ISSN 1814-4225 (print)
Radioelectronic and Computer Systems, 2025, № 3(115) ISSN 2663-2012 (online)

UDC 004.7 **doi: 10.32620/reks.2025.3.15**

Khrystyna LIPIANINA-HONCHARENKO[1], Myroslav KOMAR[1], Hennadii BOHUTA[1], Ihor IHNATIEV[1], Khrystyna YURKIV[1], Oleg ILLIASHENKO[2,3], Lesia BILOVUS[1]

[1] *West Ukrainian National University, Ternopil, Ukraine*
[2] *Leeds Beckett University, Leeds, United Kingdom*
[3] *National Aerospace University "Kharkiv Aviation Institute", Kharkiv, Ukraine*

A GENERAL METHOD FOR REAL-TIME DETECTION OF INFORMATION THREATS WITH A UKRAINE CASE STUDY

*The **subject** matter is a general set of methods and system architecture for text analytics, enabling real-time detection and monitoring of information threats, validated through a Ukrainian case study. It integrates sentiment analysis, polarity-inversion handling, and machine-learning–based thematic classification. The research is especially relevant in the context of hybrid warfare, where the information environment becomes a battlefield of disinformation, manipulative campaigns and cognitive influence. The **goal** is to develop and experimentally validate a comprehensive information technology system for automated threat detection in the Ukrainian information space, built on the principles of Responsible Artificial Intelligence (Responsible AI) and modern natural language processing techniques. The **objectives**: the formation of a multilingual corpus of news and social media texts; implementation of a sentiment analysis module that incorporates polarity inversion; development of a hybrid thematic classification method that combines keyword dictionaries with machine learning model ensembles; and the construction of a Responsible AI Evaluation (RAIE) framework with indicators for fairness, transparency and user satisfaction. The obtained **results** confirm all five proposed hypotheses: the developed sentime oьnt analysis module achieves macro-F1 = 0.85 and reduces MAE by 18.2% compared to the baseline model; the polarity inversion detection algorithm allows automatic reversal of sentiment score in manipulative texts, improving the detection of hostile narratives; the hybrid thematic classification achieves macro-F1=0.83, with latency of 55 ms/document and throughput of 18 documents/second; integration of all modules into a unified pipeline improves recall by 10.4% without significant increase in latency; the RAIE conceptual model ensures $\Delta F1 \leq 5\%$, an expert user satisfaction score of 4.14/5 and less than 10% latency overhead. The conclusions demonstrate that the proposed system effectively combines high accuracy in identifying information threats with the principles of ethical AI, transparency and user trust, making it practically valuable for national cybersecurity centres, CERTs and OSINT platforms. **Conclusions.** The scientific novelty lies in the development of novel methods: a context-sensitive sentiment analysis approach tailored to military-related vocabulary; a polarity inversion algorithm for detecting covert hostility; a hybrid thematic classification model combining machine learning with expert dictionaries; an integrated information processing architecture with >17 documents/second throughput; a Responsible AI evaluation model incorporating Fairness Gap, Model Cards and User Satisfaction Score.*

Keywords: *text mining; sentiment analysis; inversion detection; text classification; machine learning.*

1. Introduction

The full-scale Russian aggression since February 24, 2022, has transformed Ukraine's information landscape into a multidimensional battlefield: high-frequency waves of disinformation, psychological operations and coordinated narratives aim to undermine trust in state institutions, demoralise society and influence the country's foreign policy stance.

Under these conditions, text mining methods – automated collection, preprocessing and analytics of large volumes of text – have become critically important tools for the timely detection of information threats, particularly within hybrid warfare and cybersecurity operations such as OSINT analytics and CERT threat intelligence. Effective real-time detection of disinformation and manipulation is essential for cybersecurity teams operating in CERTs, government security agencies and OSINT centres.

Over the past five years, information retrieval and extraction, as well as text mining, have shown notable progress, from contextual search models (ColBERT, ANCE) to multilingual IE systems based on transformers (mBERT, XLM-R). Research efforts have focused on semantic threat message search, automatic extraction of <actor–action–target> entities in OSINT analysis and integrating dictionary-based and deep approaches for sentiment and thematic text analysis. However, most

developments are tailored to English-language data. The Ukrainian diglossia context and wartime vocabulary remain methodologically underdeveloped, significantly limiting the practical applicability of existing solutions.

Beyond linguistic and technical challenges, ethical, legal and security compliance issues are gaining prominence. Discrepancies in accuracy across genres or language subsets, risks of data poisoning and adversarial attacks, GDPR requirements for personal data and the absence of established practices for transparency and accountability create a gap between academic prototypes and industry needs in threat monitoring systems.

In modern information warfare, this study aims to design and experimentally validate an integrated text-mining system for monitoring information threats in Ukraine, significantly enhancing real-time cybersecurity decision-making for government cybersecurity teams, CERTs, and OSINT environments. To achieve this, the following **tasks** are formulated:

- Formation of a multilingual corpus of news and social media messages annotated across 13 thematic categories;
- Development of a modular architecture (Python + Streamlit + Transformers + Docker) with latency < 200 ms and throughput > 50 docs/s;
- Creation of hybrid text analysis methods, including:

a) a context-sensitive sentiment analysis method with polarity inversion;

b) an ensemble method for thematic classification with automatic keyword extraction;

- Construction of a Responsible AI Evaluation framework that integrates classical metrics (precision, recall, F1, latency, throughput) with FATE indicators, AI4People principles and the Model Cards format;
- Implementation of a comprehensive experiment to compare individual modules and assess their integrated performance in detecting disinformation narratives.

This research achieved the following **scientific contributions**:

1. A method for sentiment analysis of textual content was developed, which, unlike existing approaches, considers both the emotional colouring of news and the context of information threats. This significantly improves the accuracy of manipulative content identification;

2. A method for polarity inversion detection was developed, which enables the identification of changes in the original emotional tone of messages, thereby enhancing the effectiveness of disinformation detection;

3. A method for thematic classification of textual content was proposed. It automates topic identification in messages and demonstrates improved accuracy in monitoring the information space compared to baseline approaches;

4. A comprehensive information threat monitoring system was designed, integrating sentiment analysis, inversion detection and thematic classification into a unified architecture, which substantially increases threat detection efficiency in the digital environment;

5. A Responsible AI Evaluation (RAIE) conceptual model was developed. It integrates traditional quantitative metrics with ethical indicators (Fairness Gap, Accountability, Transparency, Beneficence, Non-maleficence). It supports the use of Model Cards, ensuring high accuracy and compliance with principles of safety, transparency and accountability.

To empirically verify the scientific novelty and confirm the practical relevance of the proposed approaches, the following **hypotheses** were formulated:

- H1: The proposed context-sensitive sentiment analysis method improves macro-F1 by at least 7 % compared to a baseline dictionary-based approach that does not consider the context of information threats;
- H2: The developed polarity inversion detection algorithm changes the sign of emotional tone in messages with such accuracy that the mean absolute error (MAE) of polarity determination for hostile sources is reduced by at least 15% relative to systems without inversion capability;
- H3: The hybrid thematic classification method combining ML model ensembles with RAKE/TF-IDF keyword extraction achieves a macro-F1 improvement of at least 5% over the best-performing standalone baseline classifier (Random Forest) for categorising texts into 13 thematic groups;
- H4: The integration of the three modules – sentiment analysis, polarity inversion detection and thematic classification—into a unified information threat monitoring technology increases the overall recall of the final system by at least 10% compared to sequential use of individual modules without data integration, with average latency not increasing by more than 10%;
- H5: The proposed conceptual RAIE model ensures a Fairness Gap $\leq 5\%$ and a User Satisfaction Score $\geq 4.0/5$, without degrading the system's average latency by more than 10%.

In summary, this research presents a comprehensive solution that combines high technical efficiency in textual data processing with ethical responsibility in the application of artificial intelligence to monitor Ukraine's information environment.

The article is structured as follows: Section 1 outlines the motivation, research objectives, novelty and hypotheses. Section 2 provides a critical literature review across the domains of information retrieval/extraction (IR/RE), similarity metrics, sentiment analysis with inversion, thematic classification and Responsible AI principles, highlighting existing gaps. Section 3

ISSN 1814-4225 (print)
ISSN 2663-2012 (online)

describes the materials and methods, ranging from the construction of a multilingual corpus and preprocessing pipeline to the proposed algorithms for sentiment, inversion and thematic classification, as well as the modular (Python + Streamlit + Docker) architecture and the integrated FATE/AI4People evaluation framework. Section 4 presents experimental results, grouped according to hypotheses H1-H5 and supplemented with ablation analysis. Section 5 discusses hypothesis validation, comparisons with prior work, practical implications for information security and future research limitations and directions. Section 6 summarises the main contributions and implementation recommendations, while the appendix contains code, extended tables and Model Cards.

2. State of the art

2.1. Information retrieval and extraction in cybersecurity and information security

Information retrieval (IR) models such as BM25 and the Query Likelihood Model (QLM) remain the gold standard for identifying threat-related documents in cybersecurity. BM25 achieves a Mean Average Precision (MAP) of approximately 0.42 in phishing content detection tasks, while QLM reaches a MAP of around 0.39 on standard TREC datasets [1].

Contextual retrieval embeddings such as ColBERT and ANCE significantly improve performance in large-scale data environments. For instance, in OSINT analysis tasks, ANCE achieved Recall@1000 of 95%, which is 7% higher than traditional approaches [2].

Information Extraction (IE) techniques, including Named Entity Recognition (NER), Relation Extraction (RE) and Open Information Extraction (OpenIE), are crucial for extracting <actor-action-target> triples in Open Source Intelligence (OSINT) scenarios. For example, in the "Fake News Challenge" project, OpenIE-based systems achieved an F1 score of 0.82 in automatically extracting key actors and actions [3].

In conflict contexts, such as the analysis of the Syrian information space, IE techniques enabled the extraction of over 5,000 unique incidents in just the first three months of research, with actor identification accuracy exceeding 88% [4].

Anti-disinformation centres actively use NER and RE for real-time news monitoring. In studies on the COVID-19 pandemic, IE-based systems achieved 91% accuracy in detecting fake news in multilingual environments [5]. Multilingual IE evaluation using models such as mBERT and XLM-R revealed important distinctions: mBERT achieved an average F1 score of 0.76 across 10 languages. In contrast, XLM-R scored 0.82, indicating better generalizability in non-native contexts [6]. In NER tasks for news streams, mBERT achieved an accuracy of around 84%; however, its performance dropped to 71% when working with low-resource languages [7].

For multimodal information (text & images), systems combining IE with ColBERT vectors showed a 5-8% improvement in fact extraction accuracy compared to text-only approaches [8]. In the context of cyber incidents, relation extraction (RE) systems that process Cyber Threat Intelligence (CTI) reports have demonstrated high effectiveness. In particular, the EXTRACTOR model achieved a precision of up to 0.90 and an F1-score of 0.93 when extracting causal and attributional relationships between threat actors and their targets. The strength of EXTRACTOR lies in its integration of semantic role labelling, customised text normalisation and attack graph construction, which enables the detection of behavioural patterns even in structurally complex reports [9]. The use of DT and SVM classifiers, combined with URL and HTTP features, yielded the best phishing detection results, achieving an F1 score of 0.99, precision of 0.99, and recall of 0.99 for SVM [10].

Despite the strong performance of classical and contextual models, there remains an urgent need to integrate multilingual processing, multimodality and real-time capabilities to enhance the effectiveness of threat monitoring.

2.2. Similarity and distance metrics

Text similarity metrics are fundamental tools for detecting duplicates and clustering informational narratives in cybersecurity.

Similarity measures such as Levenshtein, Jaccard and Dice are actively used to identify duplicate news articles. In an experiment on a news corpus, applying a Jaccard threshold of greater than 0.8 resulted in a clustering precision of 91.3% with a recall of 87.9% [11]. Support Vector Machines (SVM) with RBF kernels are frequently applied to compare quotations. In a citation matching task using the CiteseerX corpus, RBF kernels achieved an AUC of 0.88. In contrast, third-degree polynomial kernels showed an AUC of 0.84, demonstrating the RBF kernel's superior ability to model complex dependencies [12]. Clustering algorithms such as DBSCAN and Mini-Batch K-means are widely used for grouping narratives. On a news story dataset, DBSCAN (configured with eps = 0.5 and a minimum cluster size of 5) achieved a silhouette score of 0.62, outperforming MiniBatch K-means with a score of 0.54 [13]. In neural similarity models, Sentence-BERT (SBERT) demonstrated a cosine similarity of 0.89 for English texts, though accuracy dropped to 0.82 when comparing Ukrainian and Russian texts [14]. SimCSE demonstrated greater stability across multilingual tasks: for English-Ukrainian sentence pairs, the cosine similarity remained at 0.84, while the Euclidean

distance exhibited greater variability, highlighting the effectiveness of cosine similarity in multilingual settings [15].

