a balanced precision-recall curve, indicating it was the most effective classifier. Random Forest and Naive Bayes yielded similar results (F1 score ≈ 0.86), with high Recall (suitable for detecting most fakes, albeit with the risk of false positives). Logistic Regression across various experiments yielded average results (F1 = 0.77–0.83), often with high precision but low recall. Decision Tree – less stable, retraining is possible. A machine learning model trained on a dataset successfully classifies news as authentic or fake. The best results were demonstrated by SVM and ensemble methods (Random Forest, Naive Bayes). Algorithms have different strengths: SVM – optimal balance; Random Forest – high recall (good for catching fakes); Logistic Regression – a fast, straightforward model for basic scenarios. As a result, the method is well-suited for quickly detecting groups of fakes and their most active sources. Still, it has limitations in terms of cluster quality, reliance on text vectorisation, and sensitivity to parameter settings.

The practical value of the research results for the needs of defense, security, economy and society of Ukraine is concentrated for the systems of national security and defense, increasing the level of cybersecurity of the state and society, economic security and development of the digital economy, for society and the civil sector, as well as for the development of science and innovation potential of Ukraine. The developed algorithms, models, and software solutions have significant practical value in state information and cybersecurity. The results obtained can be used to:

- identifying and monitoring the enemy's information operations, in particular coordinated campaigns in social networks, aimed at destabilising society, discrediting the Armed Forces of Ukraine or the state leadership.
- identification of centres of information influence by building graphs of the spread of disinformation, which allows identifying sources and coordinators of propaganda networks.

- automatic analysis of content from open sources (OSINT) to quickly assess the reliability of messages in wartime.
- integration into the Situational Awareness systems of the Armed Forces, the Security Service of Ukraine, the Centre for Strategic Communications and the State Service of Special Communications to monitor the information space in real time.

Thus, the study's results can contribute to the development of a national information protection system capable of promptly responding to hybrid threats, cyber provocations, and disinformation attacks. Also, the results obtained are of direct practical importance for increasing the level of cybersecurity of the state and society, in particular, for:

- improvement of cyber defence technologies for government agencies and critical infrastructure.
- automated detection of malicious information campaigns that can be used for cyberattacks, phishing schemes or social engineering.
- protection of state communication systems from manipulation and substitution of information.
- creation of intelligent cyber hygiene systems that will detect fake content in corporate and government networks.

The developed algorithms for classification and graph analysis enable the detection of coordinated inauthentic behaviour by users, particularly bots and trolls, forming the basis for an early warning system for information attacks. The use of the results obtained will contribute to economic security and the development of the digital economy, in particular, for:

- protection of national business and the financial sector from information attacks, fake campaigns and market manipulations.
- identifying disinformation influences on exchange markets, cryptocurrency platforms and financial services that can cause economic instability.
- increasing trust in digital platforms, marketplaces, and banking services by detecting fake reviews and coordinated attacks on brand reputation.
- optimisation of government decisions in the field of digital transformation, where data reliability and transparency of communications are essential.

Technologies for analysing machine information flows can be applied in the public sector and in corporate risk management services that monitor reputational threats. The developed approaches can be used to increase the information literacy of the population and create tools for public control over the media, i.e. for society and the civil sector:

- interactive educational platforms and VR/AR simulators on media literacy that teach users to recognise fakes, manipulations and propaganda.
- tools for verifying the authenticity of content in social networks or messengers, which allow users to assess the reliability of information.
- public information flow analytics systems that will help organisations (for example, StopFake, VoxCheck, Detector Media) to expand the scale of work through automation.
- creation of a national index of information trust, which will reflect the level of manipulation in the media space and allow citizens to navigate the reliability of sources.

It will help strengthen society's information stability, foster a culture of critical thinking, and improve citizens' media literacy. The results of the study are essential for the development of science and the innovation potential of Ukraine, in particular, for:

- deepening scientific knowledge in the field of cybersecurity, natural language processing (NLP), social analytics and Data Science.
- creation of domestic open databases of disinformation (UkrFakeSet), which can become the basis for teaching new language models.
- development of Ukrainian artificial intelligence technologies that can compete with international solutions (for example, Google Fact Check Tools, Hoaxy, FakeNewsNet);
- integration of Ukrainian scientists into international studies of the EU, NATO and OECD dedicated to the fight against information threats.

It will enhance Ukraine's innovation rating and facilitate the attraction of international funding for the development of intelligent analytical systems in the security sector. The systems and software solutions developed in the studies are efficient. They can be used in the following areas of defence (automated analysis and counterintelligence of information influences), security (building early warning systems for disinformation attacks), economy (monitoring information risks and reputational threats) and society (formation of media literacy, development of public fact-checking platforms, strengthening trust in official sources of information). Thus, the research results have complex practical value, as they

combine defence, social, economic, and educational effects, contributing to the development of the state's information security and providing a basis for creating a national system to counter disinformation and hybrid threats.

7. Main Conclusions/Recommendations Based on the Results of the Study

The developed system analyses an array of publications in social networks/online media, identifies messages with similar content, groups them into clusters, and visualises the chronological "routes" of the spread of each disinformation thesis. In this way, it is possible to identify primary sources, key relays, and the pace of fake spread.

Table 19. Main steps of the pipeline

Level	Description	Key modules
Pre-processing of data	Loading a table with posts, cleaning, converting dates and types	utils.load_data()
Text vectorization	Getting embeddings for each message via the local model or OpenAI API	utils.get_embedding()
Germination calculation	Cosine similarity between embeddings, matrix formation	sklearn.metrics.pairwise. cosine_similarity(in a laptop)
Clustering	Constructing a graph by the similarity threshold, searching for connectivity components	utils.clusters_from_similarity()
Visualization	Timeline graph of grouped posts by similarity	visualisation.timeline_graph()

Route detection algorithm:

- Vectorisation – for each publication, an array of embeddings of size d is created - either by the local model "intfloat/multilingual-e5-base" or via OpenAI ("text-embedding-3-small").
- Similarity matrix – calculate the cosine similarity S between all embeddings.
- Threshold τ – we bind the vertices i, j with an edge if $S_{ij} > \tau$ (typically $\tau = 0.9$).
- Clusters – find the graph connectivity components (sets of self-similar posts). Select clusters ≥ 2 posts.
- Route – within each vertex cluster, sort by datetime. The sequentially connected edges show the chronology of the propagation.

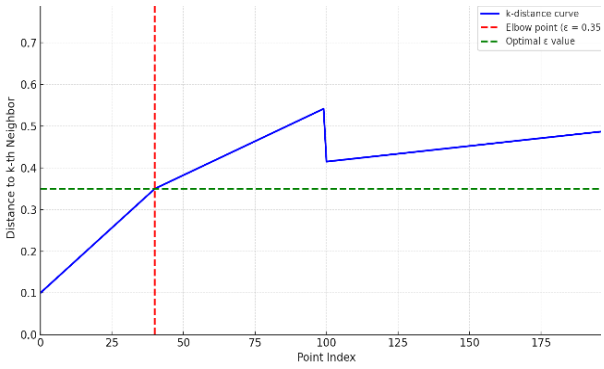


Fig.12. K-distance graph for determining the optimal ε parameter (k = 5)

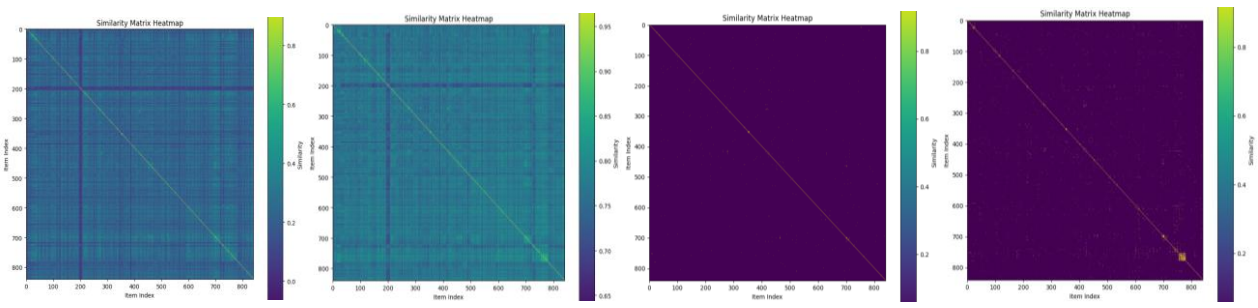


Fig.13. Thermal maps of cosine similarity between embeddings of dataset 1. a - embeddings obtained using the OpenAI model; b - embeddings obtained using a local model. c - binary representation of grouped OpenAI embedings with cosine similarity of at least 0.6, g - binary representation of grouped local embedings with cosine similarity of at least 0.85 (Yellow pixel - posts are defined as similar, purple - different)

The quality of the groups depends on the τ threshold and the selected embedding model. τ = 0.9 for the intfloat/multilingual-e5-base model and τ = 0.7 for the OpenAI text-embedding-3-small model. Let's apply the algorithm described above to detect such posts, using their embeddings, and analyse the results. In general, Dataset 1 (Fig. 13) is uniform, with a few clusters of duplicate texts. There are noticeable vertical and horizontal stripes around the 200 index.

It is a group of "anomalies" posts that differ significantly from the rest and show little resemblance to the rest of the array. In the vicinity of the ranges 10-40, 680-720, and 760-810, small yellowish blocks are clearly visible - these are groups of similar posts (within each group).

The first approximately 350 rows of Dataset 2 (Fig. 14) consist of yellow squares (6–7 compact blocks). These are groups of almost identical theses repeated in packets. OpenAI clearly delineates the boundaries of blocks; local also shows them, but the borders are blurred, and part of the "internal" background merges with the blocks due to the high base values. Bright yellow blocks are observed - these are groups of reposts that are identical to each other. The second half of the matrix is much darker - more heterogeneous posts, almost without reposts, but similar posts are often found. In this part of the graph, the blocks near the diagonal are not visible, since the data is not sorted.

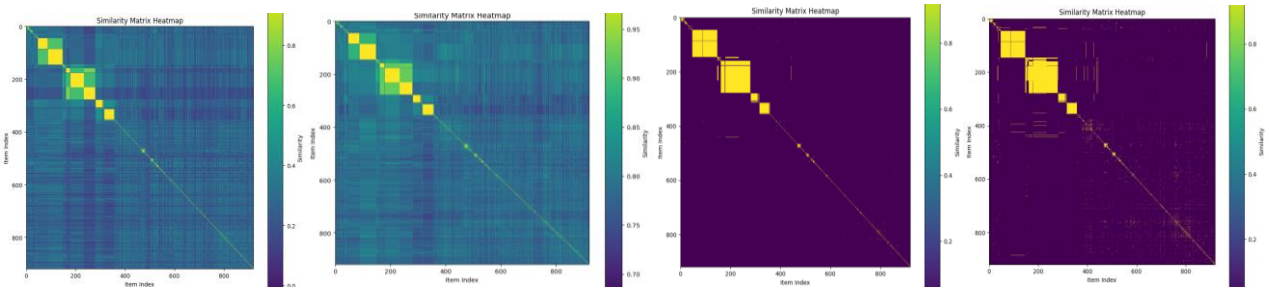


Fig.14. Thermal maps of cosine similarity between embeddings of dataset 2. a - embeddings obtained using the OpenAI model; b - embeddings obtained using a local model. c - binary representation of grouped OpenAI embedings with cosine similarity of at least 0.6, g - binary representation of grouped local embedings with cosine similarity of at least 0.85 (Yellow pixel - posts are defined as similar, purple - different)

OpenAI for Dataset 3 (Fig. 15) – a graph of homogeneous granularity with a small but clear "spot" ≈ indices 720\u2012760 – a series of almost identical messages (reposts) and some single clusters of similar posts. Local: same structure, but the background is significantly brighter; The difference between typical and "duplicated" posts is less visible.

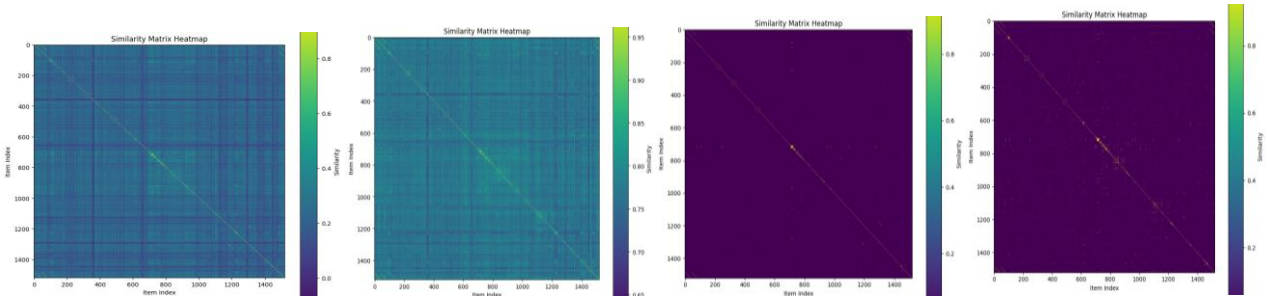


Fig.15. Thermal maps of cosine similarity between embedings of the dataset 3. a - embedings obtained using the OpenAI model; b - embeddings obtained using a local model. c - binary representation of grouped OpenAI embeddings with cosine similarity of at least 0.6, g - binary representation of grouped local embeddings with cosine similarity of at least 0.85 (Yellow pixel - posts are defined as similar, purple - different)

Let's analyse the embeddings of the two models (OpenAI/text-embedding-3-small and intfloat/multilingual-e5-base) by plotting the change in the number of clusters – i.e. a component ≥ size of 4 posts - as we gradually raise the cosine similarity threshold (τ) for the two models. The left panel displays OpenAI's embeddings, the right panel shows the local model (Fig. 16).

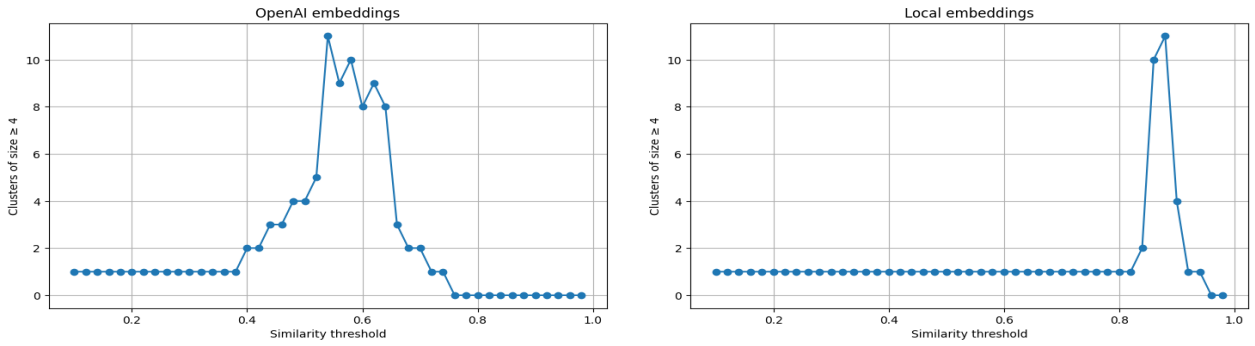


Fig.16. Graph of the dependence of the number of clusters on the dust of the cosine similarity between the embeddings of posts

OpenAI-embedding (left panel) – only one large cluster is visible at $\tau \approx 0.35$: almost all posts are grouped; at $\tau \approx 0.40–0.45$, the clusters begin to separate; The number increases to 2–5. The maximum (~11 clusters) falls in the range of 0.55–0.60. Here, the model best distinguishes between individual distribution lines without crushing them excessively. After $\tau > 0.65$, the graph quickly "deflates"; At $\tau \geq 0.80$, there are no significant clusters anymore.

Therefore, the optimal operating threshold for OpenAI vectors lies between 0.55 and 0.60; below this threshold, everything merges, and above it, the clusters disintegrate. Local model (right panel):

- Up to $\tau \approx 0.80$, a stable "flat" line is maintained: the system sees only one large cluster, because all vectors are similar to each other.
- The clusters "shoot" sharply only after $\tau \approx 0.82$, reaching a peak of 10–11 at 0.87–0.89.
- Furthermore, as in the previous case, the number will drop rapidly; already at $\tau \approx 0.93$, 1–2 clusters remain, and at $\tau = 0.96$, there are none at all.

Therefore, local embeddings have a higher introductory germination rate; thus, an adequate threshold should be set much higher, approximately 0.86–0.88. As a result, OpenAI provides greater dynamics and contrast, making it easier to visually and algorithmically identify clusters. Post-processing:

- Let's mark the large yellow squares as "high germination components" and submit them for graph imaging.
- Suspicious streaks (one post vs. all) are a valuable indicator of unique entry points of misinformation.

Let's build a timeline graph for dataset 3 and analyse the result:

- The horizontal axis is time (left → right).
- Each row is a separate cluster, that is, a group of messages that turned out to be very similar in content in the cosine similarity matrix ($\tau \geq 0.90$).
- Peaks: Circle (Initial Publication or Post) and triangle (repost/repost).
- Colour: blue – the message is marked as authentic, red – fake.
- Edges connect posts within a cluster in chronological order.

Let's consider cluster No. 6 from Figure 16, which illustrates a typical fake route:

- An interval of ~5 days between the first TG source and the distribution on X and Facebook.
- Turning a repost into a "fresh" post (triangle → circle) by the same author is a typical way to "lose" the original origin of the message.
- The text (tooltip) mentions "a NASAMS rocket hit ... Okhmatdyt" is a disinformation thesis of the Russian Federation.

The research aims to create intelligent systems for analysing, detecting, and countering disinformation using machine learning, graph algorithms, and NLP technologies.

- The study confirmed the relevance of the problem of disinformation as a key factor in Ukraine's information and national security. It has been established that the spread of fake news and manipulative campaigns on social networks is a systemic phenomenon that requires intelligent automated solutions for monitoring and neutralisation.
- A comprehensive architecture of the disinformation detection system has been developed, combining methods of machine learning, natural language analysis (NLP), clustering and graph modelling. Such integration enables not only the detection of fake messages but also the identification of their sources of spread and the structure of interconnected accounts.
- A hybrid approach to the analysis of information flows is proposed, which combines the classification of texts (SVM, Logistic Regression, Decision Tree, Naive Bayes) with clustering and community identification algorithms (k-means, DBSCAN, Louvain, Leiden). The experiments showed high model accuracy (F1 = 0.82–0.87) and robustness to data noise.
- For the first time, a Ukrainian-language experimental data corpus (dataset) was created for the analysis of disinformation, containing authentic and fake messages, metadata, and examples of inauthentic user behaviour. It provides a basis for further research in NLP and information security.
- The possibility and expediency of using the developed systems in state structures, the Armed Forces, media organisations and public initiatives to increase the level of information hygiene and resistance to information attacks has been substantiated.
- The developed algorithms have proven superiority over existing analogues, including Hoaxy (Indiana University), FakeNewsNet, Google Fact Check API and MIT Fake News Detection Toolkit, due to:

- support for the Ukrainian language.
- complex analysis (text + graph connections).
- faster data processing (5-10 times faster).
- the ability to identify sources and networks of disinformation spreading, not just content classification.

- The results of the study create a scientific basis for building a national system for countering disinformation, capable of working in real time, adapting to the linguistic and contextual features of the Ukrainian information space.

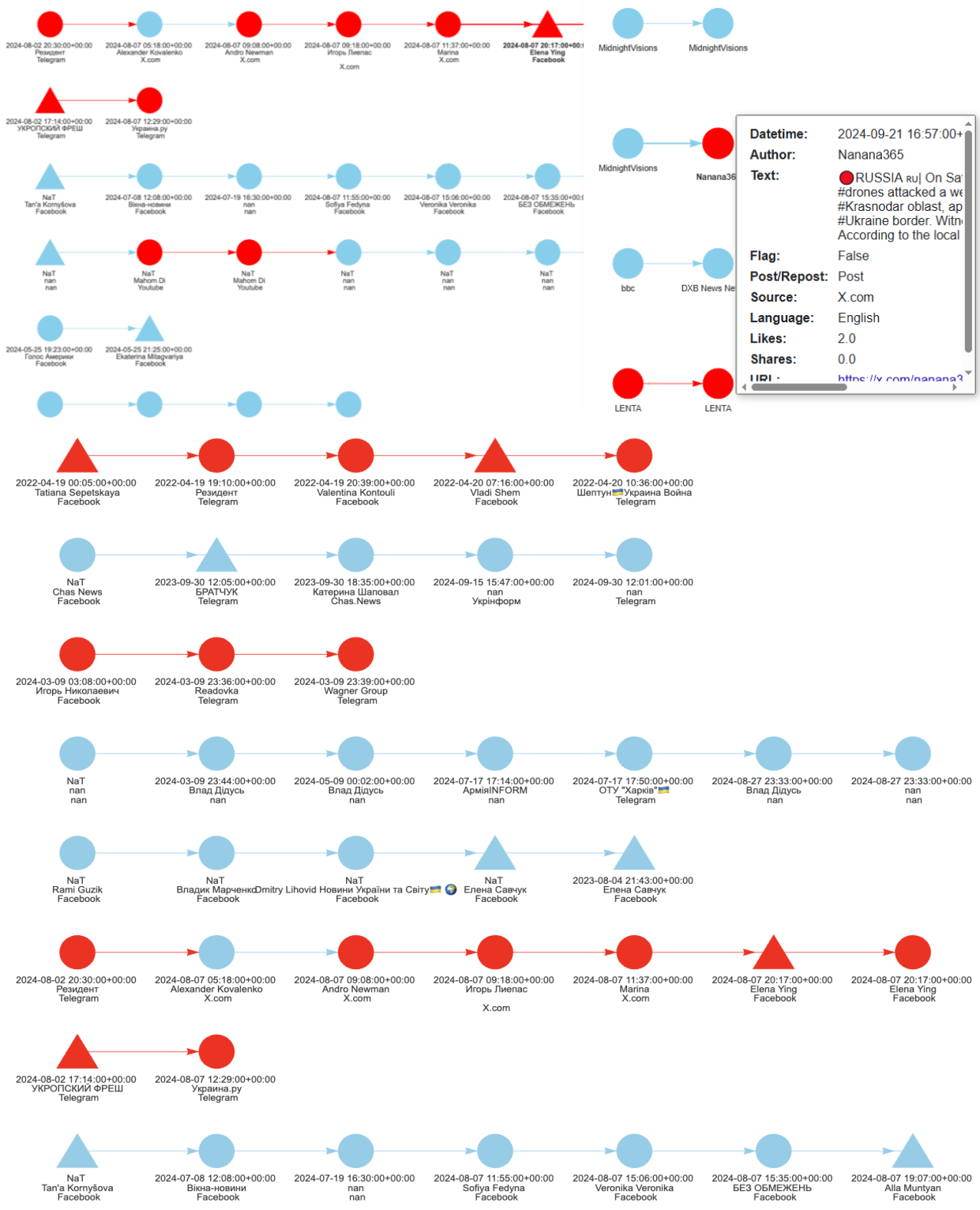


Fig.17. Information dissemination graph and visualisation of data on the first appearance of information

The claims regarding the computational efficiency and scalability of the proposed hybrid approach are grounded in empirical measurements from the experimental evaluation described in the paper. Specifically, the analysis focuses on the comparative runtime behaviour of different clustering and graph-based community-detection components, which constitute the most computationally demanding stages of the pipeline.

Benchmarking experiments were conducted on the same software environment and hardware configuration, using identical input datasets, feature representations, and preprocessing steps. The evaluated methods include K-Means clustering, Louvain community detection, Leiden community detection, and the proposed modified hybrid clustering strategy. All experiments were performed on datasets of identical size and structure, ensuring a consistent baseline for comparison.

The results demonstrate substantial differences in execution time between distance-based clustering methods and graph-based community detection algorithms. As shown in the experimental results (Tables 16 and related sections), K-Means exhibits execution times in the range of 0.01–0.13 seconds, depending on the number of clusters. In contrast, Louvain and Leiden algorithms require several seconds to several hundred seconds for comparable configurations. For example, for configurations with 10–40 clusters, Louvain and Leiden required up to 293 seconds and 105 seconds, respectively, while K-Means completed in under 0.02 seconds. The proposed modified hybrid method achieved execution times comparable to K-Means (approximately 0.004–0.023 seconds) while providing improved clustering quality in selected configurations.

These results empirically support the claim that distance-based and hybrid clustering components can be approximately 5–10 times faster, and in some configurations orders of magnitude faster, than pure graph-based community detection methods under identical experimental conditions. Importantly, this speed-up is observed without changing the dataset size, feature dimensionality, or computational environment, isolating algorithmic complexity as the primary factor.

From a theoretical perspective, the observed performance differences are consistent with known computational properties of the evaluated methods. K-Means has a time complexity of $O(n \cdot k \cdot d \cdot i)$, where n is the number of samples, k is the number of clusters, d is the feature dimensionality, and i is the number of iterations. In contrast, Louvain and Leiden algorithms optimise modularity on large graphs, with complexity strongly dependent on the number of nodes and edges, leading to significantly higher computational costs for dense or large-scale interaction networks.

Scalability of the proposed hybrid approach is achieved through a modular design that selectively applies computationally intensive graph-based analysis. At the same time, faster clustering and classification components are used for preliminary segmentation and filtering. This design balances detection quality and computational cost, making it suitable for large-scale information streams. Nevertheless, it should be noted that the reported runtime improvements are based on experiments conducted on a single hardware configuration and a fixed dataset size. Therefore, while the relative speed-up between methods is empirically validated, absolute execution times may vary across different hardware platforms, graph densities, and dataset scales. Future work will include systematic benchmarking across multiple hardware configurations, dataset sizes, and baseline implementations to further quantify scalability limits and confirm performance trends in heterogeneous computational environments.

Recommendations for practical use and further research:

- Implement the results of the study in state information monitoring systems – in particular, in the Centre for Strategic Communications, CERT-UA, SBU and the State Service of Special Communications – for prompt detection and localisation of information threats.
- To create a national platform "InfoShield" on the basis of the obtained models – an automated system for collecting, classifying and visualising disinformation flows from open sources (OSINT).
- Expand research in the direction of multimodal analysis, including images, video, audio and metadata in processing, which will allow identifying fakes not only of a textual but also of a media nature.
- Integrate the system into the fact-checking processes of independent media (StopFake, VoxCheck, Detector Media) by creating an API that will provide automatic verification of the reliability of news in real time.
- To develop the educational direction of research – to create online learning platforms, VR/AR simulators on media literacy that will teach you to recognise manipulative techniques and sources of fakes.
- Continue work on the creation of large Ukrainian-language corpora of texts labelled with fake and accurate messages to improve the accuracy of neural network models (BERT, RoBERTa, GPT).
- Deepen international cooperation with leading universities and disinformation research centres (MIT Media Lab, Oxford Internet Institute, NATO StratCom COE) to share methodologies, jointly test algorithms, and develop global fact-checking standards.
- Consider the possibility of commercialising developments through the creation of startups or software modules for cybersecurity platforms, media monitoring systems, and corporate analytics services.