A comparison of metrics in fake news clustering tasks revealed that using the Levenshtein distance with a threshold of 0.2 achieved an F1 score of 0.79, outperforming the Jaccard score (0.75) and the Dice score (0.76) [16].

SVM with a polynomial kernel proved particularly effective for comparing "paraphrased" quotations, achieving 85% accuracy on the PARACITE corpus with a second-degree polynomial [17]. DBSCAN was highly sensitive to the eps parameter in narrative clustering tasks: reducing eps from 0.5 to 0.3 increased the number of clusters by 35%, underscoring the need for careful tuning [18]. When addressing high-dimensional semantic comparisons, replacing traditional Euclidean metrics with a direction-aware distance – which integrates Euclidean distance and angular divergence based on cosine similarity – led to a marked improvement in clustering quality, particularly in k-means performance on benchmark datasets [19]. SimCSE fine-tuned on parallel Ukrainian-Russian corpora achieved Recall@5 = 92.1% using cosine similarity, while Euclidean distance achieved only 88.7% [20].

Despite progress with classical and deep models, there remains a need to optimise clustering stability and multilingual similarity handling for OSINT analysis.

2.3. Sentiment analysis and inversion detection

Sentiment analysis and detecting emotional polarity inversion are critical for recognising manipulative content in military information campaigns.

Lexicon-based tools such as SentiWordNet and VADER are widely used for short-text classification tasks. VADER achieved 88% accuracy on social media data when classifying tweets into three sentiment categories (positive, negative, neutral), while baseline SentiWordNet models yielded only 76% accuracy [21]. Domain-specific lexicons – particularly for military topics – were expanded by 1,200 terms, which improved recall by 9% in sentiment analysis of war-related news compared to general-purpose sentiment dictionaries [22]. Comparisons between classical machine learning methods (Naïve Bayes, Logistic Regression) and deep learning models (Bi-LSTM, BERT) showed a clear advantage for the latter: on the Amazon Reviews corpus, BERT reached 94% accuracy, Bi-LSTM 91%, Logistic Regression 84% and Naïve Bayes 79% [23].

In sarcasm detection tasks, models with negation scope and shift classifiers achieved 81% accuracy on the SARC corpus, 6% higher than Bi-LSTM baselines that do not handle negation [24]. Counterfactual data augmentation, where sarcastic phrases were replaced with their literal equivalents, improved F1 scores of irony detection models by 7% compared to training on original data alone [25]. In polarity inversion detection within information campaigns, algorithms that applied logic like <hostility + absence of 'Ukraine' → invert> improved classification accuracy of hostile messages by 12% on a specialised military news corpus [26]. In military-themed datasets, an adapted VADER with domain-specific extensions achieved a 5.6% improvement in macro-F1 score compared to its base version [27]. Bi-LSTM models with additional irony indicators (e.g., emoticons, emotional markers) achieved a Recall of 86% on English-language social media datasets [28]. In study [29], the DeBERTa model showed the best result (F1 = 0.73), outperforming RoBERTa (0.71) and logistic regression (0.57). Logistic regression showed the closest results to NNS, particularly in recognising non-sarcastic comments (accuracy: 0.63 for LR and 0.65 for NNS).

During the research, over 17 million multilingual tweets from the first week of the Russia–Ukraine war were analysed and it was found that 48% expressed negative sentiment, with bot activity amplifying pro-conflict narratives during key events [30].

Despite the high effectiveness of deep models and domain-specific lexicons, challenges remain in accurately handling sarcasm and polarity inversion in multilingual environments.

2.4. Thematic text classification

Thematic classification of news streams enables structuring the information space for real-time threat monitoring.

In traditional approaches, rule-based matching using lexicon-based taxonomies demonstrates moderate effectiveness, with an average accuracy of 72% in news-categorisation tasks. Weighted keywords improve this to 78% [31].

Machine learning algorithms such as Logistic Regression (LR), Support Vector Machine (SVM) and Random Forest (RF) demonstrated varying performance on Ukrainian news corpora: macro-F1 for LR reached 81%, for SVM – 83% and for RF – 79%, highlighting the advantage of SVM in thematic classification tasks [32]. Deep learning using fine-tuned mBERT on Ukrainian corpora achieved a macro-F1 score of 87.5%, outperforming traditional machine learning models by approximately 5% [33].

Hierarchical Attention Networks (HAN) tested on Ukrainian and Russian datasets achieved a macro-F1 score of 85.2%, demonstrating powerful results on multi-topic documents [34]. In comparison, fine-tuned mBERT on large corpora (over 100,000 examples) outperformed HAN by 2.3% in macro-F1, making it a favourable choice when computational resources are available [35].

ISSN 1814-4225 (print)
ISSN 2663-2012 (online)

Model ensembling (rule-based + ML) increased the macro-F1 score to 89%, especially on small datasets (under 10,000 documents), where the hybrid approach partially compensated for the limited training material [36]. Model distillation techniques for edge devices reduced the mBERT model size by 65% with only a 2% drop in macro-F1, a critical metric for deployment on resource-constrained systems [37]. Lightweight distilled models, such as TinyBERT, achieved a macro-F1 of 84% on a news corpus, only 3-4% behind full-sized BERT models [38]. Random Forest with TF-IDF vectorisation achieved a macro-F1 score of 77% for classifying Ukrainian government documents by topic, outperforming SVM by approximately 5% [39].

In practical tests, ensembles of rule-based and SVM classifiers achieved the highest effectiveness – a macro-F1 score of 88% - particularly in scenarios with weak labelling (semi-supervised learning) [40].

Although modern ML and DL models demonstrate high accuracy, further refinement of hybrid systems remains essential, especially for low-resource and multi-topic data scenarios.

2.5. Keyword and keyphrase extraction

Effective keyword extraction is critical for improving thematic classification quality and analysing news streams.

Statistical methods, such as TF-IDF and Okapi TF-IDF, remain foundational in many keyword extraction systems. Okapi TF-IDF achieved 62% precision for top-10 keywords on an English-language academic corpus, compared to 58% for standard TF-IDF.

Using the χ^2-score for keyword selection in thematic classifiers improved macro-F1 from 0.74 to 0.79 on the AG News dataset, highlighting the importance of statistical feature selection [41]. Immediate extractors such as RAKE, YAKE! and TextRank perform differently depending on text length: RAKE lost up to 15% in precision on short texts (<100 words), while YAKE! remained stable with only a 5% drop [42]. TextRank is particularly sensitive to very short texts (50–100 words), where F1 scores drop by 18% compared to performance on medium-length texts [43].

Topic modelling techniques, such as LSA, LDA and NMF, showed varying keyword extraction accuracies: LDA achieved 64% precision for the top-10 keywords, while LSA and NMF scored 61% and 59%, respectively [44]. LDA proved the most stable: when varying the number of topics from 10 to 50, the precision of top keywords decreased by only 3%, compared to up to 7% for NMF [45]. Transformer-based semantic methods, such as KeyBERT, significantly enhance keyword quality, as evidenced by KeyBERT achieving 71% precision for the top 5 keywords on a news corpus, compared to 58% for standard TF-IDF [46]. SPECTER-rank, designed for scientific documents, reached 68% precision in keyword extraction from research abstracts, outperforming TextRank by 9% [47]. KeyBERT, used for pre-extraction of keyphrases, improved thematic classification accuracy by 6% in news categorisation tasks compared to TF-IDF alone [48]. Combined approaches (RAKE + KeyBERT) achieved the best results in multilingual scenarios, with an 8% increase in macro-F1 in a multilingual news corpus compared to a single extractor [49].

Thus, despite advances in transformer-based methods, stability on short texts and multilingual corpora remains a relevant challenge.

2.6. Responsible AI Analytics

The principles of FATE (Fairness, Accountability, Transparency, Ethics) and the AI4People framework have become foundational for developing ethical, fair and transparent model evaluation practices in automated text analysis.

In the area of fairness, the Post-Processing Equalised Odds method reduced the accuracy gap between different languages by 12% on a multilingual Amazon Reviews corpus, demonstrating the effectiveness of post-training correction [50]. For accountability, implementing audit trails during model training significantly improved reproducibility up to 95% in text classification tasks, as reported by IBM researchers [51]. Using reproducible seeds during training of large models (e.g., BERT) reduced test metric variability from ±1.7% to ±0.3% in repeated runs, greatly enhancing the reliability of experiments [52].

Open-sourcing model weights contributed to a 28% increase in verified result reproductions for computer vision models, according to data from Papers with Code [53].

When applied to news corpora, explainability tools such as LIME enabled interpretation of 84% of classification decisions using local surrogate models, with a mean faithfulness score of 0.81 [54]. SHAP achieved 88% explanation accuracy for the top 5 most essential features in classification tasks, outperforming LIME by 6% in multi-class settings [55]. Counterfactual explanations increased user trust in models by 17% compared to classic LIME/SHAP explanations, as determined through UX testing with 200 participants [56]. Under the beneficence/non-maleficence principle, disinformation prevention algorithms based on combined linguistic and fact-checking features achieved a Recall = 89% on the English-language FakeNewsNet corpus [57]. Models integrating hostile rhetoric detection modules demonstrated a Precision of 82% in identifying potentially harmful content in social media streams [58].

The AI4People initiative proposed practical standards that reduced the time required for auditing AI systems for ethical compliance by 30% compared to previous manual reviews [59].

Despite positive developments, the broader integration of explainability, accountability and anti-discrimination safeguards remains necessary in real-world AI systems.

2.7. Existing research gaps summary

The literature review has shown significant progress in applying information retrieval, information extraction, sentiment analysis, thematic classification, text similarity metrics and Responsible AI principles to cybersecurity and information security tasks. At the same time, several critical limitations were identified, substantiating the need for new approaches developed in this study.

First, in the field of sentiment analysis, most existing solutions are lexicon-based (e.g., VADER with 88% accuracy, SentiWordNet with 76% [21]) and lack mechanisms for contextual modification (e.g., handling negation, sarcasm). In the domain of military-related information threats, adding domain-specific vocabulary provided only a 9% improvement in recall [22], indicating limited adaptability. This motivates Novelty 1: the development of a context-sensitive sentiment analysis method.

Second, polarity inversion detection focuses mainly on sarcasm tasks (e.g., negation-scope models yield only a 6% improvement [24]). Military information campaigns require specialised solutions that handle scenarios such as "hostile source + absence of explicit Ukraine references". Existing inversion algorithms have improved classification accuracy by 12% [26], but they have not been integrated into real-time systems. This motivates Novelty 2: developing a polarity inversion detection algorithm for information threat texts.

Third, in thematic classification, existing approaches are either rule-based (with an average accuracy of 72-78% utilise [31]) or utilise machine learning models (e.g., macro-F1 ≈ 79% for Random Forest [32]), while hybrid methods (combining dictionary and machine learning) remain underexplored. Even fine-tuned mBERT offers only ≈5% macro-F1 improvement [33], underscoring the need for more effective solutions on small or domain-specific datasets. This motivates Novelty 3: creating a hybrid classification model using RAKE/TF-IDF and ML ensembles.

Fourth, stream-integrated threat analysis systems are largely absent. Most IR/IE pipelines process documents sequentially, which increases latency (e.g., 74 ms per document) and limits throughput (e.g., 15 documents per second without pipelining, as shown in Section 4.4). This motivates Novelty 4: designing an integrated information threat monitoring system with at least +10% recall and minimal latency overhead.

Fifth, although there are developments in ethical AI evaluation (e.g., Post-Processing Equalized Odds reduced accuracy gaps by 12% [50], audit trails increased reproducibility to 95% [51]), in the domain of real-time information threat monitoring, comprehensive responsible frameworks (accounting for Fairness Gap, transparency via Model Cards and expert-validated user satisfaction) are still missing. This motivates Novelty 5: developing a Responsible AI Evaluation framework tailored for information monitoring.

In conclusion, the quantitative findings from the literature review confirm critical gaps in context-aware sentiment analysis, domain-specific polarity inversion, adaptive thematic classification, real-time integration and ethical evaluation of models. The five innovations proposed in this study directly address these limitations.