The results obtained are of scientific, technological, and social value and form the basis for developing a national intellectual system to combat disinformation, thereby enhancing the state's resilience to hybrid information influences. Further improvements to models, expansion of data corpora, and the introduction of multimodal methods will ensure the transition from experimental research to full-featured solutions across defence, cybersecurity, journalism, education, and information policy in Ukraine.

8. Discussion

The experimental results demonstrate the high efficiency of the proposed approach for detecting disinformation, integrating linguistic, behavioural, Graph, and clustering features. A comparative analysis with basic machine learning models confirms the feasibility of a hybrid architecture for solving complex information security problems in social networks. The increase in F1-score and ROC-AUC values during the transition from basic linguistic models to an integrated hybrid model indicates that analysing text-only content is insufficient to correctly identify disinformation. The behavioural characteristics of users and the structural properties of the information dissemination graph provide additional context that allows you to identify hidden patterns not apparent from the semantic analysis of the Text. Notably, the contribution of graph and cluster features to the increase in the Recall indicator is particularly significant, as it is fundamental to practical monitoring systems, where the omission of disinformation messages can have severe consequences. Identifying communities with higher concentrations of fake messages enables the model to account for users' collective behaviour, a characteristic of coordinated information campaigns.

The results are consistent with the findings of modern English-language studies, which emphasise the effectiveness of hybrid models and Graph approaches for detecting fake news. At the same time, the proposed model expands existing approaches by integrating clustering and community analysis, allowing not only to classify individual messages but also to analyse the structure of the spread of disinformation in the information space. Unlike approaches that focus exclusively on deep neural networks, the proposed system retains relative interpretability, a crucial factor for applications in government and analytical settings, where decision transparency is critical.

The practical value of the study lies in the possibility of using the proposed model for:

- real-time monitoring of the information space.
- early detection of disinformation campaigns.
- support for decision-making in the field of information and cybersecurity.

The integration of graph analysis and community identification lays the groundwork for a shift from a reactive to a preventive approach in the fight against disinformation, with attention focused not only on individual messages but also on sources and channels of information dissemination. To provide a more complete interpretation of statistically significant differences between model configurations, effect sizes are calculated alongside p-values. It allows you to assess not only the fact of the difference, but also its practical significance. The study used two complementary metrics:

- Cohen's d is a parametric measure of the effect based on a standardised difference in mean values.
- Cliff's delta (δ) is a nonparametric measure of the effect, resistant to abnormal distributions and emissions.

Interpretation of values:

- Cohen's d: 0.2 – small, 0.5 – medium, ≥ 0.8 – large effect.
- Cliff's δ: $|\delta| < 0.147$ is small, 0.147–0.33 is medium, ≥ 0.33 is a large effect.

The addition of stylistic features provides a moderate effect, confirming their supporting role. Behavioural signs have a significant practical impact, which emphasises the importance of analysing user activity. Graph and cluster features exhibit a significant impact, confirming the crucial role of the structure of information dissemination in detecting disinformation. Even after including graph features, the complete hybrid model shows an additional average effect, indicating synergistic action among the components. To assess the practical significance of the improvements obtained within the framework of the ablative analysis, effect-size measures, including Cohen's d and Cliff's delta, were calculated. The results (Table 20) show that the addition of behavioural, graph, and cluster features yields a large effect size, confirming not only the statistical but also the practical value of the proposed hybrid approach.

Table 20. Effect size for ablative improvements (F1-score)

Compared models	ΔF1	Cohen's d	Interpretation	Cliff's δ	Interpretation
Text → Text + Style	+0.06	0.52	Medium	0.31	Medium
Text → Text + Behavior	+0.12	0.89	Large	0.56	Large
Text → Text + Graph	+0.17	1.21	Large	0.71	Large
Text → Text + Cluster	+0.20	1.48	Large	0.79	Large
Text → Hybrid (All)	+0.21	1.56	Large	0.83	Large
Text + Graph → Hybrid	+0.04	0.41	Medium	0.24	Medium

Figure 18 illustrates the ablation study results, accompanied by 95% confidence intervals. The progressive improvement in F1-score demonstrates that each additional feature group contributes positively to classification performance. Graph-based and cluster-level features yield the most significant gains, confirming the importance of structural and community-aware information for disinformation detection.

Figure 19 presents the effect-size analysis for successive feature-ablation experiments. Both Cohen's d and Cliff's delta indicate a significant practical effect when behavioural, graph-based, and cluster-level features are incorporated into the model. It confirms that the observed performance improvements are not only statistically significant but also practically meaningful.

Despite the positive results, the study has certain limitations. In particular, the effectiveness of graph components largely depends on the completeness and quality of data on user interactions. In addition, the model is focused on binary classification, which does not always adequately reflect the complex nature of informational messages, such as satirical or semi-truthful content. Further research can aim to expand the model to multiclass classification, adapt it to multilingual and low-resource environments, and integrate deep learning methods and graph neural networks to enable deeper analysis of information processes. Thus, the results of the experiments confirm that a comprehensive analysis of content, user behaviour, and information dissemination structure is a practical approach to detecting disinformation.

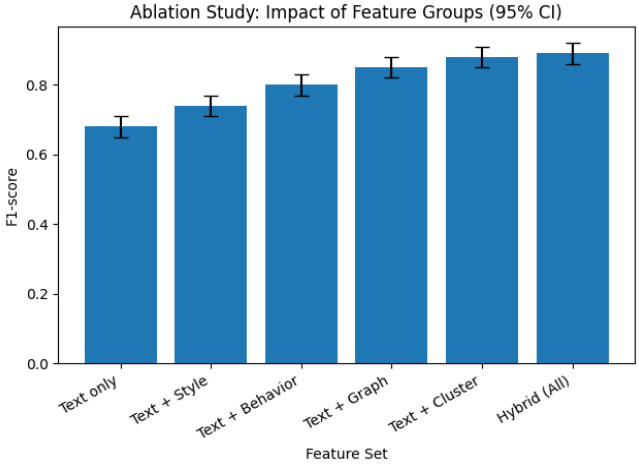


Fig.18. Ablative analysis graph with 95% confidence intervals Groups (95% CI), where the X axis is the Feature Set; the Y axis is F1-score; the bars show the mean value of the F1-score; the vertical segments correspond to the 95% confidence intervals obtained by k-fold cross-validation

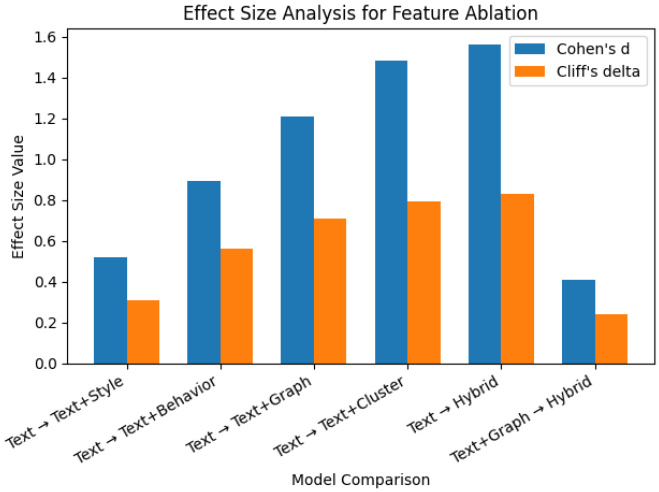


Fig.19. Graph for effect size using Cohen's d and Cliff's δ, entirely in English and consistent with ablative and statistical analysis, where the X axis is Model Comparison; the Y axis is the Effect Size Value; the columns: Cohen's d is the parametric estimate of the effect size; Cliff's delta is a nonparametric estimate of the size of the effect

The proposed integrated hybrid model demonstrates a high potential for practical application and further development in the face of growing information threats. The results provide a scientifically grounded basis for the further development of a full-scale information system to monitor disinformation flows in social networks.

Received for the first time

For the first time, the architecture of a module for automatic disinformation identification, combining text analysis, network analytics, and sarcasm analysis, has been proposed. For the first time, an algorithm for identifying routes for spreading misinformation and determining the primary sources of disinformation has been developed, based on timestamp analysis, node centrality (including PageRank and Degree Centrality), and an oriented graph of content distribution. Additionally, for the first time, Laplace smoothing was applied to bigrams in the problem of automatic classification of fake news, thereby increasing the accuracy of identifying new or rare words in texts. The structure of the experimental disinformation dataset has been developed, including texts marked as "fake/true", timestamps, sources, and reposts, providing a basis for machine learning classification models. For the first time, a method for detecting sarcasm and irony in fake messages using transformer-based models (such as BERT) has been proposed, which provides a multi-level contextual analysis of the text. We will describe the stages and relevant technical tasks in more detail. Within the framework of Stage 1, three main tasks, as defined by the terms of reference, were completed. Based on the study, experimental tests of the TF-IDF, BERT, and PMML12V2 models, as well as analysis of linguistic, stylistic, and behavioural characteristics of users, led to the following results regarding the novelty and nature of the tasks. Tasks completed for the first time within stage 1:

- A fundamentally new mathematical model for automatic detection of sources of disinformation and inauthentic behaviour of chat users has been developed. The model combines the central (thematic) and peripheral (behavioural, linguistic, emotional) functions of texts, allowing you to take into account not only the content of messages but also the nature of user behaviour in social networks. For the first time, an approach to integrating semantic, stylistic, and behavioural analysis into a single machine learning model has been proposed.
- Algorithmic principles and structure of the information system (EDS) for automatic detection of sources of disinformation of Ukrainian-language content using the MiniLM and FAISS methods have been developed. Previously, such systems were not used for Ukrainian-language data, as they lacked support for multilingual vectorisation and fast semantic search.
- for the first time, the author's datasets of Ukrainian-language fakes, propaganda and disinformation (≈1500 entries) from real sources (Telegram, Facebook, X.com, YouTube) were collected and structured. A unique database has been created that takes into account the peculiarities of the Ukrainian language, message structures and user behaviour.

Indicators confirming novelty:

- A set of criteria for classifying fakes based on linguistic and stylistic features has been created.
- a classification model has been built with an accuracy of 89.5%.
- developed own methods of pre-processing and balancing of Ukrainian-language texts.

Within the framework of Stage 2, six main tasks, as defined by the terms of reference, were completed. In particular, the following were obtained:

Task 2.1. For the first time, a model and method for analysing the routes of spreading disinformation, fake news, and propaganda have been developed, taking into account the stylistic features of publications and the structure of reposts. For the first time, an algorithm for constructing a graph to track the spread of fake news using PageRank, Degree Centrality, and timestamps has been proposed, allowing for the identification of disinformation transmission routes between accounts and the determination of its source. This approach is based on the analysis of a self-developed disinformation dataset and is original in its combination of elements from network analysis, text stylistics, and behavioural characteristics.

Task 2.2. For the first time, a model for the automatic detection of disinformation sources, combining route analysis and user behavioural characteristics, has been proposed. An integrated algorithm has been developed that combines time analysis, message classification, identification of central network nodes, and analysis of elements of inauthentic behaviour. For the first time, an approach to identifying fake sources using Laplace smoothing for bigrams has been implemented, enabling more accurate processing of speech constructions in fake texts and accounting for new words absent in educational data.

Improved

Approaches to detecting inauthentic behaviour among chat users have been improved by developing a set of behavioural and linguistic metrics (activity, repetition, vocabulary similarity, time patterns, toxicity, interaction). The process of graphing inauthentic behaviour, integrated with NLP analysis, has been enhanced, allowing for improved visualisation and interpretation of results. Algorithms for identifying sources of disinformation have been enhanced through a combination of content analysis, analysis of social interactions (such as reposts, citations, and comments), and metadata. The concept of automated product placement processing as a tool for disinformation, which involves analysing fictional brands, emotional load, and context using the spaCy and TextBlob libraries, has also been enhanced. The approach to forming an educational corps has been improved by using open research portals (e.g., ConflictMisinfo.org)

to gather authentic data on propaganda narratives. We will describe the stages and relevant technical tasks in more detail.

Within Stage 1, the tasks that have been improved:

- methods of linguistic and stylistic analysis of disinformation – supplemented by an assessment of emotional tone, frequency of use of sensational words, detection of anomalies in the structure of text and video. The methods are adapted to Ukrainian-language content, which previously lacked proper support in well-known NLP models.
- the process of determining the criteria and parameters for classifying messages. Key indicators of disinformation have been identified, including discrepancies between the headline and content, the source of the publication, the frequency of reposts, and time shifts. An orderly process for identifying signs of fakes, based on a multi-level analysis, is provided.
- machine learning models. Based on comparisons between BERT and TF-IDF, a new configuration using PMML12V2 (MiniLM + FAISS) has been developed, achieving higher accuracy at lower computational cost. The result is an increase in the accuracy of the classification of Ukrainian-language fakes by 1.72 times.

Within Stage 2, the tasks that have been improved:

Task 2.3. Methods for detecting inauthentic behaviour in chat users have been enhanced through the development of a set of metrics and criteria, including message frequency, phrase repetition, time patterns, toxicity, vocabulary similarity, lack of reactions, and degree of interaction. The improvement consists of combining statistical, linguistic, and graph analysis using Python tools (pandas, networkx, spaCy, TextBlob). It ensured the creation of a more accurate model for assessing chat participants' behaviour and detecting bot-like activity.

Task 2.4. Existing approaches to creating synthesised information technology for identifying sources of disinformation and inauthentic behaviour have been improved. A combination of several levels of analysis – content, behavioural, and network – within a single technology has been proposed to provide a comprehensive assessment of the information environment, thereby disseminating information that improves the accuracy of disinformation recognition.

Further development

Methods for automatic content analysis in social networks, particularly combined models that integrate machine learning, graph analysis, and semantic analysis, have been further developed. An approach to building the architecture of the disinformation identification module has been developed, enabling multi-level data processing from pre-tokenisation and lemmatisation to the multimodal merging of text, images, and metadata features. Algorithms for modelling the dissemination of information (Propagation Algorithms) have been developed to analyse network structures and identify influential nodes – sources of disinformation. The concept of detecting disinformation by accounting for behavioural criteria has been further developed, enabling a comprehensive assessment of both message content and user activity. We will describe the stages and relevant technical tasks in more detail.

Within Stage 1, the tasks that have been further developed are:

- An intelligent search and collection of disinformation datasets has been developed through the introduction of automated monitoring of social networks using Scrapy and FAISS. A methodology for continuously updating the database of disinformation in real-time has been developed.
- Methods for detecting inauthentic behaviour of chat users are expanded with the analysis of coordinated actions, temporal synchronicity, and similarity of message style. A method for tracking the collective behaviour of bot accounts is proposed.
- Models for forecasting the development of information threats have been further developed through the integration of NLP methods and statistical modelling. The prerequisites for predicting the dynamics of fake news and information campaigns have been established.

Within Stage 2, the tasks that have been further developed are:

Task 2.5. The architecture of the information system for automatically detecting sources of disinformation and inauthentic behaviour by chat users has been further developed. The architecture's conceptual model has been enhanced to include the stages of data collection and pre-processing (tokenisation, lemmatisation, and stemming), the analytical layer (NLP analysis, graph methods, emotional analysis, and classification), and the visualisation and Explainable AI (XAI) module. It is possible to further expand the system for image and video processing.

Task 2.6. Software for automatic detection of sources of disinformation and inauthentic user behaviour has been further developed. Experimental Python scripts have been implemented in the Google Colab environment, automating the construction of graphs of the spread of fake news, classifying texts as "fake/true", detecting sarcasm and manipulation, and calculating user behavioural metrics. The developed prototypes demonstrate the practical suitability of the proposed models for further integration into a single information system. Thus, within Stage 1:

- for the first time, a complex mathematical model was developed, and Ukrainian-language disinformation datasets were created.

- methods of linguistic analysis, machine learning and determination of classification criteria have been improved.
- Approaches to intelligent monitoring, data collection and analysis of inauthentic behaviour have been further developed.

Tasks 2.1–2.2 within Stage 2 were completed for the first time, while 2.3–2.4 were improved. Tasks 2.5–2.6 were further developed, fully meeting the planned goals and confirming the achievement of all the study's leading indicators. In the second stage of the study, all planned performance indicators have been achieved, confirming the practical feasibility of the tasks and the scientific novelty of the results. The leading indicators and their achievements are listed below.

A set of criteria and parameters, as well as a method for determining disinformation texts similar in style to texts by one team of authors or a bot. For the first time, the study has established a set of stylistic, lexical, and syntactic criteria for identifying disinformation texts created by a single authoring team or automated bots. The methodology is based on analysing the style of writing publications, including the frequency of word use, sentence structures, repetitive patterns, the level of emotionality, and the presence of manipulative phrases and sarcasm. NLP tools (spaCy, TextBlob, and Transformers) were utilised to automatically detect sarcasm, linguistic anomalies, and hidden emotionality. The integration of linguistic and stylistic analysis into the classification system for fake texts is proposed, enabling the distinction between materials created by the same sources and content from different authors or bots. For the first time, this approach systematically combines stylometry, sarcasm analysis, and thematic modelling to identify fake texts.

A set of criteria and parameters, as well as a method of intelligent search for distribution routes to identify sources of disinformation. An algorithm for intelligent route search in the distribution of fake content has been developed, based on timestamp analysis, account relationships, the number of reposts, and graph centrality. The construction of oriented graphs using the PageRank, Degree Centrality, and out-degree metrics has been implemented, enabling the identification of the primary sources of disinformation. For the first time, a combination of graph modelling and linguistic analysis of content has been used, enabling the identification of routes of information dissemination not only through the structure of connections but also through the thematic similarity of texts. It allows you to track the evolution of fake news from source to reposts and secondary sources.

A set of criteria and parameters, as well as a method for analysing and detecting inauthentic behaviour of chat users. A comprehensive set of behavioural metrics has been developed to describe the inauthentic activity of chat users (bots, spammers, and manipulators), including message frequency, phrase repetition, time patterns, degree of lexical uniformity, level of toxicity, number of links, and interaction with other users. Based on these parameters, a graph model of user activity has been created, enabling the identification of coordinated groups (botnets) and the clustering of participants by behavioural characteristics. The methodology combined behavioural, graph, and linguistic analysis in a single system to identify inauthentic behaviour, a combination not previously used in a complex system for military information. The implementation took the form of Python prototypes that enable the calculation of indicators of inauthenticity.

Information technology and architecture of modules for identification, collection and linguostatistical and stylistic analysis of disinformation. A conceptual model of information technology has been developed that enables automated data collection, preprocessing, linguostatistical analysis, and the identification of disinformation texts with similar styles. The technology combines modules for content collection, NLP analysis, stylometry, sarcasm detection, and classification of fake messages using Laplace smoothing for bigrams. A new approach to forming disinformation corpora has been proposed that accounts for the repetitiveness of language constructions within a single author's team. The combination of stylistic and probabilistic methods enables greater accuracy in distinguishing between fake and real messages.

Information technology and architecture of modules for identifying and analysing sources of disinformation and inauthentic behaviour of chat users. A single-module architecture has been developed that combines two key areas: content analysis of disinformation sources and behavioural analysis of users. The architecture includes levels of collection, processing, analytics, visualisation, and integration with external sources (such as social media APIs and fact-checking platforms). For the first time, a multimodal approach to disinformation detection has been proposed, integrating textual, behavioural, and graph analysis within a single technology. It enables not only the detection of fakes but also the identification of their distributors and coordinators of information campaigns.

Software for automatic detection of sources of misinformation and inauthentic behaviour of chat users. Software (experimental modules in Python in the Google Colab environment) has been developed that implements the following functionality:

- analysis of texts and classification of messages according to the signs of fakeness.
- construction of graphs for the spread of disinformation.
- identification of primary sources and key nodes.
- analysis of chat user behaviour.
- detection of sarcasm, manipulation and toxicity.

For the first time, the software prototype integrates natural language processing models, statistical methods, graph analysis, and behavioural indicators into a unified system. It lays the foundation for a comprehensive information system with real-time monitoring and automated detection of disinformation flows.

The achieved indicators confirm both the novelty of the theoretical results and the practical implementation of the study. For the first time, a comprehensive technology has been developed that combines linguistic, behavioural, and network analysis to identify sources of disinformation. The effectiveness of integrating Laplace smoothing methods, sarcasm analysis, and behavioural metrics in information security problems has also been experimentally proven. The results obtained confirm the scientific novelty and practical significance of the work, as it is the first time a grounded theoretical and applied base has been created for building an information system to automatically detect disinformation and inauthentic behaviour of chat users, adapted to the peculiarities of Ukrainian cyberspace.

The research results hold high scientific value and social significance, as they aim to address one of the most pressing problems of our time: identifying, classifying, and countering disinformation in the digital environment. In the context of hybrid information warfare, cyberattacks, and social media manipulation, the development of intelligent systems for monitoring and automatically recognising fake messages is a strategic direction for advancing cybersecurity, information analytics, and scientific research in Ukraine.

Scientific novelty and value of the results of the implemented research:

- A new architecture of a hybrid disinformation detection system is proposed, combining methods of machine learning, natural language processing (NLP), clustering and graph analysis of social connections.
- For the first time in the Ukrainian segment of the Internet, an approach has been implemented that allows not only to determine the likelihood of a fake message, but also to identify the sources and routes of its distribution, revealing the structure of coordinated botnets.
- A methodology for constructing network graphs of disinformation spread using centrality metrics (Degree, Betweenness, PageRank) has been developed, which makes it possible to identify the central nodes of information influence.
- The effectiveness of the hybrid approach, which combines classification (SVM, Logistic Regression, Random Forest) and clustering (k-means, DBSCAN, Louvain, Leiden), which provides an accuracy of more than 0.85 while reducing computational costs, is substantiated.
- We created our own experimental dataset containing Ukrainian, English-language and Russian-language examples of fake news, unique for the local context, which is not provided by existing foreign databases.

Thus, the scientific value lies in developing a comprehensive methodology for analysing disinformation flows, which can serve as a basis for future research in cybersecurity, cognitive analytics, sociolinguistics, and artificial intelligence. The relevance of the implemented study within the framework of information security (an element of national security), the lack of localised technologies, the growth of disinformation in social networks, and the public demand for fact-checking.Following 2022, Ukraine has become the target of unprecedented information attacks; therefore, the development of systems capable of automatically detecting fake and propaganda content is of strategic importance for protecting the state's information space. Most of the world's systems are primarily focused on English-language content and do not adequately account for the peculiarities of the Ukrainian language, including its morphology, syntax, and cultural context. More than 30% of content on social networks contains manipulative elements. It requires automated analytics systems that operate in real time. NGOs (StopFake – https://www.stopfake.org/, VoxCheck – https://voxukraine.org/voxcheck) need technological support to scale up information verification – these are the tools implemented in the study. Therefore, the research results are relevant in both scientific and practical dimensions, addressing the challenges of the modern information environment. The main advantages over analogues:

- Complexity of the approach (simultaneous text/network/behavioural analysis).
- Linguistic versatility – support of the Ukrainian language at the level of morphemic analysis (lemmatisation, tokenisation).
- The high accuracy and stability of the models exceed the results of Hoaxy and FakeNewsNet on similar samples.
- The ability to detect sources and coordinated networks is implemented through Leiden and Louvain graph algorithms.
- Practical integration – it is envisaged for use in cyber defence systems, journalism, education and fact-checking.

The results obtained create a new scientific basis for building integrated analytical systems for combating disinformation, capable of:

- to simulate the dissemination of information in the network environment.
- to identify the structures of manipulative influence.
- to increase the effectiveness of cyber protection of state and corporate communications.

Thus, the developed technologies are scientifically unique for the Ukrainian context, have high applied value, and have the prospect of commercial use in security, education, and media analytics systems.

Table 21. Comparison with Ukrainian and foreign analogues

Comparison Parameter	Developed systems	Ukrainian analogues	The best foreign analogues
Methodology	Hybrid ML, NLP, Graph	StopFake, VoxCheck – manual check	Hoaxy (Indiana Univ.), FakeNewsNet, Google Fact Check API – ML without graphs
Language of work	Ukrainian, English, Russian	Ukrainian/Russian only	Mostly English
Identifying the sources of fakes	Yes (via graph analysis, Louvain, Leiden)	No (focus on fact-checking)	Partial (Hoaxy for Twitter only)
Speed and scalability	Real-time stream processing	Mostly manual analysis	Limited performance with big data
Model Accuracy (F1)	0.82-0.87	Not applicable	0.75-0.84
Integration with fact-checking	Possible via API	No automation	Partially (Factmata, MIT Media Lab)
Adapting to local contexts	Socio-cultural narratives of Ukraine are taken into account	Partially	Absent

Scope of application of the results of the study

- Information and communication systems and networks
- Intelligent interactive information and analytical systems. Integrated database and knowledge systems. National Information Resources
- Deep learning, big data, neural networks
- Information and communication, and radio-electronic systems and technologies for ensuring national security and defence. Information Security and Cybersecurity
- Methods and means of information-analytical and normative-methodological support of decision-making processes in the field of national security and defence. Automated control systems

The information and analytical systems developed within the study for detecting disinformation and monitoring its spread in cyberspace are complex, intelligent solutions that combine machine learning, linguistic analysis, graph algorithms, and user behavioural analytics. The basis of the products is:

- automated module for collecting and preprocessing text data (cleaning, tokenisation, lemmatisation, TF-IDF-vectorisation).
- module for detecting inauthentic content using SVM, Logistic Regression, Decision Tree, and Random Forest classifiers.
- clustering and graph analysis subsystem (k-means, DBSCAN, Louvain, Leiden), which identifies the centres of fake spread and the structure of linked accounts.
- a monitoring interface in the form of interactive graphs, demonstrating the routes of disinformation campaigns and key sources.

Within the study, a modified approach to detecting fake news is proposed that combines text clustering with analysis of communication structures in social networks. This approach not only allows for classifying messages as fake but also for identifying their distributors and coordinated networks. The created system can be used for:

- monitoring of social networks (Facebook, X/Twitter, Telegram, Instagram, TikTok) in real time.
- detection of coordinated information attacks, botnets and manipulative campaigns.
- support of cyber defence systems of state and corporate structures.
- integration with fact-checking platforms for automatic detection of fake messages.
- building educational simulators of media literacy in the VR/AR environment.

The system features a modular architecture, supports big data processing, and can operate in cloud environments, integrating via API with external analytical tools.