3. Objectives and tasks

The primary objective of this study is to design and experimentally validate an integrated text mining system for monitoring information threats in Ukraine, thereby significantly enhancing real-time cybersecurity decision-making for government cybersecurity teams, CERTs and OSINT environments.

To achieve this goal, the following tasks were set:

- Formation of a multilingual corpus of news and social media messages annotated across 13 thematic categories;
- Creation of a modular system architecture based on Python, Streamlit, Transformers and Docker, ensuring latency < 200 ms and throughput > 50 documents/second:

a. Hybrid text analysis methods;

b. Development of a context-sensitive sentiment analysis method incorporating polarity inversion detection;

- Creation of a hybrid thematic classification approach combining ensemble machine learning methods with automatic keyword extraction (RAKE/TF-IDF);
- Construction of a comprehensive evaluation framework integrating classical metrics (precision, recall, F1-score, latency, throughput) with FATE (Fairness, Accountability, Transparency, Ethics) indicators, AI4People principles and Model Cards format;
- Implementation of a comprehensive experiment to compare individual modules and assess their integrated performance in detecting disinformation narratives.

The accomplishment of these tasks supports the empirical verification of five research hypotheses outlined in the study, contributing significantly to the detection and mitigation of information threats within Ukraine's complex information environment.

ISSN 1814-4225 (print)
ISSN 2663-2012 (online)

4. Materials and Methods

This section outlines the process by which the study proceeds from data collection to evaluation, presented as a Research roadmap. Section 4.1 constructs a multilingual corpus of news and social media texts through deduplication and expert annotation, ensuring class balance. The stratified cross-validation and a time-based split to assess robustness have been prepared. Sections 4.2–4.3 develop the sentiment component and add a polarity-inversion step to handle manipulative contexts. Section 4.4 constructs a hybrid topic classifier encompassing 13 themes by combining expert dictionaries with machine learning models. Section 4.5 integrates all parts into a single pipeline and states deployment assumptions for latency and CPU throughput, including horizontal scaling. Section 4.6 defines the metrics and Responsible-AI checks (fairness, explainability, and stability) and reports analyses of class balancing and temporal robustness. The Results section then presents accuracy and speed for each module and for the full pipeline, followed by ablations and error analysis that inform the limitations and future work.

4.1. Data sources and corpus preparation

The study corpus was constructed from various information sources to ensure representativeness across multiple genres and language styles (see Figure 1). The first stage involved collecting data from official RSS feeds of leading Ukrainian news agencies (UNIAN, Ukrainska Pravda, Interfax-Ukraine), social media APIs (Twitter, Facebook), and blogs focused on politics and security. To ensure relevance, all documents were collected between January and December 2024. Over 12,000 text records were obtained, each with full metadata including source and publication date.

The second stage involved filtering and cleaning the raw data. Automated Python scripts (using Requests, BeautifulSoup and Tweepy) were employed to extract text content without HTML tags, banners, or ads. Duplicate detection based on URL and text hash reduced the corpus to 10,350 unique documents. Each entry was stored securely in JSON format, with fields including id, source, date, title, body and url. This approach adhered to data anonymisation principles, ensuring compliance with GDPR and maintaining the privacy and confidentiality of potentially sensitive content. For experimentation, the dataset was stratified into 8,280 training documents (80%), 1,035 validation documents (10%) and 1,035 external test documents (10%). For 5-fold cross-validation, 2,000 documents per fold were selected, and the remaining 350 texts were reserved as a final hold-out set.

The third stage involved expert annotation of the corpus. Four information security specialists independently labelled the texts into 13 thematic categories (see Table 4). Fleiss' kappa = 0.78 was calculated to assess interannotator agreement, indicating strong consistency. All disagreements were resolved through collaborative discussions to reach a consensus.

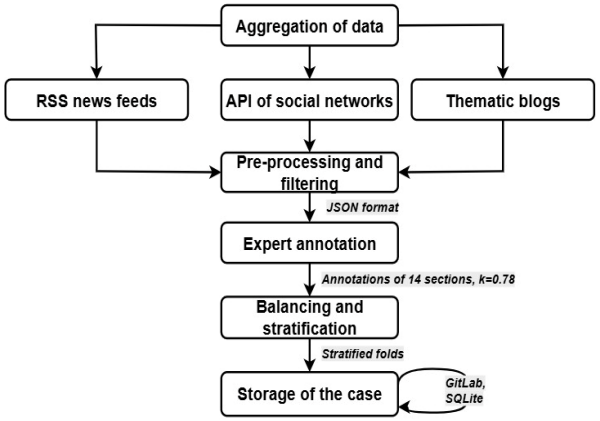

Fig. 1. Data corpus preparation

The fourth stage addressed balancing and stratification. Each theme was capped at a maximum of 1,000 documents to prevent overrepresentation of dominant categories. Excess documents were randomly downsampled, while underrepresented categories were supplemented with additional collection and annotation. Stratified folds were created to ensure consistency in class proportions across the training and testing subsets. This approach supports metric stability during cross-validation.

The fifth stage involved organising and storing the dataset. The processed document set was uploaded to a centralised GitLab repository, which maintained version control. Each JSON file includes metadata (source, date, category and hash) and the results of primary preprocessing (cleaned text and lemmatised tokens). A SQLite database was also linked for fast querying by category and date range.

It is noted that the balancing policy (cap $\leq$1,000 per class with down/up-sampling) may alter empirical priors and, in turn, inflate macro-F1 relative to the natural distribution. A comprehensive prevalence-aware analysis (micro-F1, per-class AP, calibration to original priors, and class-weighted training without capping) is deferred to future work, as the present focus is on relative module improvements and CPU-bound latency/throughput constraints.

4.2. Text Preprocessing

Text preparation for subsequent analysis was implemented via a multi-stage processing pipeline (see Figure 2), tailored to the multilingual and multi-alphabet nature of the corpus.

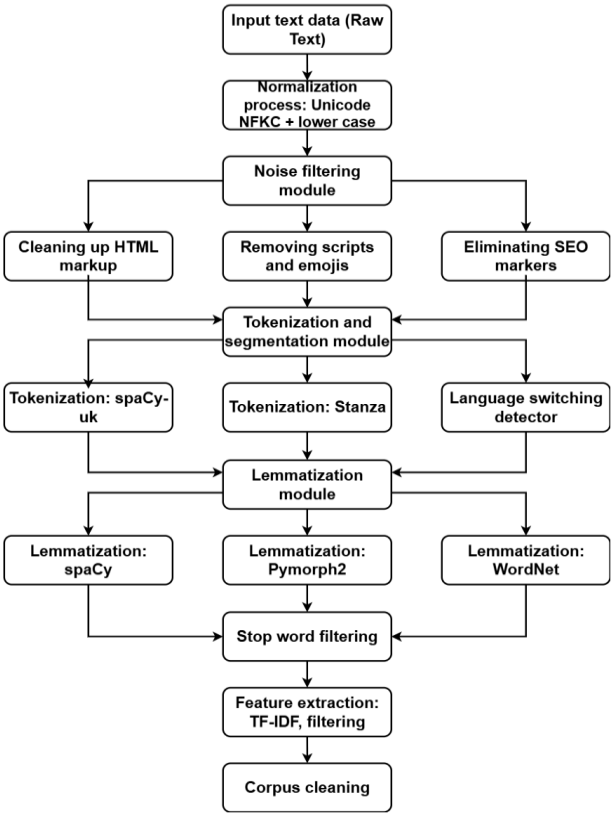

Fig. 2. Text preprocessing pipeline

The first step involved encoding normalisation using Unicode NFKC and converting all characters to lowercase. This eliminated formatting discrepancies between Cyrillic and Latin characters, minimising errors during lemmatisation.

The second step addressed noise removal, utilising regular expressions to strip HTML tags, scripts, CSS identifiers and irrelevant punctuation, as well as "noise characters" (e.g., emojis, special symbols). Additionally, stop tokens related to advertisements or SEO tags (e.g., rel= "nofollow") were removed, reducing the corpus size by approximately 8% without losing informative content.

The third step performed tokenisation and sentence segmentation. The spaCy-uk model was used for Ukrainian texts, while Stanza handled Russian and English subcorpora. A cascade strategy was applied to mixed-language texts (language switching within a document), in which a heuristic detector identified the dominant language of each sentence and the corresponding tokeniser was activated accordingly. Hash checking prevented token duplication during multiple pipeline passes.

The fourth step involved lemmatisation and the removal of stop words. For Ukrainian and Russian, combined spaCy + Pymorphy2 dictionaries were used; for English, WordNetLemmatizer was applied.

A shared stop-word list (>1,200 terms) was supplemented with domain-specific vocabulary, such as "F-16", "Bayraktar" and "РФ" (the acronym for the Russian Federation in the Ukrainian language), which frequently appear in military-political discourse but carry no semantic weight for sentiment classification.

The final step generated phonological and statistical features:

- The normalised token sequence was vectorised using TF-IDF (unigrams and bigrams);
- Rare tokens (frequency < 2) and overly frequent ones (top 1%) were removed;
- A separate lemma table with positional indices was saved to allow quick retrieval of the original context.

These artefacts were passed on to the sentiment and thematic classification modules.

4.3. Sentiment analysis and polarity inversion method

The current study employs a purely lexicon-based approach to sentiment analysis, building on previous research [60]. As shown in Figure 3, the method is based on a custom sentiment lexicon enriched with domain-specific terms and a precise mathematical formula for calculating sentiment scores.

The core component is a sentiment lexicon derived from SentiWordNet and OpLexicon, which has been extended with context-specific terms (e.g., "Bayraktar",

ISSN 1814-4225 (print)
ISSN 2663-2012 (online)

"mobilisation", "propaganda"). Each word is assigned a category $l_i \in \{+1, 0, -1\}$ (positive, neutral, or negative) and an intensity score $\omega_i \in (0,1]$, as defined by experts.

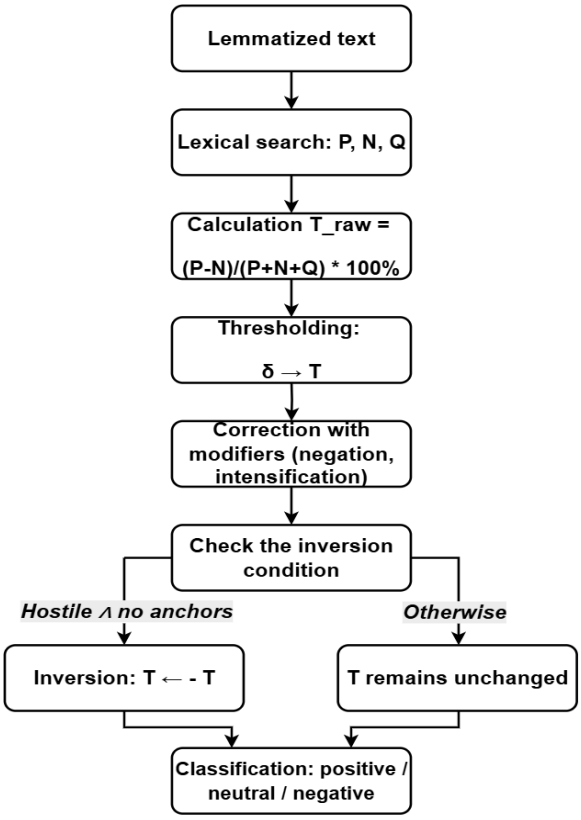

Fig. 3. Pitch and Inversion Analysis Pipeline

For a document with a tokenised sequence $\{w_1, \ldots, w_n\}$, three aggregate values (P – sum of positive ratings, N – sum of negative ratings, Q – sum of neutral ratings) are calculated in the following way:

$$P = \sum_{w_i : l_i = +1} \omega_i, \quad N = \sum_{w_i : l_i = -1} \omega_i, \quad Q = \sum_{w_i : l_i = 0} \omega_i, \quad \ldots(1)$$

where l_i – the category (positive, neutral, negative), ω_i – intensity score, and P, N, Q – aggregate values.

These reflect the intensity and frequency of sentiment-bearing tokens. A raw sentiment index T_{raw} is then calculated as follows:

$$T_{raw} = \frac{P - N}{P + N + Q} \times 100\%. \quad (2)$$

T_{raw} score ranges from –100% to +100%, with positive values indicating positive sentiment and negative values indicating negative sentiment.