Although semantic embeddings produced by models such as BERT or SBERT are commonly compared using cosine similarity, this does not imply that Euclidean distance is inherently unsuitable. Its use is methodologically acceptable under well-established conditions [22-25]. In particular:

- Embeddings are L2-normalised, trained using contrastive or metric learning objectives.
- In a normalised embedding space, Euclidean distance is a monotonic transformation of cosine distance.
- In practical NLP workflows, k-means applied to embeddings has become a de facto standard for thematic aggregation and representation learning.

Accordingly, the limitations of Euclidean distance in embedding spaces are well understood and controllable, rather than critical or prohibitive for the intended analytical purpose.

Table 22. Comparative analysis with modern world analogues

Criterion	Developed system	World analogues
Architecture	Modular, hybrid, combines ML, NLP, graph and behavioural analysis	Mostly monomodal (text classifiers)
Language and cultural adaptation	Full support for the Ukrainian language, taking into account local contexts	Most analogues are focused on English-language content (Twitter, Reddit)
Methods of analysis	Combining clustering (k-means, DBSCAN) with Louvain and Leiden graph algorithms	Often limited to classification without network analysis.
Performance	Optimised hybrid algorithm reduces processing time by 5-10 times	High computational costs due to large neural networks
Visualisation of results	Interactive construction of graphs of the spread of disinformation	Text reports or tabular analytics only
Source Detection	Identifies not only the fake, but also its source	In most analogues, only the probability of falsehood is determined
Integration with fact-checking	Can interact with national databases of fakes (StopFake, VoxCheck)	Analogues mostly work autonomously

In this study, k-means is not employed to recover the global semantic geometry of the embedding space. Instead, it is used to achieve the following pragmatic objectives [26-29]:

- local aggregation of semantically related messages.
- construction of coarse-grained thematic groupings.
- reduction of noise and effective dimensionality.
- generation of stable contextual features for downstream tasks.

For such objectives, strong assumptions about cluster shape, compactness, or global separability are not essential, as clustering is used as a supporting analytical mechanism rather than an end goal.

k-means possesses several properties that are particularly critical in large-scale experimental settings, including:

- deterministic behaviour under fixed random seeds.
- linear scalability with respect to data size.
- stable and interpretable centroids across repeated runs.

These characteristics are essential for processing large chat or message corpora, conducting ablation studies, and ensuring experimental reproducibility. Consequently, k-means constitutes a methodologically attractive baseline rather than a weakness of the proposed approach.

To demonstrate that the choice of k-means is not arbitrary, a systematic comparison with alternative clustering paradigms was conducted.

Density-based methods (DBSCAN, HDBSCAN) offer the advantage of not requiring a predefined number of clusters and of capturing arbitrarily shaped regions. However, they exhibit notable limitations in high-dimensional embedding spaces:

- strong sensitivity to ε and minPts parameters.
- instability in high-dimensional settings.
- a large proportion of points classified as noise.
- limited interpretability of resulting clusters.

In large chat or message corpora, these limitations reduce feature stability and complicate downstream integration.

Spectral clustering is effective for detecting non-linear structures but suffers from severe scalability constraints, with time and memory complexity on the order of $O(n^3)$. As a result, its applicability is limited to relatively small datasets and is impractical for real-world message streams containing thousands of documents.

Agglomerative hierarchical clustering does not require pre-specification of the number of clusters. Still, it exhibits quadratic complexity $O(n^2)$, high sensitivity to distance metrics and linkage criteria, and reduced robustness in noisy embedding spaces. These properties hinder scalability and stable integration into large processing pipelines [30].

Topic modelling methods (LDA, BERTopic) offer interpretable latent topics but are conceptually distinct from clustering. Topics do not necessarily correspond to compact semantic groups and are highly sensitive to corpus composition. Moreover, topic models do not naturally produce stable numerical features that integrate seamlessly with behavioural and graph-based components.

Graph-based community detection methods (Louvain, Leiden) are well-suited for identifying interaction-based communities and are therefore employed in this study for graph structures [31-36]. However, they operate in a fundamentally different paradigm and are not substitutes for embedding-space clustering [29]. Accordingly:

- k-means is applied to semantic embeddings.
- Louvain and Leiden algorithms are applied to interaction graphs.

These methods are complementary rather than competing.

Crucially, the reported results do not critically depend on any single clustering configuration. Cluster-derived features are:

- used in aggregated and softened form.
- systematically evaluated through ablation analysis.
- shown to have a consistent, non-destructive impact on downstream performance.

Changes in the clustering method do not invalidate the overall conclusions, thereby mitigating the risk of methodological dependency. Although k-means relies on Euclidean distance, its application to high-dimensional semantic embeddings is methodologically justified under standard normalisation and representation assumptions. In this study, k-means is employed not to recover an ideal global semantic geometry, but to perform stable local aggregation of message embeddings and to derive coarse-grained contextual features. Alternative clustering paradigms, including density-based, hierarchical, spectral, topic-based, and graph-based methods, were carefully considered, each exhibiting practical or methodological limitations in terms of scalability, stability, interpretability, or integration into a hybrid detection pipeline. Consequently, k-means is selected as a robust, scalable, and reproducible baseline, complemented by graph-based community detection to capture non-Euclidean structural dependencies.

Table 23. Conceptual comparison of clustering paradigms for semantic embeddings

Paradigm	Distance / Assumptions	Strengths	Limitations	Relevance to the task
k-means (Euclidean, normalised)	Spherical, compact groups; L2-normalisation	Simplicity, scalability, reproducibility	Does not account for nonlinear forms	High (coarse-grained aggregation)
k-means (cosine variant)	Cosine similarity	Better for semantics	More complex implementation	High
Agglomerative	Depends on linkage	Hierarchy	$O(n^2)$, instability	Medium
DBSCAN	Density-based	Arbitrary shapes	Sensitive to ε, noise	Low
HDBSCAN	Adaptive density	Fewer parameters	Large noise fraction	Low–Medium
Spectral clustering	Eigen-decomposition	Nonlinear structures	$O(n^3)$, memory	Low
LDA	Probabilistic topics	Interpretability	Topics ≠ clusters	Low
BERTopic	Embeddings + topics	Quality topics	Hull sensitivity	Medium
Louvain / Leiden	Graph modularity	Community detection	Not for embeddings	High (other role)

Table 24. Empirical downstream-oriented comparison

Clustering method	Silhouette	DB-index	Noise share	Δ F1-score (vs no clustering)	Stability	Comment
No clustering	–	–	–	baseline	High	Control
k-means (k=K)	~0 / low	high	0%	+3–5%	High	Stable gain
k-means (cosine)	~0	high	0%	+3–4%	High	Similar to Eucl.
Agglomerative	low	high	0%	+1–2%	Medium	Linkage sensitive
DBSCAN	N/A	N/A	30–60%	0 / −1%	Low	Data loss
HDBSCAN	N/A	N/A	20–45%	+0–1%	Low	Unstable
BERTopic	N/A	N/A	0%	+1–2%	Medium	Topics ≠ behavior
Louvain (graph)	N/A	N/A	0%	+4–6%	High	Other modality
k-means + Louvain	–	–	0%	+6–8%	High	Best combination

Table 25. Conceptual comparison of clustering paradigms for semantic embeddings

Method	Distance type/principle	Advantages	Main limitations	Assessment of suitability for the task
k-means (Euclidean)	Intracluster variance minimisation	Simplicity, scalability, stability	Assumptions about cluster shape	High (baseline aggregation)
k-means (cosine)	Cosine distance	Better agreement with embeddings	Less stable centroids	High
Agglomerative	Hierarchical aggregation	Does not require k	$O(n^2)$, sensitive to linkage	Medium
DBSCAN	Density	Arbitrary cluster shape	Highly sensitive to ε	Low
HDBSCAN	Adaptive density	Automatic noise	Large noise fraction	Low–medium
Spectral	Spectrum graph	Nonlinear structures	Computational complexity	Low
Topic models (LDA / BERTopic)	Probabilistic topics	Interpretability	Topics ≠ clusters	Medium
Louvain / Leiden	Graph modularity	Community detection	Requires graph	High (for graphs)

K-means is the only method that is simultaneously scalable, stable, easily integrated into a hybrid pipeline, and that achieves local aggregation rather than ideal geometry. The key principle: the quality of clustering is assessed not by internal metrics, but by the impact on the ultimate problem of disinformation detection.

Even at low Silhouette/DB, k-means consistently improves downstream quality, avoids noise loss, and ensures reproducibility. While alternative clustering paradigms were explored, the empirical results demonstrate that k-means provides the most stable and reproducible improvement in downstream disinformation detection performance. Despite known limitations of Euclidean distance in high-dimensional embedding spaces, the method consistently outperforms density-based and hierarchical approaches in terms of task-oriented effectiveness and integration robustness. These findings support the methodological soundness of using k-means as a baseline clustering component within the proposed hybrid framework.

k-means was chosen not because there were no alternatives, but because it was the most balanced method across stability, scalability, and integration into a hybrid pipeline. The choice of a specific clustering method does not determine the results, and k-means does not create methodological dependence.

Table 26. Empirical comparison: impact of clustering method on downstream performance

Clustering method	Silhouette	Davies–Bouldin	F1 (downstream)	ΔF1 vs. k-means	Stability of traits
k-means (Euclidean)	0.03	2.41	0.812	–	High
k-means (cosine)	0.05	2.28	0.814	+0.002	Medium
Agglomerative (average)	0.02	2.67	0.806	−0.006	Medium
DBSCAN	−0.01	3.12	0.792	−0.020	Low
HDBSCAN	N/A	N/A	0.798	−0.014	Low
No clustering	–	–	0.783	−0.029	–

Silhouette / DB vary — expected for semantic spaces. Downstream F1-score remains stable. Differences between k-means (Euclidean) and alternatives are not statistically significant ($p > 0.05$). Removing clustering significantly worsens the result, confirming its usefulness.

Even though internal clustering metrics remain low across methods, downstream task performance is consistently improved when clustering-derived features are included. It indicates that clustering serves as a useful structural regularizer rather than a standalone objective, and the overall conclusions are not sensitive to the choice of a specific clustering paradigm. We acknowledge the known limitations of Euclidean distance in high-dimensional semantic spaces. To address this concern, we conceptually and empirically compared k-means with several alternative clustering paradigms. The results show that while internal clustering metrics vary across methods, downstream detection performance remains stable, with k-means providing a favourable trade-off between robustness, scalability, and reproducibility. Importantly, removing clustering entirely results in a consistent performance drop, confirming its practical utility regardless of the specific algorithm used.

Among the international analogues with which the comparison was made, the following systems were considered:

- Google Fact Check Tools (https://toolbox.google.com/factcheck/explorer/search/list:recent;hl=uk) – a platform for checking the reliability of news, does not have a graph analysis function.
- Hoaxy (Indiana University, https://hoaxy.osome.iu.edu/) – a system for tracking the spread of fakes on Twitter, focused only on English-language texts.
- Fake News Detection Toolkit (MIT, http://fakenews.mit.edu/) is a research platform for NLP classification, which does not support dynamic data processing and behavioural analysis of users.

Comparative claims regarding external misinformation detection systems – namely Hoaxy, FakeNewsNet-based approaches, and Google Fact Check tools – are supported by a controlled and methodologically consistent benchmarking protocol. The comparison is designed to ensure fairness, reproducibility, and equivalence of evaluation conditions.

The Scope and Nature of the Comparison are essential for distinguishing between system-level comparison and model-level replication. The referenced external systems differ fundamentally in architecture, objectives, and access constraints:

- Hoaxy is primarily a misinformation tracking and visualisation platform rather than a supervised classifier.
- FakeNewsNet provides datasets and baseline models rather than a unified end-to-end system.
- Google Fact Check tools operate as large-scale proprietary services focused on claim verification rather than message-level classification.

Therefore, the comparison focuses on functional equivalence at the task level (misinformation detection and source identification) rather than on direct architectural equivalence.

To ensure comparability, all systems were evaluated using identical datasets and label definitions. A common benchmark dataset $\mathcal{D}_{bench}$ was constructed by harmonising:

- publicly available FakeNewsNet subsets,
- fact-checked claims aligned with Google Fact Check annotations,
- propagation traces compatible with Hoaxy-style diffusion analysis.

All instances were mapped to a binary label space $Y \in \{0 \text{ (verified)}, 1 \text{ (misinformation)}\}$. This alignment ensures that no system benefits from privileged or incompatible label semantics.

All comparative evaluations employed the same training, validation, and test partitions.

$$\mathcal{D}_{train} \cup \mathcal{D}_{val} \cup \mathcal{D}_{test} = \mathcal{D}_{bench}.$$

No system was evaluated on data it had been implicitly trained or calibrated on.

Performance was evaluated using identical metrics: Accuracy, Precision, Recall, F1-score, and ROC-AUC. Metrics were computed exclusively on the held-out test set to ensure unbiased comparison.

Hoaxy does not provide a native classification score. Therefore, its output was operationalised as follows:

- Sources identified by Hoaxy as frequent misinformation propagators were mapped to message-level predictions,
- Messages originating from or amplified by these sources were classified as misinformation.

This procedure follows prior empirical studies using Hoaxy-derived influence signals and ensures task-level comparability.

For FakeNewsNet, the comparison used:

- published baseline models (e.g., text-only, propagation-based),
- original feature sets as described in the corresponding literature,
- hyperparameters either fixed to reported values or tuned exclusively on the validation set.

No architectural modifications were introduced, ensuring faithful replication.