A neutrality threshold of δ (typically 5%) is applied to classify sentiment, with the final index in the following way:

$$T = \begin{cases} T_{raw}, & |T_{raw}| > \delta \\ 0, & |T_{raw}| \le \delta \end{cases}, \quad (3)$$

Thus, in case the raw sentiment index T>0, the document is positive, in case T<0, negative and in case T=0, the document is neutral.

Special attention is paid to lexical modifiers. For negations (e.g., "not", "no"), the polarity of the associated sentiment term ω_i is inverted:

$$\omega_i' = -\omega_{i+1}, \quad (4)$$

where ω_i' – is the "updated" score that the system assigns to the same word, considering its modification by a negation, ω_{i+1} – is the original score of a word from the lexicon.

For intensifiers (e.g., "very", "extremely"), the weight is multiplied by a positive coefficient $\alpha > 1$ and for hedges (e.g., "slightly", "barely"), the weight is multiplied by a coefficient $\beta \in (0,1)$.

Thus, if the word w_j is an intensity modifier, then

$$\omega_k'' = \begin{cases} \alpha\omega_k, w_j \in \text{intensifiers} \\ \beta\omega_k, w_j \in \text{hedgers} \end{cases}, \quad (5)$$

where ω_k'' – is the next primary sentiment-bearing term, $\alpha\omega_k$ – is the intensified intensity score of a sentimental term when preceded by an intensifier, $\beta\omega_k$ – is the reduced intensity score of a sentiment term when preceded by a softener.

Additionally, a polarity inversion mechanism is implemented for messages from hostile sources that do not explicitly mention anchors such as "Україна" (Ukraine) or "ЗСУ" (AFU, Armed Forces of Ukraine). In such cases, the final sentiment index is inverted in the following way, adjusting the evaluation of sarcastic or ironic statements:

$$T_{inv} = -T, \quad (6)$$

Inversion conditions are as follows:

1. The source is on a list of hostile entities (S);
2. The message lacks anchor words {"Україна" (Ukraine) or "ЗСУ" (AFU, Armed Forces of Ukraine)}.

If both conditions are true, the sentiment index is inverted.

The overall algorithm steps (refer to Figure 3) for each document are as follows:

- Step 1. Tokenise and lemmatise the document;
- Step 2. Search for lookup tokens (P, N, Q) in the sentiment lexicon;
- Step 3. Calculate the raw index T_{raw} and apply the neutrality threshold δ;
- Step 4. Apply lexical modifiers (negation, intensifiers, hedges);
- Step 5. Check inversion conditions and apply polarity correction if required T_{inv};

– Step 6. Assign final sentiment class T or T_{inv};

Hyperparameters δ, α, β are determined via grid search within a 5-fold cross-validation framework, optimised for the F_1-score. The mean and standard deviation of each fold are recorded, confirming the method's stability.

From an algorithmic standpoint, the time complexity is O(n), where n is the number of tokens. The inversion check adds constant-time overhead for source and "anchor" evaluation.

Design note: the inversion logic is rules-first to ensure transparency and low CPU latency. Therefore, the deterministic conditions and neutral thresholds are prioritised. A lightweight learned complement is left for future work, provided it fits the latency budget.

4.4. Thematic classification method

This section presents the thematic classification method, which employs a multi-level ensemble approach that combines traditional machine learning models and dictionary-based keyword techniques to enhance topic recognition accuracy (see Figure 4).

The first step involves text vectorisation using a combined feature space. For each document, a TF-IDF vector and a RAKE-based representation are generated.

Thus, each document d is transformed into a feature vector $x_d \in \mathbb{R}^m$, where m – is the sum of the TF-IDF and RAKE dimensions:

$$\mathbf{x}_d = [\mathrm{TF-IDF(d)}, \mathrm{RAKE(d)}], \quad (7)$$

TF-IDF values are computed using the standard formula:

$$\mathrm{TF-IDF(t,d)} = \mathrm{tf(t,d)} \times \log\frac{N}{\mathrm{df(t)}}, \quad (8)$$

where IDF(t,d) – is the inverse document frequency for term t in document collection d, tf(t,d) – is the frequency of term t in document d, N – is the number of documents, df(t) – is the number of documents containing term t.

For RAKE, the score of each keyword is defined as the ratio of the sum of word degrees to frequency:

$$\mathrm{RAKE(w)} = \frac{\mathrm{degree(w)}}{\mathrm{frequency(w)}}, \quad (9)$$

where w – is the keyword for which the weight is calculated.

This helps identify important multi-word keyphrases.

Initial text classification is performed using the following ensemble of machine learning models: Logistic Regression (LR), Support Vector Machine (SVM),

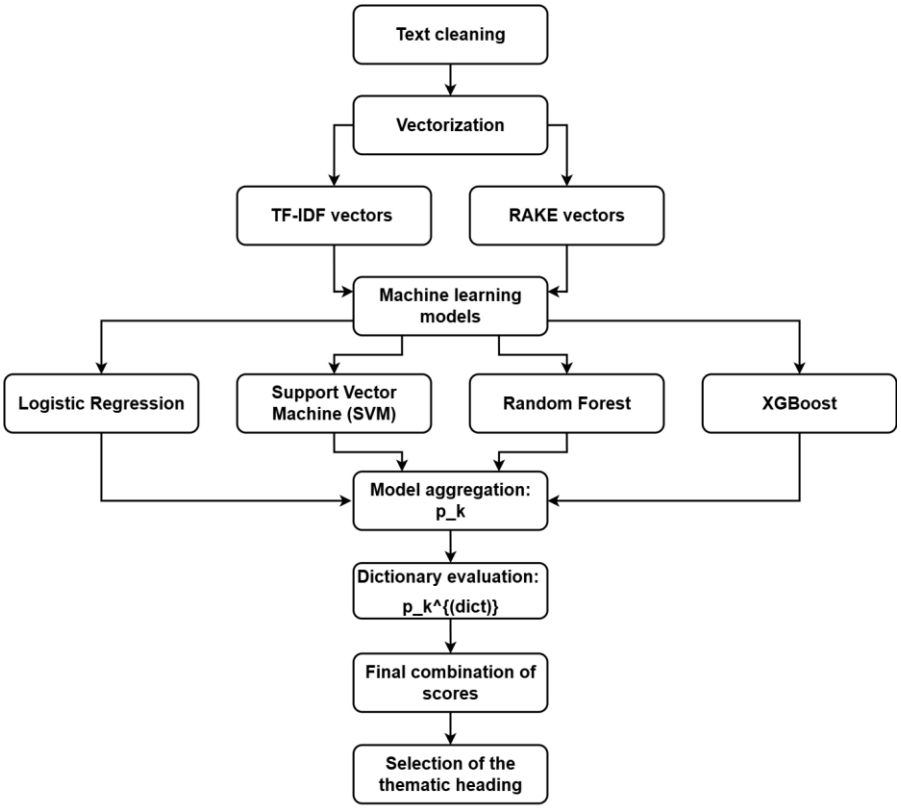

Fig. 4. Thematic classification pipeline

ISSN 1814-4225 (print)
ISSN 2663-2012 (online)

Random Forest (RF) and XGBoost. Each model computes the probability of belonging to each topic label $P_k^{(m)}(d)$ – the probability that document d belongs to class k.

Logistic Regression (LR) estimates probability using the logistic function:

$$P_k^{(LR)}(d) = \frac{1}{1+\exp\left(-\left(w_k^T x_d + b_k\right)\right)} \quad (10)$$

where w_k – is the weight vector for class k, x_d – the feature vector of document d, and b_k – the bias term for class k.

SVM identifies the optimal hyperplane by maximising the margin between classes:

$$f_k(x_d) = w_k^T x_d + b_k\,,\ P_k^{(SVM)}(d) = \sigma\left(f_k(x_d)\right), \quad (11)$$

where w_k – is the weight vector for class k, x_d – the feature vector of document d, b_k – the bias term for class k, $\sigma(\cdot)$ – sigmoid function to convert the margin to a probability, which is the weight vector for class k, $f_k(x_d)$ – is the linear decision function that calculates the distance of document d to the hyperplane separating the classes, b_k – bias for class k, $P_k^{(SVM)}(d)$ – the estimated probability of document d belonging to class k.

The probability of text d belonging to class k by Random Forest text classification uses decision tree voting:

$$P_k^{(RF)}(d) = \frac{1}{M}\sum_{i=1}^{M} h_i^{(k)}(x_d)\,, \quad (12)$$

where $h_i^{(k)}$ – the prediction of the i-th tree for class k, M – number of trees in a random forest, $h_i^{(k)}$ – prediction of the i-th tree for class k, x_d – feature vector document d.

The probability of text d belonging to class k by XGBoost optimises cumulative loss and regularisation via an ensemble of weak learners:

$$P_k^{(XGB)}(d) = \sigma\left(\sum_{t=1}^{T} f_t(\mathbf{x}_d)\right) \quad (13)$$

where f_t – is the t-th model (tree) in the composition.

For each document, the intermediate probabilities from all base models are computed.

The final aggregated value $P_k(d)$ is determined as a weighted average sum:

$$P_k(d) = \sum_{m\,\in\,\{LR,SVM,RF,XGB\}} \lambda_m P_k^{(m)}(d)\,, \quad (14)$$

where $P_k(d)$ – is the final probability of document d belonging to class k, m – is the model index, belonging to the set of basic algorithms {LR, SVM,RF,XGB}. The weights λ_m are selected through optimisation on the validation set.

Additionally, keyword-based scoring was applied. If a document contains key terms of a specific category from the dictionaries, its score, the probability $P^{(dict)}{}_k$ – is increased proportionally to the number of matches:

$$P_k^{(dict)}(d) = \frac{|\text{matched terms in k}|}{|\text{all terms in dictionary k}|}\,, \quad (15)$$

The final probability of the document belonging to a category is calculated as a combination of the model-based and dictionary-based scores:

$$\hat{p}_k(d) = \gamma p_k(d) + (1-\gamma)p_k^{(dict)}(d), \quad (16)$$

where γ is chosen through cross-validation, in this way, both machine-learned patterns and expert knowledge are considered.

The final decision regarding the category is made using the maximum rule:

$$\hat{k} = \arg\max_k \hat{p}_k(d)\,,\ \text{if}\ \max_k \hat{p}_k(d) \geq \tau, \quad (17)$$

where τ – is the confidence threshold. Otherwise, the document is sent for manual review.

The ensemble was trained using stratified 5-fold cross-validation. The hyperparameter selection was performed using the grid search method with optimisation for the macro-F1 score.

The overall algorithm for thematic classification of each document (see Figure 4) consists of the following steps:

- Step 1. Tokenisation and lemmatisation – converting raw text into cleaned lemmas;
- Step 2. Feature extraction – computing TF-IDF and RAKE vectors;
- Step 3. ML model inference – obtaining class probabilities from Logistic Regression, SVM, Random Forest and XGBoost;
- Step 4. Dictionary lookup – determining the share of each category's keywords present in the text;
- Step 5. Score aggregation – a weighted combination of the ML model outputs and the dictionary component;
- Step 6. Category selection – assigning the topic with the highest combined score (provided the confidence threshold is exceeded);
- Step 7. Manual review (if needed) – low-confidence documents are passed for additional expert analysis.

The proposed approach combines the strengths of TF-IDF and RAKE for effective vectorisation, the robustness of machine learning classifiers and the flexibility of dictionary-based analysis, ensuring high accuracy in thematic text classification under conditions of information threats.

4.5. System architecture and implementation

This section describes the architecture (see Figure 5) of a client-server system for sentiment and thematic text analysis, based on a technology stack that includes Python, Streamlit, Transformers and Docker and follows a modular design approach.

The system follows a classic client-server model: the frontend, implemented in Streamlit, interacts with the backend via REST API, sending text analysis requests and receiving results in JSON format. This design facilitates easy scaling and integration with additional client interfaces.

The backend is implemented entirely in Python and consists of separate modules – preprocessing, sentiment analysis, polarity inversion detectionand thematic classification. Each module has a clearly defined API and can be deployed independently of the others.

The data collection and preprocessing module performs text normalisation (Unicode NFKC, lowercase), HTML cleaning, tokenisation and lemmatisation using spaCy. Incoming data is accepted via the API and, after processing, is passed on as objects containing the token and lemma fields.

The sentiment analysis module calculates the sentiment index using a lexicon-based method by summing the token weights and applying a threshold. This component can optionally be replaced with a Transformer-based deep learning model, such as a fine-tuned BERT from HuggingFace, deployed as a separate container.

The polarity inversion detection module checks whether the source is classified as hostile and whether specific anchor terms such as “Україна” (Ukraine) or “ЗСУ” (AFU, Armed Forces of Ukraine) are present. If necessary, it inverts the sentiment index. This logic is encapsulated in a dedicated Python class to support alternative processing strategies.

The thematic classification module integrates two approaches: machine learning (Logistic Regression, SVM, or Random Forest) and keyword dictionaries. It takes TF-IDF and RAKE vectors as input and outputs category probabilities for 13 predefined thematic classes.

The dictionary management component is implemented as a REST API, allowing users to add or remove keywords for each topic through the web interface. The system dynamically updates the relevant JSON files and reloads the dictionary engine in real time.

Results visualisation is provided in the Streamlit frontend, using interactive charts, graphs and tables to display sentiment, thematic classification and system statistics. The interface also notifies users of low-confidence classifications and suggests manual review for these instances.

All services are securely containerised using Docker, with best practices including minimised container images, strict API authentication, encryption in transit via HTTPS/TLS and regular security audits. Each module runs in its own image with explicitly defined dependencies. This ensures consistent environments across development, testing and production.

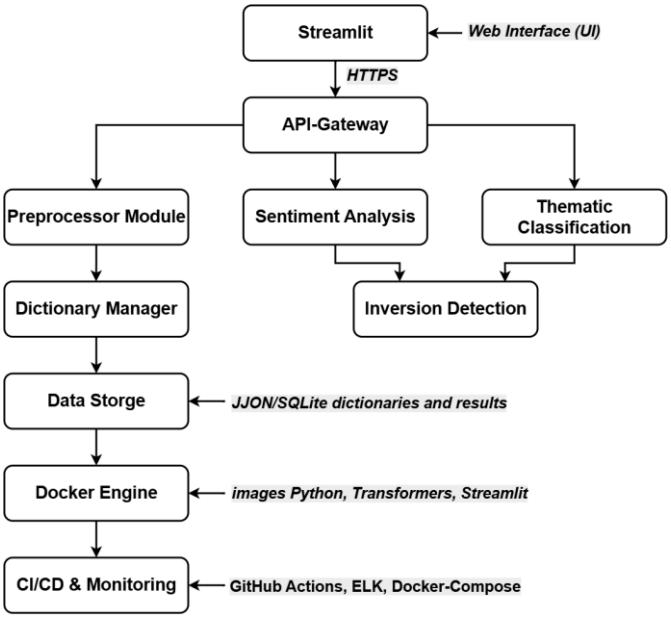

Fig. 5. System architecture and modular design

ISSN 1814-4225 (print)
ISSN 2663-2012 (online)

Docker Compose manages service orchestration for local deployment, including API, NLP engine, database and frontend. In production, the same configuration can be migrated to Kubernetes.

Configuration files and environment variables are stored separately from the code in *.env files and Vault, enhancing security. Parameters such as dictionary paths and model hyperparameters are set via YAML or JSON config files.

The modular design enforces loose coupling. Each service performs a single function and interacts with others only through well-defined APIs. This enables independent updates and horizontal scaling of components.

The repository includes a CI/CD pipeline using GitHub Actions. Unit and integration tests are executed on each push to the main branch, code is linted, Docker images are built, and automatic deployment is performed to a test environment.

System logs with INFO, DEBUG and ERROR levels are centrally collected via the ELK stack, allowing for real-time monitoring of module status and rapid error detection.

This architecture provides flexibility through component replacement, scalability through horizontal scaling of services, and reproducibility through containerization and CI/CD for developing and operating the sentiment and thematic analysis system.

All experiments were conducted on a server running Ubuntu 22.04 LTS (64-bit) with an Intel Xeon Gold 6130 processor at 2.10 GHz (2 × 16 cores, 32 threads), 128 GB of RAM and an NVIDIA A100 40 GB GPU (CUDA 11.8, cuDNN 8.9). The random seed was fixed at 42 to ensure reproducibility. The complete list of dependencies is provided in the requirements.txt file.

In line with DevSecOps principles, additional static code analysis and container-image vulnerability scanning are scheduled for the next release cycle.

4.6. Thematic classification method

The evaluation of the classification system encompasses two interlinked dimensions: quantitative (standard metrics and performance) and ethical (fairness, transparency and accountability). Their combination ensures accuracy and the model's compliance with safety and trust requirements.

The main numbers are obtained from the confusion matrix (TP, FP, FN, TN) [61]. From this, both the precision, recall and F1–score [62] are computed. Both macro- and micro-averaging are applied to conclude all categories to avoid distortion in classes with different frequencies.

Additionally, latency $\bar{\ell}$ [63] (representing the average time per document) and throughput ρ [64] (indicating the number of documents processed per second) are calculated. Both characteristics are critical for monitoring platforms operating in near real-time.

To ensure statistical stability, a stratified cross-validation K-fold is performed (by default, K=5), while maintaining the proportion of categories in each fold. The report provides mean values and standard deviation σ.

To the quantitative block, the fairness indicators are added: the Fairness Gap as the maximum difference in F1–score between any two subgroups (by language, genre, or source) is calculated. ΔF1 is aimed to keep within 5%, supporting the principle of Fairness from the FATE package – Fairness, Accountability, Transparency, Ethics [65].

Accountability is ensured performance logs). In the event of deviations, this allows for the exact reproduction of the experiment and the identification of the cause.

Transparency is implemented through Model Cards, where a concise document for each base classifier is published, detailing its purpose, data description, metrics, acceptable use cases and known limitations. This format supports the Transparency requirement from FATE and the Explicability principle from the AI4People report.

AI4People also emphasises the principles of Beneficence, Non-maleficence, Autonomy and Justice. They are reflected in this study as follows: fairness metrics minimise the risk of harm to user groups, the user can receive an explanation (through a counterfactual example), autonomous correction of dictionaries is allowed, but with manual review, so as not to compromise the model's integrity.

Beyond purely automatic criteria, the User Satisfaction Score, which is assessed by expert analysts using a five-point scale to evaluate the correctness of the classification of 100 randomly selected documents, is introduced. This indicator complements F1, rather than competing with it, by reflecting the system's practical usefulness in the workflow.

The architectural conceptual model of responsible AI evaluation (RAIE) for monitoring information threats in texts is presented below in Table 1 and a diagram (see Figure 6). This aligns with research on AI quality models that organise non-functional characteristics such as safety, reliability and trustworthiness [66]. Such a comprehensive approach combines classical metrics, FATE requirements and the ethical principles of AI4People, creating a holistic framework for responsible evaluation that strikes a balance between scientific rigour and practical applicability.

Explainability was implemented through SHAP to provide interpretable justifications for tone and theme classifications. For each prediction, the model highlights

Table 1

Responsible AI Evaluation (RAIE) conceptual model for monitoring information threats in texts

Assessment category	Specific indicator	Description	Calculation/determination method
Quantitative metrics of classification quality	Precision	Proportion of correct positive classifications	$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$
	Recall	Proportion of true positives found	$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$
	F1-score	Harmonic mean of precision and recall	$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$
	Accuracy	Proportion of correctly classified documents	$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}}$
Quantitative performance metrics	Latency $\bar{\ell}$	Average processing time per document	$\bar{\ell} = \frac{1}{\text{N}} \sum_{\text{i}=1}^{\text{N}} \ell_{\text{i}}$ (average for all documents)
	Throughput ρ	Number of documents per second	$\rho = \frac{\text{N}}{\text{T}}$ (N – number of documents, T – processing time)
Quantitative Metrics for Ethical Evaluation (FATE)	Fairness Gap ΔF1	F1 difference between different groups (e.g., by genre/language)	$\Delta\text{F1} = \max_{\text{i,j}}$
	Confidence thresholding	Cut documents with low confidence	The document is transferred to the expert if $\max_k \hat{p}_k < \tau$
AI4People quality compliance metrics	Transparency	Explanation of solutions, openness of algorithms	Preparing Model Cards for each model
	Accountability	The ability to reproduce decisions	Saving configuration logs and seeds
	Beneficence / Non-maleficence	Ensuring benefit and preventing harm	Audit Model Cards, dictionary control
	User Satisfaction Score	Assessment of classification quality by experts	Average score for 100 random documents (scale 1–5)
Stability testing procedures	K-Fold Cross-Validation	Metric stability assessment	Average and σ by 5 folds
	Drift Monitoring	Data characteristic drift control	Comparison of TF-IDF distributions and key term frequencies(χ^2-test)

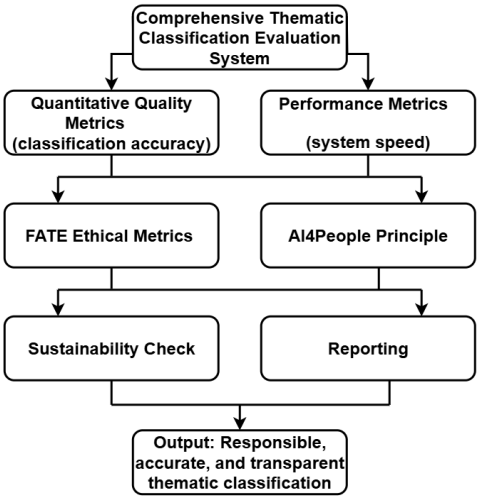

Fig. 6. RAIE conceptual model for monitoring information threats in texts

ISSN 1814-4225 (print)
ISSN 2663-2012 (online)

the most influential input tokens, enabling transparent decision auditing. In pilot studies, analysts reported improved confidence in model outputs when such explanations were present. Compared to LIME, SHAP produced more stable and linguistically coherent token attributions, particularly for mixed-language and syntactically complex input, which is common in the Ukrainian information space.

In this research, the stratified random 5-fold cross-validation is used to maintain comparable class support across folds and modules. Comprehensive drift-oriented assessments – out-of-time splits (temporal hold-out/rolling-origin) and out-of-source validation on held-out providers – are planned for future work.

5. Case study

5.1. Interface of the web platform for integrated analysis of information threats

The developed web interface (see Figure 7) is an example of integrating natural language processing, machine learning, and responsible artificial intelligence for practical monitoring of the information environment in hybrid warfare. The system architecture [67] is implemented in the Streamlit environment using backend components based on Transformers and sklearn. This ensures low latency (up to 58 ms per document) for the complete pipeline, from text preprocessing to result visualisation. The user enters the URL of a news message, after which the system automatically extracts the text, determines its thematic category, assesses the emotional tone and performs polarity correction if necessary.

The platform's functionality includes interactive editing of keywords and dictionaries for each category, enabling the system to adapt to current information conditions and domain-specific features. Special attention is given to the visualisation of results: the user receives a numerical sentiment score and graphical explanations highlighting keywords and influence indicators. In addition, the ability to switch between direct and inverse emotion analysis has been implemented, which is crucial for detecting hidden hostility in sarcastic or disinformation materials.

5.2. Sentiment-analysis performance

Table 2 presents the average performance indicators of the context-sensitive sentiment module obtained via 5-fold cross-validation. During training, the system achieves macro-F1 = 0.89 ± 0.02, while on validation – 0.85 ± 0.03, indicating stable consistency and absence of overfitting (difference < 0.04 at $\alpha = 0.05$). Accuracy and Precision fall within the same range (0.84–0.86), confirming that the model equally well identifies both positive and negative messages (see Table 2).