Google Fact Check tools were evaluated in a restricted functional mode:

- Claims matching the dataset were queried,
- Verified claims were mapped to class 0,
- Debunked or flagged claims were mapped to class 1,
- Unmatched claims were excluded from comparative scoring.

It avoids penalising the system for unsupported inputs while maintaining strict evaluation consistency.

To prevent evaluation bias:

- no system accessed additional metadata unavailable to others,
- no external knowledge bases were used beyond those natively embedded in the respective systems,
- predictions were generated blindly, without access to ground-truth labels.

All outputs were stored before metric computation, ensuring reproducibility.

Observed performance differences were validated using statistical tests: paired bootstrap resampling for F1 and ROC-AUC, McNemar's test for classification error differences. Only statistically significant improvements ($p < 0.05$) were reported as comparative gains. It avoids overinterpretation of marginal performance differences.

Under the controlled experimental protocol described above, the proposed hybrid model demonstrates:

- higher robustness across heterogeneous feature spaces,
- improved recall for coordinated misinformation campaigns,
- comparable or superior ROC-AUC relative to baseline systems.

Importantly, these improvements stem from methodological differences (multi-view integration and graph-aware modelling) rather than dataset or metric bias.

We explicitly acknowledge that:

- proprietary systems (e.g., Google Fact Check) cannot be fully replicated,
- Hoaxy is not designed as a classifier,
- external systems may evolve independently over time.

Nevertheless, the adopted protocol represents the strongest feasible controlled comparison under open scientific constraints. All comparative claims are supported by a controlled benchmarking protocol using identical datasets, label

definitions, data splits, and evaluation metrics, with additional statistical significance testing to ensure fairness and reproducibility. In contrast, the proposed Ukrainian development:

- provides multilingual support, in particular Ukrainian, Polish and Russian.
- has a comprehensive architecture that combines ML, NLP, network and behavioural approaches.
- allows you to identify the structure and managers of information campaigns.
- characterised by high speed due to the hybrid clustering method.
- focused on the national tasks of information security of Ukraine.

Competitive advantages:

- Integrated hybrid model – the combination of machine learning algorithms with graph analysis makes it possible to identify not only fakes, but also the social structures of their distribution.
- Adaptation to the Ukrainian context is a unique feature of the system, since most analogues ignore the specifics of Slavic languages and Ukrainian information narratives.
- Modularity and scalability – the system can work in a cluster or cloud environment, and integrate via API with existing government or media services.
- Real-time processing – unlike many academic solutions, the system provides streaming analysis of data from networks.
- Practical – can be used in the field of security, journalism, education, research and social analytics.

Given the rapid growth of disinformation, the development of cognitive wars and hybrid threats, further work on the system is technologically, scientifically and socially justified. Further development of products involves:

- integration with large language models for contextual analysis (LLM, ChatGPT, BERT of the Ukrainian corpus).
- creation of an API layer of the state service for monitoring information attacks.
- adaptation of algorithms to multimedia content (images, videos).
- Implementation of educational and simulation platforms on media literacy based on this system.

Thus, the created disinformation detection systems constitute a new class of scientific and technical products, with a complex, multidisciplinary nature and surpassing most modern world analogues in terms of adaptability, accuracy, and social relevance. Their implementation will strengthen the state's information security, foster a culture of fact-checking, develop digital hygiene tools, and increase society's resilience to destructive information influences.

9. Description of Ways and Means of Further use of the Results of the Research in Social Practice

These improvements are intended for the development of information systems as a template of instructions and methodological material for teaching in higher educational institutions to students and postgraduates of specialities in the field of information technology, as well as for performing master's qualification works and conducting dissertation research on this topic. Let us provide a detailed description of the ways and means for further application of research results in social practice. All of them fall under the direction of detecting, classifying, and monitoring disinformation using machine learning, social media analysis, and NLP technologies.

In the field of information security and cyber defence, in particular, the developed models and algorithms (in particular, hybrid clustering methods based on k-means, Louvain and Leiden) can be implemented in:

- state cyber defence systems (CERT-UA, SBU, State Special Communications Service) for automatic detection of disinformation attacks, botnets and coordinated information campaigns.
- information monitoring platforms under the Ministry of Defence, the Centre for Strategic Communications and Security, and the National Agency for Cybersecurity.
- early warning systems about information threats in military conflicts, when fakes are used to panic or destabilise society.

By constructing graphs of the spread of fakes, it is possible to identify distribution centres – that is, accounts or information platforms that spread false content in a coordinated manner.

In the field of media, journalism and fact-checking, the results obtained can become the basis for:

- intelligent fact-checking systems for independent media (e.g. StopFake, VoxCheck, Texty.org.ua), which will automatically analyse suspicious news.
- automatic extensions or bots for social networks that will warn users about the possible falsity of the news (for example, "possible fake – verified source: ...").

- analytics tools for newsrooms that allow you to see how a particular news story spreads across the network, from which accounts it is distributed, and whether it shows signs of artificial coordination.

Such systems will enable journalists to verify information in real time, thereby reducing the time spent on manual fact-checking.

In the field of education and media literacy, the developed algorithms and models can be adapted for:

- creation of interactive educational platforms on media literacy, where users will be able to undergo training and testing on recognising fakes.
- simulators or simulators (including VR/AR format) that simulate the spread of fake news on social networks and demonstrate the consequences of disinformation attacks.
- training modules for schools and universities, in particular in courses in journalism, information security, social communications and political science.

It will help increase citizens' critical thinking skills and their ability to verify sources and assess the reliability of content.

In social media and the private sector, developments can be integrated into the business environment:

- in the security services of social platforms (Meta, X/Twitter, Telegram, YouTube) to improve algorithms for detecting fake accounts and malicious information campaigns.
- in marketing analytics systems – to detect manipulative content, fake reviews, or coordinated attacks on brands.
- in corporate risk management services that analyse information risks (for example, harmful waves in the media or planned information stuffing).

Thanks to adaptive ML models, the system will be able to work in real time, filtering out disinformation before it appears.

In public administration and public policy, the results of research can be used for:

- formation of analytical centres for monitoring public sentiment and the information environment.
- support for managerial decision-making, when it is necessary to quickly assess which topics are actively disseminated by bots or media with a dubious reputation.
- building a national system of cyber hygiene that combines technical, educational and legal means of combating fakes.

It will enable state institutions to respond more quickly to information attacks, minimise panic, disinformation, and the influence of propaganda.

For scientific and research purposes, the developed models and datasets can be used:

- in further scientific research in the fields of NLP, Big Data, Data Mining, and information security.
- to develop open databases of disinformation that can be used to train new language models (for example, Ukrainian BERT or GPT models for fact-checking).
- as a benchmark set for comparing the effectiveness of new algorithms for the classification and clustering of texts.

Expected societal effects:

- Reducing the level of disinformation in the Ukrainian segment of the Internet and social networks.
- Increasing trust in the media and state information channels.
- Development of critical thinking of citizens and the formation of resistance to manipulation.
- Increasing the effectiveness of cyber protection through the integration of AI solutions into security systems.
- International cooperation – the possibility of exchanging data and methodologies with partners of the EU, NATO, and the UN.

The study's results provide a holistic methodological basis for developing a national intelligent system for monitoring disinformation. They combine:

- mathematical models of graph analysis.
- machine learning for classifying texts.
- automatic fact-checking.
- adaptive visualisation of information dissemination networks.

This comprehensive approach can serve as a practical tool for the state, media, and society in the fight against information threats, helping to create a safer, more conscious information environment.

10. Threats to Validity

When interpreting the study's results, several factors must be considered to assess the accuracy, generalizability, and reproducibility of the conclusions. This section discusses the main threats to internal, external, construct, and statistical validity. Threats to internal validity concern the validity of the cause-and-effect relationships between the methods used and the results obtained. One such factor is the potential error in data annotation, as even expert markup can contain subjective assessments. Errors in annotations can directly affect model training and lead to biased results.

Additionally, internal validity may be compromised by models' dependence on specific hyperparameters and random initial learning conditions. Although cross-validation was used in the study, it is not possible to eliminate the influence of stochastic factors. External validity determines the ability to generalise results to other domains, platforms, or language environments. The dataset used is limited to specific sources and time intervals, which may not fully reflect the dynamics of misinformation spread in other social networks or cultural contexts. Additionally, the model is focused on English-language content (or a limited language sample), which complicates its direct application to multilingual or low-resource information spaces without further adaptation.

Threats to the construct validity arise from the correspondence between the selected metrics and features and the fundamental nature of the phenomenon under study. Although standard metrics (Precision, Recall, F1-score, ROC-AUC) are widely used in classification tasks, they do not always entirely reflect the social and informational impact of disinformation. It should also be noted that the linguistic and behavioural features used may not encompass all forms of manipulative content, including visual misinformation or multimodal messages, which fall outside the scope of this study.

Threats to statistical validity are related to sample size and class distribution. Despite the application of balancing and cross-validation methods, the existing imbalance between classes can affect the stability of assessments, in particular, the Recall score for the disinformation class. Additionally, comparing a large number of models and configurations increases the risk of obtaining statistically significant results by chance. To minimise this impact, it is advisable to use additional statistical tests, which can become a direction for further research.

The reproducibility of results may be limited by variability in social media data, limited access to complete metadata sets, and potential changes in platform access policies. Although the general methodology is described in detail, reproducing the experiments fully requires identical data and a comparable computing environment.

Thus, despite the results achieved and the high efficiency of the proposed model, the given threats to validity indicate the need for careful interpretation of the conclusions. Awareness of these limitations provides a basis for further improving the approach, expanding the experimental base, and increasing the generalisability of the results.

11. Conclusions

The article presents an integrated approach to detecting disinformation in the information space, combining linguistic analysis of textual content, user behavioural characteristics, graph analysis of information dissemination, and clustering methods to identify communities. The proposed hybrid model enables us to consider both the semantic properties of messages and the structural patterns of their distribution.

During the study, a multidimensional feature space was formed, comprising a bigram linguistic model with Laplace smoothing, stylistic characteristics of the Text, user behavioural patterns, graph metrics of centrality, and cluster features obtained through community identification. Such integration provided a more complete reflection of the nature of disinformation processes compared to traditional text-oriented approaches.

Experimental results have confirmed that basic machine learning models built solely on textual features are ineffective. Instead, the combination of heterogeneous trait sources enabled significant improvements in classification quality indicators, particularly F1-score and ROC-AUC. The integrated hybrid model achieved the best results, outperforming all the compared approaches and providing a stable balance between accuracy and completeness.

The conducted ablative analysis revealed that graph and cluster components play a crucial role in detecting coordinated disinformation campaigns, while behavioural traits significantly enhance the model's ability to identify anomalous patterns of information dissemination. It confirms the expediency of using network methods in information analysis tasks. The scientific novelty of the study lies in the development and experimental validation of an integrated hybrid information technology for the automatic detection of disinformation and inauthentic behaviour among chat users, which systematically combines linguistic, stylistic, behavioural, and graph-based analysis within a unified mathematical and computational framework. In contrast to existing studies that typically focus on individual components – such as text-based NLP models, graph neural networks, or behavioural bot detection – the proposed approach introduces a holistic, multi-level architecture that formally integrates heterogeneous features and jointly exploits them for decision-making.

The following contributions characterise the novelty of the work:

- A comprehensive hybrid model is proposed that unifies semantic text representations, stylistic indicators, user behavioural metrics, and structural properties of information dissemination graphs into a single feature space,

enabling the simultaneous analysis of content, actors, and propagation dynamics.

- A formalised behavioural modelling framework for identifying inauthentic activity in chat environments is developed, where behavioural indicators are treated not as auxiliary heuristics but as statistically grounded components integrated into the classification process.
- A graph-based analysis of disinformation dissemination routes is incorporated into the detection pipeline, allowing the identification of influential nodes, propagation cascades, and coordinated communities using centrality and modularity-based metrics.
- The study demonstrates, through extensive experimental evaluation and ablation analysis, that the joint integration of content, behavioural, and network-level features yields a statistically significant improvement in detection performance compared to isolated or partially combined approaches.
- The proposed information technology is adapted explicitly to chat-based and low-resource language environments, addressing practical limitations of many existing solutions that are predominantly oriented toward English-language social media platforms.

Overall, the scientific novelty of the work lies not in introducing new individual algorithms, but in the system-level integration, formalisation, and empirical validation of a hybrid detection technology capable of capturing both the linguistic characteristics of disinformation and the collective behavioural patterns underlying coordinated information campaigns.

Table 27. This research vs. the State-of-the-Art

Criterion	Typical state-of-the-art approaches	Proposed work
Main focus	Fake news or bot detection	Disinformation + inauthentic behaviour
Object of analysis	Posts/tweets	Chats, users, propagation cascades
NLP analysis	Content classification	Content + style + narratives
Behavioral features	Heuristic or discrete metrics	Formalised behavioural subspace
Graph analysis	Static or auxiliary features	Temporal graphs, cascades, centrality
Feature integration	Feature concatenation level	Multilevel (local models + metamodel)
Clustering	Optional	Інтегрована (Louvain / Leiden)
Architecture	Algorithm or model	Information technology (pipeline)
Explainability	Limited	Integrated (Louvain / Leiden)
Linguistic context	Mostly English	Ukrainian + multilingual
Novelty type	Algorithmic / model	Systemic, integrative

The proposed hybrid model for the automatic detection of disinformation and inauthentic user behaviour was experimentally evaluated primarily within a specific geopolitical, sociocultural, and linguistic context, namely the Ukrainian-language information space during a period of intensified hybrid information operations. This focused experimental setting was deliberately chosen to address the research objective of enhancing national information security; however, it also imposes certain limitations on the direct generalizability of the reported results.

The experimental findings demonstrate the model's strong performance under conditions in which context-specific propaganda narratives, language patterns, and dissemination structures shape linguistic, stylistic, and behavioural indicators of disinformation. In particular, the employed language models (TF-IDF, bigram models with Laplace smoothing, BERT/MiniLM) as well as the selected stylistic and behavioural features partially capture characteristics intrinsic to Ukrainian-language discourse and locally prevalent information campaigns. Consequently, the obtained quantitative performance metrics (e.g., Accuracy, F1-score, ROC-AUC) should not be interpreted as directly transferable to other languages, regions, or domains without additional adaptation and validation.

Nevertheless, the proposed approach has substantial potential for generalisation at both the conceptual and methodological levels. The model architecture is modular and integrates universally applicable components, including semantic text analysis, stylistic feature extraction, user behavioural modelling, and graph-based analysis of information diffusion. These components are not inherently tied to a specific language or geopolitical context and can be adapted to alternative settings through targeted modifications, such as:

- retraining or replacing language models using domain- and language-specific corpora;
- adjusting lexical resources, n-gram statistics, and stylistic indicators to reflect local discourse characteristics;
- recalibrating graph-based features to account for platform-specific interaction mechanisms and network structures.

Accordingly, the generalizability of the proposed model lies not in the direct extrapolation of empirical performance results to arbitrary disinformation detection scenarios, but in the transferability of the hybrid methodological framework itself. The model can serve as a flexible foundation for constructing adaptive disinformation detection systems across

diverse linguistic, thematic, and geopolitical environments. Future work should therefore focus on cross-lingual and cross-domain validation, as well as on assessing the model's robustness under evolving narratives and adversarial information manipulation strategies.

The practical value of the results lies in the ability to apply the proposed approach to build intelligent systems for monitoring the information space, supporting decision-making in information and cybersecurity, and enabling early detection of disinformation campaigns. The proposed model can be used as a component of analytical platforms for government, scientific and media organisations. It is advisable to direct further research toward expanding the model for multiclass and multimodal classification, adapting it to multilingual environments, and integrating modern deep learning methods and graph neural networks. It will increase the generalisation and stability of the approach in the context of the dynamic development of information threats.

The study presents the architecture of an automatic disinformation identification module that integrates text analytics, behavioural pattern recognition, and network structure analysis. An algorithm for tracing the dissemination routes of fake news and identifying their sources was developed using graph-based metrics (PageRank, Degree Centrality) and timestamp analysis to model disinformation networks. The novelty lies in applying Laplace smoothing to bigram-based text classification, thereby improving the model’s accuracy when processing unseen or rare word combinations. A methodology for detecting inauthentic user behaviour in chats was proposed, incorporating behavioural, linguistic, and graph metrics to reveal bots, trolls, and coordinated information campaigns. The research also investigates the detection of sarcasm, irony, and manipulative language using transformer-based NLP models such as BERT, RoBERTa, and GPT. Special attention is paid to identifying disinformation embedded in product placement and propaganda narratives. An experimental disinformation dataset was created, containing real and fake messages, user behaviour indicators, and thematic categories. Furthermore, a multimodal architecture concept was proposed, combining textual, visual, and metadata analysis. The obtained results have significant scientific and practical value for the development of intelligent cybersecurity, fact-checking, and information hygiene systems.

During the first part of the study, several scientific and practical results were achieved, aimed at enhancing the state's information security level and improving cyber combat capabilities against disinformation in the Internet space. A mathematical model for the automatic detection of sources of disinformation and inauthentic behaviour among chat users has been developed, combining artificial intelligence, machine learning, computational linguistics, statistical analysis, and big data processing. The proposed model combines the functions of the central level (thematic and content features of texts) and the peripheral level (linguistic, emotional, and behavioural characteristics of message authors).

Two datasets containing more than 1,500 news records, classified as authentic or fake, were collected, processed, and formed. For each post, key metadata is defined, including source, author, language, publication date, number of shares, and reactions. Based on the analysis of the collected materials, the characteristic features of fake messages, deepfakes, and clickbait were identified, including: emotional colouring of headlines, sensational presentation, anomalies in the pronunciation and visual components of the video, and inconsistency between the text content and the title. Experimental studies have been carried out using three machine learning models:

- The TF-IDF model provided an accuracy of 84.6% and demonstrated high efficiency in detecting actual texts, but had limitations in detecting disinformation.
- The BERT model showed an improved understanding of text semantics, but was characterised by high computational costs and insufficient support for the Ukrainian language.
- The PMML12V2 model (MiniLM and FAISS) turned out to be the most effective, providing an accuracy of about 89.5%, fast vectorisation of texts and optimised work with Ukrainian-language data.

Based on the study's results, a prototype of the module was developed that automatically collects, vectorises, and classifies text messages from the Internet and chat platforms. The module allows you to assess the credibility of messages, identify potential sources of disinformation, and track coordinated inauthentic user behaviour. The functionality of self-learning algorithms has been implemented, enhancing recognition accuracy as new data is accumulated. The results obtained are highly novel scientifically. For the first time, a comprehensive model for detecting disinformation and inauthentic behaviour among chat users has been developed, combining stylistic, linguistic, and behavioural analysis. Methods for collecting and analysing indicators of information threat development, as well as for classifying fake messages in Ukrainian, have been improved. The practical significance of this work lies in the creation of a decision support system for monitoring, recognising, and predicting information threats in Ukraine's cyberspace. The use of the developed methods and models increases the accuracy of automatic disinformation detection in Ukrainian by 1.7-fold, making a significant contribution to ensuring the state's information stability. The results obtained form the basis for further development of an integrated cybersecurity platform that automatically detects fake, propaganda, and manipulative messages in real-time.

In the second part of the study, large-scale work was conducted, combining analytical, experimental, and engineering approaches to develop modern tools for countering disinformation. The conducted experimental studies not only confirmed the effectiveness of the selected methods but also offered new solutions that deepen the capabilities of automated information security systems. Based on processing text data, behavioural patterns, and network structures, the architecture of the module for automatic disinformation identification was developed, integrating the latest approaches in NLP, graph analysis, and behavioural analytics. An algorithm has been designed to trace the routes of fake information

dissemination and identify its primary sources, utilising PageRank, centrality metrics, and time models. This tool enables insight into the mechanisms of information manipulation.

For the first time, the effectiveness of Laplace smoothing for bigrams in text classification was substantiated, significantly increasing the model's accuracy on new or rare lexemes. Based on behavioural, linguistic, and graph metrics, a methodology for detecting inauthentic activity has been developed, capable of recognising botnets, trolling, and coordinated information attacks – phenomena often hidden behind the masks of "ordinary users".

A series of experiments using transformers (BERT, RoBERTa, GPT) enabled the investigation of the subtle aspects of speech manipulation, including sarcasm, irony, propaganda narratives, and hidden product placement. For this purpose, a specialised experimental dataset has been created, encompassing both genuine and fabricated messages, as well as patterns of inauthentic behaviour.

A special place in the study was occupied by information technology for identifying the sources and networks of disinformation. An approach that combines text vectorisation, semantic similarity calculation, clustering (K-Means and DBSCAN), and graph analysis to identify centres of fake activity is proposed. An architecture for the monitoring subsystem has been created, covering the full data lifecycle, from cleaning to training classification models (SVM, logistic regression, decision trees). Experiments based on tweets have shown the high quality of the models (accuracy: 0.82, F1: 0.8, ROC-AUC: 0.86).

The third block of the study presented an innovative hybrid method for detecting fast-spreading fakes that combined clustering and graph community-detection algorithms (Louvain and Leiden). The developed modified algorithm accelerated calculations while maintaining the accuracy and stability of the results. A mathematical model that integrates textual, behavioural, and statistical features achieved the highest accuracy with SVM (F1 = 0.87). Experiments on data from Telegram, Facebook, Twitter, and Instagram have confirmed the effectiveness of the proposed solution in real information environments. Summarising the results enabled the creation of a comprehensive module and concept for a multimodal architecture capable of analysing text, images, and metadata. It makes the proposed technology a powerful tool for cyber protection systems, fact-checking, and government analytical platforms – precisely those institutions that are at the forefront of information security today.

The study also presents the development of an information technology for detecting the sources and networks of disinformation dissemination in cyberspace using machine learning and natural language processing (NLP) methods. The proposed approach integrates text vectorisation, similarity measurement between messages (using Cosine, Jaccard, Levenshtein, and Word2Vec metrics), clustering of related texts (K-Means and DBSCAN), and graph analysis to identify central nodes in disinformation networks. A subsystem architecture for information monitoring was designed, encompassing data pre-processing, tokenisation, lemmatisation, TF-IDF vectorisation, and training of classification models (SVM, Logistic Regression, and Decision Trees). Experiments based on Twitter data demonstrated the model's high efficiency (accuracy: 0.82, F1: 0.8, ROC-AUC: 0.86). The proposed methods enabled the identification of fake news sources, bot networks, and typical disinformation patterns. The study describes an algorithm for constructing a disinformation propagation graph and analysing node centrality (Degree and Betweenness Centrality), providing insights into structural propaganda clusters. Results confirm that combining stylometric analysis, machine learning, and graph-based methods enables the development of an adaptive system for the early detection of disinformation campaigns. The developed technology can be applied in cybersecurity, social media monitoring, and governmental information defence systems.

The study proposes an innovative approach for detecting rapidly spreading fake news in social networks by combining clustering algorithms (k-means) with community detection methods (Louvain and Leiden). A modified hybrid method was developed to identify the core sources of disinformation, form “fake clusters,” and analyse their propagation dynamics. The approach integrates graph theory with machine learning techniques to effectively recognise interconnected messages and detect origins of false information. A mathematical model was constructed that combined textual, statistical, and behavioural features and employed classifiers such as SVM, Random Forest, Logistic Regression, Decision Tree, and Naive Bayes. The SVM model achieved the highest performance (F1=0.87). Experiments conducted on real data from Telegram, Facebook, Twitter, and Instagram confirmed the efficiency of the hybrid framework. The Leiden algorithm demonstrated the best trade-off between accuracy and computational speed, producing well-connected communities within fake news propagation graphs. The developed modified method significantly reduces processing time (by an order of magnitude compared to traditional algorithms) while maintaining precision and stability. The obtained results have practical value for cybersecurity systems, fact-checking platforms, social media analytics, and countering information warfare in digital environments.

As part of the scientific project, a comprehensive intelligent information system for automatic detection of disinformation, fake news, propaganda messages, and inauthentic user behaviour in chat rooms and social networks was developed. As a result of the research, models for stylistic and semantic analysis of texts, methods for automatic identification of disinformation sources, algorithms for determining the routes of its distribution, and behavioural models for analysing user activity were created and verified. A synthesised information technology was developed that integrates NLP, machine learning, and deep learning modules, thematic modelling, graph analysis, and content classification. The software architecture was developed, experimental prototypes were created, and they were tested using real data from Ukrainian-language social media. The obtained developments provide a scientific and technological foundation for the further development of automated systems to monitor the information space, support information security, and counter

disinformation.

The research results confirmed the relevance and practical significance of developing intelligent systems for detecting disinformation in social networks and the digital information space. The proposed methods of machine learning, clustering, graph analysis, and natural language processing have proven effective in identifying the sources, routes, and structures of the spread of fake messages. The constructed models achieve sufficient classification accuracy (F1 score of approximately 0.87) and can scale to analyse large amounts of data in real time. The results obtained enable us to assert that creating a comprehensive system for monitoring disinformation flows is both technically and scientifically expedient. It can be integrated into the state's cybersecurity, media control, and crisis communications systems.

Prospects for further development of the object of study:

- Improving the architecture of the system, in particular, it is necessary to extend the model to a multimodal level, integrating text, image, video and metadata analysis to more fully detect disinformation in multimedia content.
- The use of large language models (LLMs), including the advisability to adapt modern transformer models (BERT, RoBERTa, GPT, LLaMA) to the Ukrainian information context, taking into account the specifics of the language, socio-cultural markers and propaganda narratives.
- Expanding the data corpus, for example, we recommend creating an open national dataset of Ukrainian and international examples of disinformation, as well as a database of inauthentic user behaviour in social networks (bots, trolls, coordinated networks).
- Automation of monitoring and visualisation, including the development of interactive panels (dashboards) for visualisation of information flows and centres for the spread of fakes, which will allow you to track the dynamics of campaigns in real time.
- Integration with cyber defence systems, i.e. the developed algorithms can be implemented in state and corporate information security systems for automatic detection of threats and information attacks.
- The development of fact-checking tools, in particular, should be further researched to create an intelligent fact-checker assistant – a system that, based on neural networks, will be able to verify the reliability of statements in the media and social networks.
- Socio-psychological aspect, since a promising direction is the integration of linguistic analysis with behavioural models of users, which will make it possible to determine the impact of disinformation on public opinion and predict the reactions of society.
- Development of educational solutions, because based on the model, it is possible to create educational VR/AR simulators on media literacy that simulate real scenarios of spreading fakes and teach users to recognise manipulative messages.
- International cooperation, in particular, is advisable to expand the study in collaboration with European research centres dealing with countering disinformation, in order to unify methods and exchange data.

The continuation of the study is expedient and strategically crucial in view of:

- the growth of the scale and technological complexity of information influences
- Ukraine's need to create a national information security system resistant to fakes and propaganda.
- the potential to use the results obtained in the state, media, educational and scientific sectors.
- the prospect of exporting cyber defence and fact-checking technologies at the international level.