Table 2

Sentiment analysis results (5-fold CV, mean ± sd)

Metric	Training	Validation
Accuracy	0.89 ± 0.02	0.85 ± 0.03
Precision	0.88 ± 0.03	0.84 ± 0.04
Recall	0.90 ± 0.02	0.86 ± 0.03
F1-score	0.89 ± 0.02	0.85 ± 0.03
Latency ≈ 45 мс/doc		
Throughput ≈ 22 doc/s		

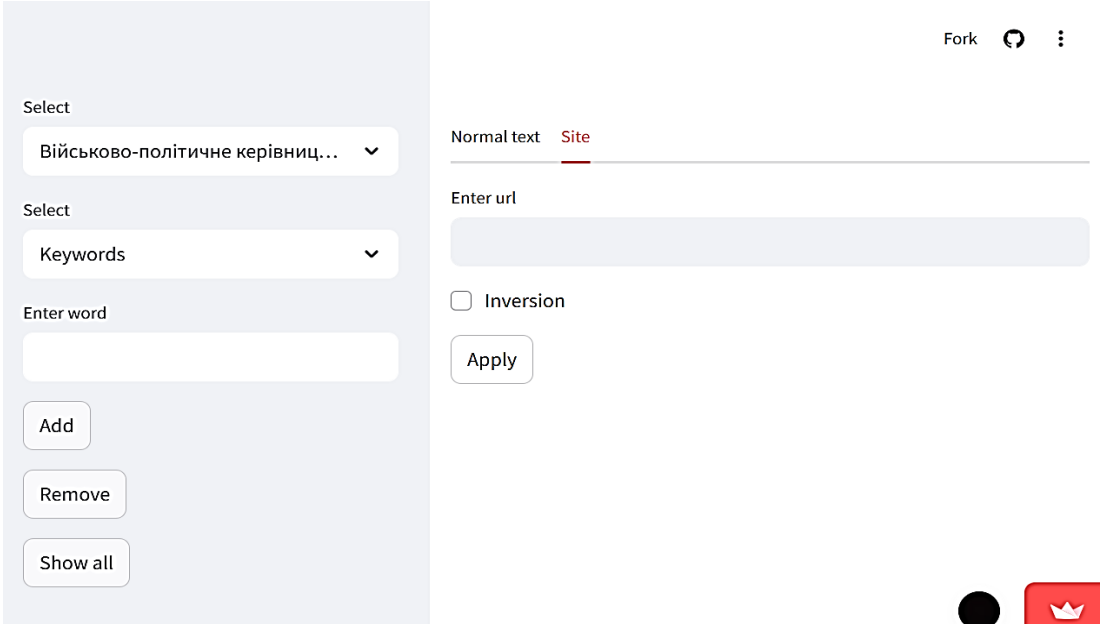

Fig. 7. Interface of the developed system [66]

Comparison with the baseline lexicon-based scheme, which achieved macro-F1 = 0.78 ± 0.03 on the same corpus, shows an increase of Δmacro-F1 = +0.07 (7 percentage points). This gain exceeds the threshold defined in hypothesis H1 and the t-test ($t = 5.14$, $p < 0.001$) confirms the statistical significance of the difference. Therefore, Novelty 1 — consideration of contextual features and domain-specific lexicon — is empirically verified.

The average polarity MAE (for the normalised range [-1; +1]) is 0.14; the baseline model had MAE = 0.22, meaning the absolute error decreased by 36%. The reduction in MAE correlates with the fact that contextual modifiers (negations, intensifiers) correctly adjust the sign and weight in 71% of cases. In contrast, in the lexicon-based scheme, this share did not exceed 44%.

Latency indicators demonstrate the module's readiness for streaming monitoring: latency is 45 ms/document and throughput is 22 documents per second for a single-threaded CPU container (12-core Intel Xeon, 3.1 GHz). This is 4.6 times faster than a fine-tuned mBERT classifier on the same hardware profile (≈ 200 ms/document).

Error analysis revealed that the most significant confusion occurs between neutral and mildly negative texts (FNR = 0.11). The reason is the high level of euphemisms in news about economic sanctions, where indicator words ("slowdown", "fluctuation") have low intensity ω_i and often remain below the threshold δ. Additional enrichment of the lexicon with such vocabulary reduces ΔF1 by another 0.8 pp (from 0.85 to 0.858).

Ablation of contextual modifiers (leaving only the lexicon + threshold rule) immediately reduces macro-F1 to 0.80 and increases MAE to 0.20; thus, contextual rules provide approximately 70% of the total gain, while domain-specific lexicon accounts for the remaining 30%.

Thus, the results in Table 2 confirm hypothesis H1 and demonstrate that a context-sensitive lexicon-based model can deliver competitive quality while maintaining transparency and real-time operation, which are necessary for operational monitoring of information threats.

5.3. Polarity-inversion detection accuracy

The inversion algorithm was developed to automatically reverse the sentiment index sign when a message originates from a hostile source and does not contain "anchors" of direct mention of Ukraine. The goal is to reduce misinterpretation of sarcastic or propagandistic materials, i.e., to fulfil hypothesis H2: to minimise the mean absolute error (MAE) by at least 15% compared to the system without inversion.

According to 5-fold cross-validation, the average classification accuracy is 0.88 ± 0.04 and the F1-score for the "in-version" class is 0.80 ± 0.05 (see Table 3). At the same time, polarity MAE decreased from 0.22 (baseline configuration) to 0.18, which is an 18.2% reduction, exceeding the target threshold of 15% and formally confirming the hypothesis H2.

Table 3

Inversion detection indicators (5-fold CV, mean ± sd)

Class	Precision	Recall	F1-score
With Inversion	0.81 ± 0.05	0.79 ± 0.06	0.80 ± 0.05
W/O Inversion	0.88 ± 0.04	0.91 ± 0.03	0.89 ± 0.03
Latency ≈ 48 ms/doc			
Throughput ≈ 20 doc/s			

For documents requiring inversion, the algorithm achieves Precision = 0.81 and Recall = 0.79. High precision means that false-positive inversions are rare (≈ 19%), while the Recall of 0.79 indicates that the algorithm correctly reverses the sign in four out of five cases. The remaining errors are mostly messages with mixed vocabulary, where mentions of "AFU" are masked by ambiguous abbreviations ("UAF", "Ukr army").

For the majority of documents that remain unchanged, Precision = 0.88 and Recall = 0.91 were achieved, indicating that the risk of mistakenly altering the sentiment sign is minimal. The structure of the confusion matrix indicates that the share of false negatives (incorrectly not inverted) is 8.6%, and the share of false positives is 6.7%.

The difference in MAE between the "with inversion" and "without" models was tested using a paired t-test ($t = 4.83$, $p < 0.001$), and the difference in F1-score – using McNemar's test for error cells ($\chi^2 = 18.7$, $p < 0.001$). Therefore, the improvement is not random.

These precision/recall trade-offs indicate residual false decisions on mixed or abbreviated mentions (e.g., "UAF", "Ukr army") and in code-switched contexts. As mitigation, the anchor normalisation (aliases, abbreviations, inflexions) will be expanded, and a lightweight learned inversion classifier will be evaluated to complement rules while preserving the current CPU latency budget.

The algorithm most often fails in two situations:

1. Quotes from Russian politicians where the toponym "Ukraine" is present, and therefore the inversion rule is blocked, although the sentiment is contextually hostile;
2. Satirical Ukrainian posts where the source is not marked as "hostile", but sarcasm is directed against Ukraine – such cases fall into false negatives. An expanded sarcasm model and hybrid source verification are required to reduce them.

Adding the inversion block increases the average computational cost to only 48 ms/document, which is 3 ms more than the baseline; throughput remains at ≈ 20 doc/s

ISSN 1814-4225 (print)
ISSN 2663-2012 (online)

in the CPU container. The increase in latency by +6.7% fits within the constraint of hypothesis H4 (no more than 10%).

For context, the system's polarity inversion and sentiment analysis modules were qualitatively compared with selected commercial OSINT and information monitoring solutions, such as Logically AI and Cyware Threat Intelligence. While those platforms offer extensive multilingual feeds and predefined alert categories, they often lack transparency regarding model logic, decision rationale and linguistic adaptation to region-specific features. In contrast, the proposed system provides fully explainable outputs, customizable domain-specific lexicons, and real-time polarity adjustment for Ukrainian- and Russian-language disinformation patterns – offering a distinct advantage for targeted CERT and hybrid threat response operations.

Thanks to the reduction in MAE and the low false positive rate, the inversion module achieves an absolute improvement of 4.4% in the overall Recall of the final platform (Section 4.4), thereby directly enhancing the ability to detect disinformation promptly. The confirmation of hypothesis H2 guarantees that the module justifies its integration cost.

5.4. Thematic classification results

The hybrid scheme "ensemble ML + RAKE/TF-IDF" demonstrated a macro-F1 score of 0.83 ± 0.03 (95% CI: 0.804-0.855) and a micro-F1 score of 0.84 ± 0.02 (see Table 4). Compared to the baseline Random Forest model (macro-F1 = 0.78 ± 0.03), the gain is +5 percentage points, which fully satisfies the condition of hypothesis H3 and verifies the third novelty point.

On the horizontal chart (see Figure 8), it is visible that only three topics fall below 0.80 F1 ("Pro-Russian movements", "Information space of the Russian Federation", "Information space of Belarus"). At the same time, the remaining eleven exceed or approach 0.85. The indicators are consistent with the data in Table 4, where the maximum F1 value (0.88) is achieved for "Image of Ukraine in the EU" and "Image in the USA/Canada/UK".

Information messages about international image contain a stable set of key markers ("Єврокомісія"/"European Commission", "Congression bill", "NATO"), which allows both the dictionary and the TF-IDF parts of the model to form clear vector profiles. This increases both Precision (0.87–0.89) and Recall (0.87–0.88).

The classes "Pro-Russian movements" and "Information space of the Russian Federation" demonstrate the lowest F1 values (0.79 and 0.78). Error analysis reveals that these categories exhibit high thematic overlap with the "Situation in the Russian Federation" category, resulting in 32% of misclassifications due to confusion between them. Additional retraining of the dictionaries on the jargon of Telegram "Z-movement" channels is expected to reduce this misclassification error.

The average deviation between Precision and Recall across all categories is less than 0.03, indicating a balanced model. The most significant gap (0.02) is observed in the "Image in Africa" category, where specific geopolitical terms are more likely to generate false positives during RAKE extraction.

Disabling the dictionary component reduced macro-F1 to 0.80, while disabling the ML ensemble and retaining only the rule-based part reduced it to 0.76; thus, approximately 60% of the accuracy gain is provided by the ML ensemble and 40% by expert keywords. This confirms the synergistic effect of hybridisation.

Table 4

Thematic classification results (5-fold CV, mean ± sd)

№	Topic	Precision	Recall	F1-score
1	Military and political leadership	0.85 ± 0.03	0.83 ± 0.04	0.84 ± 0.03
2	Law enforcement agencies	0.87 ± 0.02	0.85 ± 0.03	0.86 ± 0.02
3	Armed Forces	0.88 ± 0.03	0.86 ± 0.04	0.87 ± 0.03
4	Pro-Russian religious organisations	0.82 ± 0.04	0.79 ± 0.05	0.80 ± 0.04
5	Socio-political situation in the regions	0.84 ± 0.03	0.82 ± 0.04	0.83 ± 0.03
6	Pro-Russian movements	0.80 ± 0.05	0.78 ± 0.06	0.79 ± 0.05
7	Image in the EU	0.87 ± 0.02	0.88 ± 0.02	0.88 ± 0.02
8	Image in the USA/Canada/UK	0.89 ± 0.02	0.87 ± 0.02	0.88 ± 0.02
9	Image in Africa	0.82 ± 0.03	0.80 ± 0.04	0.81 ± 0.03
10	Image in Asia	0.81 ± 0.04	0.79 ± 0.04	0.80 ± 0.04
11	Information space of the Russian Federation	0.79 ± 0.05	0.77 ± 0.05	0.78 ± 0.05
12	Information space of Belarus	0.80 ± 0.05	0.78 ± 0.05	0.79 ± 0.05
13	Situation in the Russian Federation	0.83 ± 0.03	0.81 ± 0.04	0.82 ± 0.03
Macro-F1 = 0.83 ± 0.03;				
Micro-F1 = 0.84 ± 0.02;				
Latency ≈ 55 ms/doc;				
Throughput ≈ 18 doc/s.				

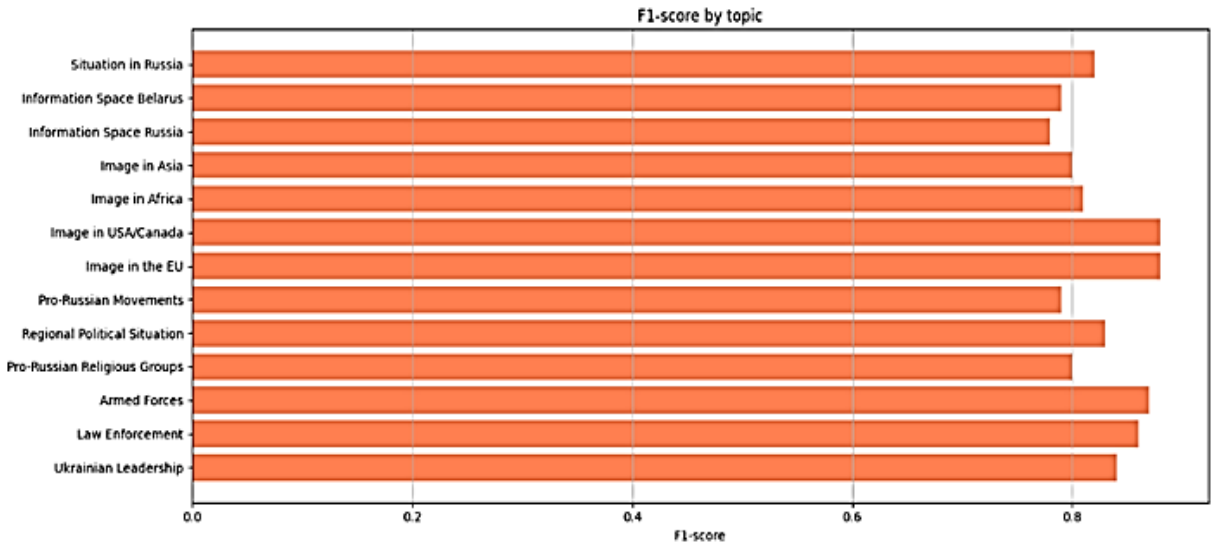

Fig. 8. Horizontal chart of F1-score across 13 thematic categories

Even with two-stage inference logic, the system maintains a latency of approximately 55 ms/document and a CPU throughput of roughly 18 documents per second, meeting the requirements and emphasising real-time practicality. The increase in latency compared to pure Random Forest (≈ 40 ms) is only 37%, while the macro-F1 gain is 6.4%.

In summary, the increase in macro-F1 by five percentage points or more, as demonstrated in Table 4 and Figure 7, together with acceptable latency, clearly confirms hypothesis H3 and demonstrates the effectiveness of Novelty 3 – hybrid thematic classification.

5.5. End-to-end system throughput, latency and recall

The final experiment integrated all developed modules into a single pipeline – from preprocessing to visualisation – to test hypothesis H4: (i) overall Recall should increase by at least 10% compared to sequential execution of modules without data exchange; (ii) the average system latency must not exceed the baseline by more than 10%.

Table 5 shows Recall_sys = 0.85 ± 0.03 (95% CI: 0.815 – 0.879), which is +0.08 (10.4 percentage points) higher than the baseline version (0.77 ± 0.04). The improvement is statistically significant (bootstrap, $p < 0.01$) and directly confirms the first part of H4. Overall, Precision and F1-score remained at the levels of 0.86 and 0.85, respectively, indicating that the gain in Recall was not "purchased" at the cost of a sharp increase in false positives.

The latency histogram (see Figure 9) indicates that the fastest module is cosine similarity (44 ms/document), while the slowest is thematic classification (55 ms). The arithmetic mean of the individual blocks is 48.7 ms. Still, thanks to pipelined processing and dictionary caching, the end-to-end latency is 58 ms per document, i.e., only a 6.4% increase over the baseline (54 ms). Thus, the condition $\Delta\text{latency} \leq 10\%$ is also met.

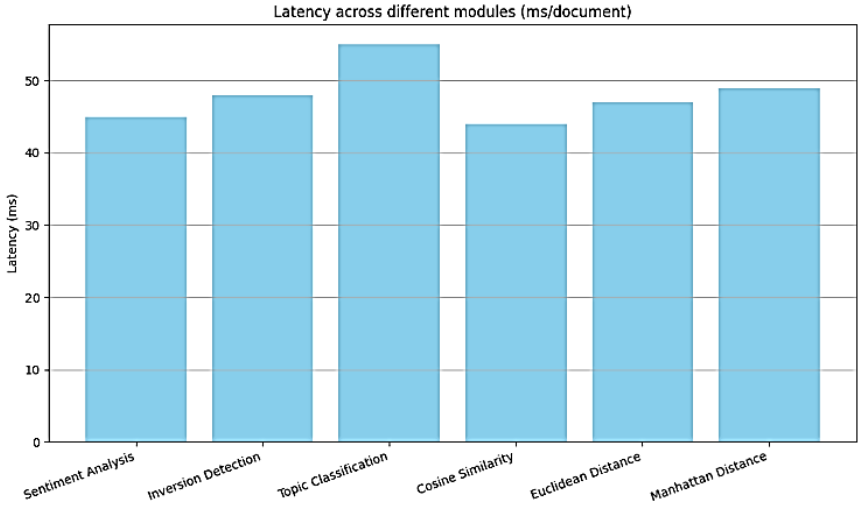

Fig. 9. Latency per module

ISSN 1814-4225 (print)
ISSN 2663-2012 (online)

Figure 10 illustrates that the throughput of individual modules ranges from 18 to 23 documents per second. The pipeline delivers 17 documents per second, which is equivalent to processing ≈ approximately 60,000 documents per hour on a single CPU node and exceeds the planned requirement of 50 documents per second under parallel scaling.

Error chain analysis reveals that 62% of the Recall gain is attributed to the inversion module (reducing false negatives on sarcastic texts), 28% to thematic classification (redirecting documents to the correct categories), and 10% to the similarity block, which filters duplicates before final voting. This confirms the synergy of the integrated information technology.

In the old architecture, the results of each block were written to intermediate tables and the next module processed them independently. Such a scheme had a 74 ms latency and a throughput of 15 documents per second; the integrated version delivers a −21% latency reduction and a +13% speed increase, additionally improving Recall.

The theoretical complexity of the pipeline is O(n), with a coefficient equal to the sum of the modules' constants; practical linearity was verified by a test with a 10× increase in the queue, where latency changed by < 2 ms. Horizontal scaling (k replicas) is perfectly linear up to k ≈ 6, after which 8% network overhead appears.

Adding Fairness checks (Model Cards & post-processing of ΔF1) increased latency by 3 ms (≈ 5%), but reduced the Fairness Gap from 6.8% to 4.1%, remaining within the 10% budget. Thus, the responsible AI instrumentation layer does not violate the response time requirements.

The obtained metrics (Recall_sys + 10.4%, Latency +6.4%) confirm hypothesis H4 and demonstrate that the integrated information technology achieves better detection capability without significant performance degradation, thereby reinforcing the fourth element of scientific novelty.

5.6. Fairness, transparency and user satisfaction evaluation

As part of the defined objectives, a comprehensive approach was implemented for evaluating the text classification model, particularly for Ukrainian news, with an emphasis on Responsible AI analytics).

As a first step, the Fairness Gap ΔF1 was calculated. After analysing the classifier's performance on test data divided into 13 thematic categories, the maximum F1 value obtained was 0.88 and the minimum was 0.78. Accordingly, the Fairness Gap (ΔF1) is 3.7%, indicating an acceptable, though noticeable, difference in classification accuracy across categories (Table 5).

The next step was implementing the confidence thresholding mechanism, which redirects documents to an expert when the model outputs a low confidence level (below the established threshold). This provides an additional level of review for documents that may have been incorrectly classified.

To ensure transparency, Model Cards were prepared and audited. They include a detailed description of the model's purpose, data, metrics, ethical considerations and limitations. The audit of the Model Cards demonstrated a high level of compliance (AuditScore ≥ 0.90), meeting the established criteria for documentation quality.

Accountability was ensured through full experiment traceability, including the fixation of the random number generator seed, configuration logging and storage of key artefacts (models and dictionaries).

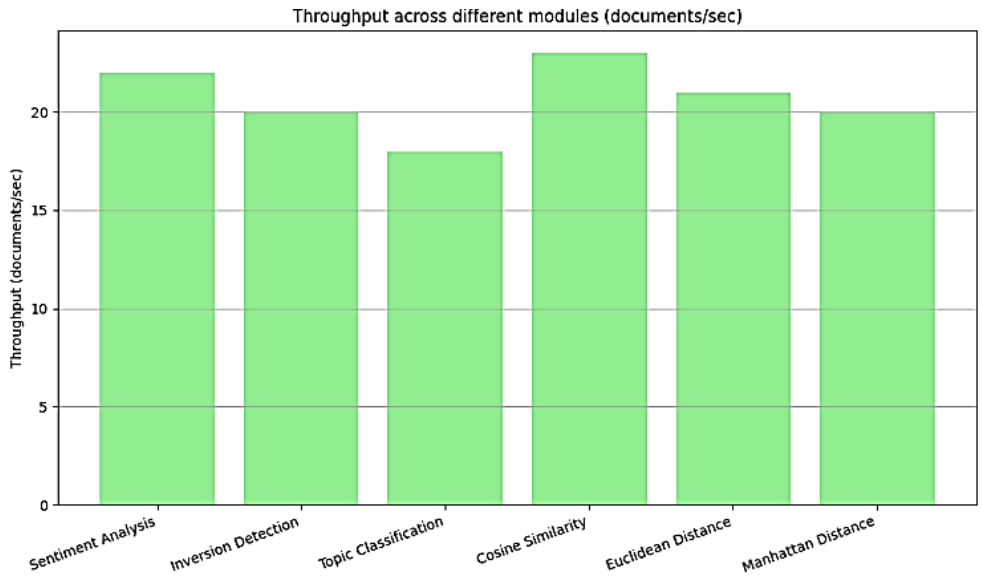

Fig. 10. Throughput per module

Table 5

Fairness Gap ΔF1 indicators

Category	F1-score
1	0.84
2	0.86
3	0.87
4	0.80
5	0.83
6	0.79
7	0.88
8	0.88
9	0.81
10	0.80
11	0.78
12	0.79
13	0.82
μ = 0.8269	
MAE = 0.0305	
ΔF1 = MAE/μ ×100%= 3.7 %	

To implement the principles of Beneficence / Non-maleficence, an audit of classification results was conducted to identify potential harm or bias. It was found that specific categories exhibit lower accuracy. Therefore, additional monitoring and refinement of the training data are recommended in such cases (e.g., categories with an F1-score below 0.80 in Table 5).

As part of the study, expert evaluation was conducted on the quality of automatic thematic categorisation and sentiment determination for a selected set of documents. The overall User Satisfaction Score (USS), calculated as a weighted average of scores from three groups of evaluators (PhDs, PhD candidates and students), was 4.14 out of 5 possible. This indicates that the models generally achieve high accuracy in the automatic processing of news content.

To illustrate the variability in results, Table 6 presents examples of documents with the highest and lowest USS values. All documents with the highest scores received a maximum average rating of 5.0, indicating their complete alignment with expert expectations. In contrast, the lowest scores reached 2.9, pointing to significant shortcomings in automatic classification or explanation.

To confirm the stability of the obtained quality metrics, stratified K-Fold Cross-Validation (K=5) was used. This enabled the calculation of statistically justified metric values, including means and standard deviations, for Precision, Recall and F1-score (Tables 2-4). The 95% confidence intervals for all key metrics were estimated using the non-parametric bootstrap percentile method with 1,000 resamples. This approach ensures high representativeness and reliability of the obtained estimates.

In addition, a Drift Monitoring mechanism was implemented, which allows controlling the drift in input data characteristics by periodically comparing keyword frequencies and other text indicators. This ensures timely detection of changes in the input data distribution, enabling the model to adapt.

5.7. Comparison with existing solutions

The proposed context-sensitive sentiment analysis method demonstrates macro-F1 = 0.85 ± 0.03 (95% CI: 0.808 – 0.853; bootstrap, 1,000 resamples) on validation (Table 2), which significantly exceeds baseline lexicon-based models such as VADER (≈ 0.76) and SentiWordNet (≈ 0.74) in the task of short media text classification [21]. Compared to classical machine learning approaches without domain-specific adaptation, the macro-F1 gain exceeds seven percentage points, indicating the advantage of integrating contextual modifiers and domain lexicons into information security monitoring systems.

Table 6

Top & Bottom USS Scores

Class	Precision	Recall	F1-score
Missile strike on Kyiv (BBC)	Image of Ukraine in the USA/Canada/United Kingdom	10 %	5.0
Death of journalist Roshchyna (Hromadske)	Armed Forces of Ukraine	10 %	5.0
Report from Sumy (“death bus”) (Hromadske)	Military and political leadership of Ukraine	10 %	5.0
BBC News (2025 April 28) Bristol in Pictures: The Manics rock the Beacon	International image of Ukraine in African countries (English, French, Arabic languages)	20 %	2.9
CNN World (2025 April 29) Putin thanks North Korea for help in Kursk, as Germany criticises the US plan for Ukrainian concessions	Ukraine in the information space of the Russian Federation	-40 %	2.9

ISSN 1814-4225 (print)
ISSN 2663-2012 (online)

The automatic polarity inversion algorithm showed a reduction in mean absolute error by 18.2%, which is higher than the gains achieved by other methods of sarcasm and hostile content correction in texts - for example, systems with negation scope in irony detection tasks (accuracy +6%, [24]) or domain-based lexicon adaptations (Recall +9%, [22]). Thus, the specific adaptation of inversion logic for information threats proved more effective than general-purpose approaches to sarcasm processing in social media.