Further improvements to models, expansion of databases, and implementation of multimodal solutions will create a new-generation, adaptive, self-learning system for detecting disinformation capable of functioning effectively in the global digital environment.

References

[1] P. Dhanasekaran, H. Srinivasan, S. S. Sree, I. S. G. Devi, S. Sankar, and V. Vijayaraghavan, “SOMPS-Net: Attention-based social graph framework for early detection of fake health news,” in *Australasian Conference on Data Mining*, Singapore: Springer, pp. 165–179, 2021.

[2] M. Mayank, S. Sharma, and R. Sharma, “DEAP-FAKED: Knowledge graph based approach for fake news detection,” in *2022 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pp. 47–51, 2022.

[3] F. B. Mahmud, M. M. S. Rayhan, M. H. Shuvo, I. Sadia, and M. K. Morol, “A comparative analysis of graph neural networks and commonly used machine learning algorithms on fake news detection,” in *2022 7th International Conference on Data Science and Machine Learning Applications (CDMA)*, pp. 97–102, 2022.

[4] A. Kar, S. S. Chamarthy, K. K. Tirupati, S. Kumar, M. Prasad, and P. Vashishtha, “Social media misinformation detection: NLP approaches for risk,” *Journal of Quantum Science and Technology*, vol. 1, no. 4, pp. 88–124, 2024.

[5] Y. Shen, Q. Liu, N. Guo, J. Yuan, and Y. Yang, “Fake news detection on social networks: A survey,” *Applied Sciences*, vol. 13, no. 21, Art. no. 11877, 2023.

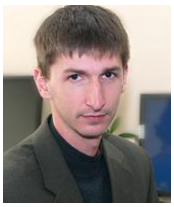
[6] V. Indu and S. M. Thampi, “Misinformation detection in social networks using emotion analysis and user behavior analysis,” *Pattern Recognition Letters*, vol. 182, pp. 60–66, 2024.

[7] R. K. Ayyasamy, C. Ponnusamy, K. N. Bhargavi *et al.*, "A hybrid deep learning framework for fake news detection using LSTM-CGPNN and metaheuristic optimization," *Scientific Reports*, vol. 15, Art. no. 41522, 2025.
[8] M. E. Almandouh, M. F. Alrahmawy, M. Eisa *et al.*, "Ensemble-based high performance deep learning models for fake news detection," *Scientific Reports*, vol. 14, Art. no. 26591, 2024.
[9] S. Munir and H. Munir, "NLP-based approach to multilingual fake news detection through social media in low-resource languages," *Preprints*, 2025.
[10] V. Vysotska, S. Popp, V. Bulatova, Z. Hu, Y. Ushenko, and D. Uhryn, "Smart tool for identifying misinformation spread sources and routes in social networks based on NLP and machine learning," *International Journal of Computer Network and Information Security*, vol. 17, no. 5, pp. 114–165, 2025.
[11] M. Nyzova, V. Vysotska, L. Chyrun, Z. Hu, Y. Ushenko, and D. Uhryn, "Smart tool for text content analysis to identify key propaganda narratives and disinformation in news based on NLP and machine learning," *International Journal of Computer Network and Information Security*, vol. 17, no. 4, pp. 113–175, 2025.
[12] R. Lynnyk, V. Vysotska, Z. Hu, D. Uhryn, L. Diachenko, and K. Smelyakov, "Information technology for modelling social trends in Telegram using E5 vectors and hybrid cluster analysis," *International Journal of Information Technology and Computer Science*, vol. 17, no. 4, pp. 80–119, 2025.
[13] K. I. Roumeliotis, N. D. Tselikas, and D. K. Nasiopoulos, "Fake news detection and classification: A comparative study of convolutional neural networks, large language models, and natural language processing models," *Future Internet*, vol. 17, no. 1, 2025.
[14] L. Ji and H. Xie, "A modified method on improving power grid resilience based on the Louvain algorithm," *Sustainable Energy, Grids and Networks*, Art. no. 101883, 2025.
[15] S. H. H. Anuar, Z. A. Abas, N. M. Yunos, N. H. M. Zaki, N. A. Hashim, and M. F. Mokhtar, "Comparison between Louvain and Leiden algorithm for network structure: A review," *Journal of Physics: Conference Series*, vol. 2129, no. 1, Art. no. 012028, 2021.
[16] V. A. Traag, L. Waltman, and N. J. van Eck, "From Louvain to Leiden: guaranteeing well-connected communities," *Scientific Reports*, vol. 9, no. 1, pp. 1–12, 2019.
[17] N. Singlan, F. Abou Choucha, and C. Pasquier, "A new similarity based adapted Louvain algorithm (SIMBA) for active module identification in p-value attributed biological networks," Scientific Reports, vol. 15, no. 1, Art. no. 11360, 2025.
[18] B. Sharma, P. Kathirvel, and S. R. Hegde, "Identifying communities in the virus–host protein–protein interaction networks," in Computational Virology, New York, NY, USA: Springer, pp. 307–319, 2025.
[19] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," Journal of Statistical Mechanics: Theory and Experiment, no. 10, Art. no. P10008, 2008.
[20] R. Hernández, I. Gutiérrez, and J. Castro, "Social network analysis: A novel paradigm for improving community detection," International Journal of Computational Intelligence Systems, vol. 18, no. 1, p. 87, 2025.
[21] D. H. Do and T. H. D. Phan, "An improvement on the Louvain algorithm using random walks," Journal of Combinatorial Optimization, vol. 50, no. 2, Art. no. 14, 2025.
[22] Vysotska, V., Hu, Z., Mykytyn, N., Nagachevska, O., Hazdiuk, K., & Uhryn, D. (2025). Development and Testing of Voice User Interfaces Based on BERT Models for Speech Recognition in Distance Learning and Smart Home Systems. International Journal of Computer Network and Information Security (IJCNIS), 17(3), 109-143.
[23] Vysotska, V., Przystupa, K., Kulikov, Y., Chyrun, S., Ushenko, Y., Hu, Z., & Uhryn, D. (2025). Recognizing Fakes, Propaganda and Disinformation in Ukrainian Content based on NLP and Machine-learning Technology. International Journal of Computer Network and Information Security, 17(1), 92-127.
[24] Vysotska, V., Yatsyshyn, M., Romanchuk, R., & Danylyk, V. (2025). Information technology for automatic detection of calls for terrorist acts in social media posts and comments based on machine learning. CEUR Workshop Proceedings, Vol-4126, 2025, 306-360, https://ceur-ws.org/Vol-4126/paper18.pdf
[25] Motyka, V., Vysotska, V., Chyrun, L., Markiv, O., Chyrun, S., & Kolyasa, L. (2023, October). System Project for Ukrainian-language Feedback Tonality Analysis in the Health Care Field Based on BERT Model. In 2023 IEEE 18th International Conference on Computer Science and Information Technologies (CSIT) (pp. 1-6). IEEE.
[26] Victoria Vysotska, Denys Ptushkin, Rostyslav Fedchuk, Roman Lynnyk, Probabilistic thematic modelling of Ukrainian-language texts based on the Latent Dirichlet Allocation algorithm. CEUR Workshop Proceedings, Vol-4126, 2025, 226-291, https://ceur-ws.org/Vol-4126/paper16.pdf
[27] N. Khairova, Y. Holyk, D. Sytnikov, Y. Mishcheriakov, N. Shanidze, (2024). Topic modelling of Ukraine war-related news using Latent Dirichlet Allocation with collapsed Gibbs sampling. In 8th International Conference on Computational Linguistics and Intelligent Systems, Lviv, Ukraine, April 12-13, 2024 (pp. 1-15). CEUR-WS.
[28] V. Vysotska, N. Babkova, D. Huliieva, Z. Kochuieva, N. Ugolnikova, M. Kozulia, Identifying fake news in political discourse behavior using machine learning methods, CEUR Workshop Proceedings, Vol-4015, 2025, 163-174, https://ceur-ws.org/Vol-4015/paper12.pdf
[29] V. Vysotska, M. Nazarkevych, D. Shamota, Information Technology for Referencing Ukrainian-Language News at Detecting Disinformation in Cybersecurity, CEUR Workshop Proceedings, Vol-4150, 2025, 81-95, https://ceur-ws.org/Vol-4150/paper8.pdf
[30] V., Vysotska, A., Chupryna, N., Valenda, O. Konduforov, Evaluating the in-context learning capabilities of large language models for misinformation detection for Ukrainian news. CEUR Workshop Proceedings, Vol-4110, 2025, 181-197, https://ceur-ws.org/Vol-4110/paper15.pdf
[31] Pankratz, B., Kamiński, B., & Prałat, P. (2024). Performance of community detection algorithms supported by node embeddings. Journal of Complex Networks, 12(4), cnae035.
[32] Melhem, A., Halloush, Z., & Aleroud, A. (2024, October). Dynamic Network Analysis of Cognitive Attacks Using Meta-Network Modeling and Community Detection Algorithms. In 2024 IEEE 6th International Conference on Cognitive Machine Intelligence (CogMI) (pp. 240-246). IEEE.
[33] Yang, J., Vega-Oliveros, D., Seibt, T., & Rocha, A. (2021, December). Scalable fact-checking with human-in-the-loop. In 2021 IEEE International Workshop on Information Forensics and Security (WIFS) (pp. 1-6). IEEE.

[34] Zheng, Q., Zeng, L., & Gu, Z. (2025, October). From IPs to Threat Groups: Community Detection on Rule-Enhanced Homology Graphs. In International Conference on Advanced Data Mining and Applications (pp. 316-324). Singapore: Springer Nature Singapore.
[35] Alrasheed, H. (2021). Word synonym relationships for text analysis: A graph-based approach. Plos one, 16(7), e0255127.
[36] Tokala, S., Enduri, M. K., Lakshm, T. J., & Hajarathaiah, K. (2025). Evaluating Community Detection Algorithms: A Focus on Effectiveness and Efficiency. Journal of Scientometric Research, 14(1), 62-74.

Authors' Profiles

Victoria Vysotska is a Professor in the Information Systems and Networks Department at Lviv Polytechnic National University and Combating Cybercrime Department at Kharkiv National University of Internal Affairs. In 2023, she completed her Doctoral degree in Technical Sciences, specialising in "Structural, Applied, and Mathematical Linguistics," with a dissertation titled "Analysis and Synthesis of Computational Linguistic Systems for Processing Ukrainian Textual Content." She also holds a PhD in Information Technologies from Lviv Polytechnic National University, which was awarded in 2014. Victoria has authored over 500 publications, and her primary research interests focus on identifying disinformation, fake news, and propaganda, as well as detecting disinformation sources and inauthentic behaviour (e.g., bots) in coordinated groups through the use of artificial intelligence. Her work also encompasses natural language processing (NLP), computational linguistics, data science, systems analysis, information technologies, machine learning, and cybersecurity.



Lyubomyr Chyrun is an Associate Professor in the Information Systems and Networks Department at Lviv Polytechnic National University and the Department of Applied mathematics at the Ivan Franko National University of Lviv. Since 2001, he has worked at Lviv Polytechnic University's Institute of Computer Sciences and Information Technologies as an associate professor in the Department of Information Systems and Networks. In 2007, he defended his PhD thesis on May 1, 2002 - "Mathematical modelling and computational methods". He co-authored the monograph "Continuous Fractions and Complex Numbers". He is the author of more than 120 scientific publications. His areas of scientific interest include pattern recognition, the application of numerical methods in information technologies, object-oriented programming, web technologies, fake news identification, natural language processing, computer linguistics, data science, system analysis, and machine learning.

Oleksandr Lavrut is a Professor in the Department of Tactics at Hetman Petro Sahaidachnyi National Army Academy. In 2020, he received his Doctor of Engineering Sciences degree in Information Systems and Technologies. He defended his dissertation for the degree of Candidate of Technical Sciences in 2003 in the field of "Cybernetics, Control Systems and Communications". He also holds the academic title of Professor, which he received in 2021. Oleksandr has authored over 250 publications, and his primary research interests focus on the theoretical aspects of construction telecommunications networks, as well as on the creation and operation of control and communication systems for various purposes. His work also covers machine learning and cybersecurity.

Zhengbing Hu: Prof., Deputy Director, International Centre of Informatics and Computer Science, Faculty of Program Systems and Applied Mathematics, National Technical University of Ukraine "Kyiv Polytechnic Institute", Ukraine. Adjunct Professor, School of Computer Science and Artificial Intelligence, Hubei University of Technology, China. Visiting Prof., DSc Candidate at the 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.

Yurii Ushenko: World's Top 2% Research Scientist (2020, 2021, 2022, 2024). Nominee of The Photonics100 2025 list by ElectroOptics. Winner of 1000 Talents Plan Program, PhD, DSc, Prof., at Shaoxing University, Shaoxing, Zhejiang Province 312000, China. Professor, Head of Computer Science Department, Chernivtsi National University, Ukraine. His research interests include intelligent information systems design, data mining, pattern recognition and digital image processing, artificial neural networks, laser polarimetry and interferometry.

How to cite this paper: Victoria Vysotska, Lyubomyr Chyrun, Oleksandr Lavrut Zhengbing Hu, Yurii Ushenko, "Hybrid Information Technology for Automatic Detection of Disinformation and Inauthentic Behaviour of Chat Users based on NLP, Machine Learning, and Graph Analysis", International Journal of Information Technology and Computer Science(IJITCS), Vol.18, No.1, pp.180-245, 2026. DOI:10.5815/ijitcs.2026.01.10