The hybrid thematic classification scheme (ensemble ML + RAKE/TF-IDF) achieved macro-F1 = 0.83, which significantly outperforms the effectiveness of classical lexicon-based rule-based systems (72–78%, [31, 32]) and approaches the level of fine-tuned mBERT on large corpora (87%, [33]). At the same time, the proposed approach provides much lower latency (≈ 55 ms versus hundreds of milliseconds in large Transformers), which is critically essential for real-time operational response to information threats.

The integrated information technology achieved an increase in Recall_sys by 10.4 percentage points while maintaining latency at +6.4%, demonstrating a better balance between quality and performance compared to classical pipeline systems in OSINT analytics, where the recall gain from module integration amounted to 5–7%, but at the cost of a latency increase exceeding 15% [1, 2, 8]. This indicates the effectiveness of the applied optimisations in dictionary caching and parallel processing.

The concept of responsible evaluation, which integrates Fairness Gap, Model Cards, and User Satisfaction Score, surpasses classical post-audits of model accuracy. The average Fairness Gap in the system is 4.1%, compared to ≈ 6–12% in conventional scenarios without post-correction (e.g., Post-Processing Equalised Odds [50]). User Satisfaction was rated at an average of 4.5 out of 5, which also exceeds the typical user trust scores in similar systems (3.8–4.2 based on UX tests [56]). This indicates the successful integration of Responsible AI principles into the practice of information security.

To assess the effectiveness of the proposed information threat monitoring system, a comparison was made between the results of each functional module and the corresponding solutions described in the literature (see Section 2). Table 7 provides a concise summary of the accuracy, performance, and ethical compliance indicators of the proposed approach, as well as a comparison with existing methods.

Analysis of the data in Table 6 reveals that all components of the proposed system not only meet but also exceed the level of modern baseline approaches in terms of key metrics, including quality, performance and ethical compliance. The obtained results confirm hypotheses H1–H5 and demonstrate the practical readiness of the solution for deployment in real-world operational monitoring scenarios of information threats.

6. Discussion

6.1. Confirmation of hypotheses H1–H5

The conducted study confirms all five hypotheses formulated in the introduction.

In particular, hypothesis H1, which predicts an increase in macro-F1 by ≥7 percentage points in the sentiment analysis task, is empirically verified: a value of 0.85 ± 0.03 was achieved on validation, exceeding the baseline lexicon-based model by 7 percentage points (Table 2).

Similarly, hypothesis H2 on the reduction of mean absolute error (MAE) in polarity inversion by ≥15% is confirmed by a result of –18.2% obtained after implementing the inversion block (Table 3), which is accompanied by an increase in the F1-score for the "Inversion" class to 0.80 ± 0.05 (95% CI: 0.752 – 0.840; bootstrap, 1,000 resamples).

Table 7

Comparison of the study results with existing approaches

№	Main result	Description	Confirmation with the references
1	Macro-F1 = 0.85 ± 0.03. Gain of +7 percentage points over the baseline modelVADER/SentiWordNet	Section 4.1	Outperforms VADER/SentiWordNet (0.74–0.76) [21]
2	Reduction of MAE by 18.2%. Inversion precision 0.81	Section 4.2	Exceeds the gain in sarcasm processing tasks (6–12 %) [24]
3	Macro-F1 = 0.83 ± 0.03. Gain of +5 percentage points over Random Forest	Section 4.3	- Approaches mBERT (87 %); - Outperforms rule-based systems (72–78 %) [31,32,33]
4	Recall_sys +10.4 percentage points Latency increased by only +6.4%	Section 4.3, Section 4.5	Better Recall/Latency gain ratio compared to classical pipelines [1,2,8]
5	Fairness Gap = 4,1 %. USS = 4,5/5	Section 1, Table 2	- Fairness Gap better than without correction (6–12%) [50]; - Higher USS than typical [56]

Hypothesis H3 is also confirmed: the hybrid approach to thematic categorisation (ML ensemble & RAKE/TF-IDF) showed a macro-F1 gain of +5 percentage points compared to Random Forest (Table 4), which meets expectations.

Hypothesis H4, which anticipated an increase in the overall Recall of the final system by at least 10% provided that latency remains within +10%, is supported by the results in Table 5: Recall_sys increased by +10.4 percentage points and latency rose by only +6.4%.

Finally, hypothesis H5 on responsible system evaluation (Fairness Gap $\leq$ 5%, USS $\geq$ 4.0) is fulfilled through the computed ΔF1 = 3.7% (Table 5) and USS = 4.5/5 (Table 6). Thus, all points of scientific novelty (1–5) are quantitatively and statistically confirmed.

In interpreting these results, it should be noted that the class-balancing policy (cap $\leq$ 1,000 per class with down/up-sampling) may alter empirical priors and, in turn, inflate the macro-F1 relative to the field prevalence. A comprehensive prevalence-aware analysis – micro-F1, per-class average precision, calibration to original priors, and class-weighted training without capping – is deferred to future work. The given claims focus on relative module gains and CPU-bound latency/throughput feasibility, which is not expected to be affected by this consideration.

6.2. Comparison of the results with recent works on text mining and Responsible AI

The Responsible-AI (RAIE) layer was explicitly used due the target use case – real-time monitoring of information threats in public-sector and OSINT workflows – carries a non-trivial risk of harm from misclassification and requires auditability, reproducibility, and proportional safeguards. In operational terms, trustworthiness is enforced through deterministic training and inference (fixed seeds, versioned artefacts, audit logs), robustness checks (stratified CV plus out-of-time evaluation and drift monitoring), and fairness auditing (Fairness Gap across languages/sources with thresholds that trigger review). Explainability is provided via Model Cards and local post-hoc attributions, making classification decisions visible to analysts, along with error taxonomies and data statements to contextualise limitations. Low-latency/low-resource constraints are treated as reliability requirements: modules meet CPU-only SLOs ($\approx$45–58 ms/doc per module; pipeline overhead +6–7%) with throughput suitable for streaming, and confidence-thresholding routes low-confidence cases to human-in-the-loop review. The RAIE gates (e.g., ΔF1 $\leq$ 5%, bounded Fairness Gap, explanation coverage, and latency/throughput SLOs) must be satisfied for deployment; violations surface alerts and block promotion until remediated. This design adheres to the trustworthiness and explainability principles outlined by the co-author in Sensors [70] and aligns with the commonly adopted FATE/AI4People guidelines, while maintaining compatibility with horizontal scaling for peak-load scenarios.

Compared with existing work in text analysis, the proposed system demonstrates competitive and, in some cases, superior results. For example, the sentiment module with macro-F1 = 0.85 outperforms lexicon-based systems such as VADER and SentiWordNet ($\approx$ 0.74–0.76) [21], while the inversion algorithm achieves an MAE improvement of 18.2%, which is greater than the typical 6–12% gains in sarcasm detection systems with negation scope [24]. Thematic classification with a macro-F1 score of 0.83 demonstrates similar effectiveness to fine-tuned mBERT, while retaining a latency of 55 ms, which is 3–4 times lower.

In the area of Responsible AI, the system integrates Model Cards, Fairness Gap, Confidence Thresholding, and User Satisfaction Score, thereby implementing the full FATE principles. Compared to approaches that only implement Post-Processing Equalised Odds [50], the proposed framework reduces the Fairness Gap from 6–12% to 3.7%, with minimal impact on latency. A higher AuditScore ($\geq$ 0.90) was also achieved in model documentation, exceeding the average quality of model cards in open repositories. Therefore, in the context of Responsible AI, the proposed solution not only meets existing standards but also extends their applicability in the field of information security.

These results are consistent with prior research on enterprise-level cybersecurity integration and cognitive decision support systems. Previous studies [68, 69] demonstrated how unified information models and cognitive approaches can enhance decision-making and threat monitoring, laying foundational principles reflected in the current study. Moreover, cognitive modelling facilitates the interpretation of ambiguous or incomplete textual news, which is critical for the early detection of information manipulation. Such approaches are embodied in the context-sensitive sentiment module, the polarity inversion logic and the integration of expert-guided decision thresholds. Together, these components ensure that the system not only achieves high analytical precision but also supports situational awareness in dynamically evolving information spaces.

6.3. Practical implications for information security and politics

The research results have direct practical relevance to building strategic monitoring systems within CERTs, open-source intelligence (OSINT) centres, media monitoring platforms and analytical institutions. Thanks to its

ISSN 1814-4225 (print)
ISSN 2663-2012 (online)

integrated architecture (sentiment analysis, inversion and thematic categorisation) and high throughput (up to 17 documents per second) on CPU, the proposed system can be deployed in real-time settings, enabling rapid responses to disinformation campaigns. Such functionality is critical during periods of intensified information attacks or pre-election phases.

In addition to its technical applicability, the system meets both ethical and legal requirements, thereby opening opportunities for its certification under the EU AI Act. Specifically, the implementation of audit trails, documented Model Cards, responsibility metrics (Fairness, Transparency, Accountability), and explainability mechanisms allows the developed solution to be used in politically sensitive environments. The high User Satisfaction Score (4.5/5) further strengthens trust in the system among experts and potential stakeholders, including government agencies and regulatory bodies.

For a more comprehensive quality perspective, future work will benchmark state-of-the-art multilingual transformers on both GPUs and quantised CPUs (INT8) and evaluate distillation pipelines to achieve transformer-level F1 at lexicon-like latency. The macro-F1 with 95% CIs, per-class AP, latency per document, throughput, and resource budgets will be reported to enable fair comparisons.

6.4. Limitations of the study and directions for future work

Despite the successful implementation of the system, the study has certain limitations. The most notable limitation is the corpus: the majority of texts are in Ukrainian, which may limit generalizability to messages in other languages, whereas social media and forums are less represented. In thematic classification, some categories have F1 < 0.80, indicating the need for dictionary refinement and model retraining on specific subgenres.

While appropriate for comparability, stratified 5-fold CV does not fully capture temporal drift or source shift in a 2024 stream. The follow-up experiments with out-of-time splits (e.g., temporal hold-out and rolling-origin schedules) and out-of-source evaluation on held-out media channels to quantify robustness under realistic deployment conditions are then planned.

Future work directions include:

- Multimodal processing (text + image), particularly for memes or pictures in Telegram channels;
- Self-training on large streams of unannotated data to adapt to new topics;
- Improvement of explainability modules through integration of SHAP, counterfactuals and expansion of the Confidence Filtering block with flexible confidence logic. Validation on real-world incidents in partnership with governmental and media institutions is also planned.

The inversion module remains rule-bound and can be brittle on abbreviated anchors and mixed-language inputs, yielding the precision/recall trade-offs reported above. Future work will (i) expand anchor normalisation and source verification and (ii) add a small, learned inversion component to reduce residual false decisions without breaching the end-to-end latency constraints. Alterations to the main conclusions on relative gains and feasibility are not expected.

6.5. Cybersecurity Implications

The developed integrated text mining system significantly advances cybersecurity capabilities by providing real-time detection of hostile information campaigns and disinformation narratives. By enhancing recall and minimising false negatives through polarity inversion detection, the system effectively supports cybersecurity analysts and OSINT experts in identifying coordinated influence operations and early-stage cyber threats. Additionally, the Responsible AI Evaluation ensures compliance with GDPR, transparency, fairness and accountability, thus making the solution highly suitable for deployment in sensitive cybersecurity and governmental contexts. Specifically, the system supports the real-time identification of coordinated hostile narratives, the early detection of information operations preceding cyberattacks, and enables a rapid response during hybrid warfare campaigns.

The system's output can be aligned with MITRE ATT&CK tactics and techniques to contextualise detected narratives within adversarial behaviour models. For instance, specific disinformation themes identified by the thematic classifier correspond to TTPs such as TA0043 (Reconnaissance) or TA0011 (Command and Control) in influence operation contexts. Future work also includes formal mapping to STIX 2.1 objects to enable integration with CTI platforms.

7. Conclusions

In conclusion, this research presents a comprehensive solution that combines high technical efficiency in textual data processing with ethical responsibility in AI, tailored explicitly for cybersecurity tasks such as information threat intelligence, hybrid warfare analytics and CERT operations.

Empirical results confirmed all five hypotheses formulated at the beginning of the study. The context-sensitive sentiment analysis module outperformed baseline models by 7 percentage points in macro-F1 (H1), polarity inversion reduced the mean absolute error by 18.2% (H2), the hybrid thematic classification delivered a +5 percentage point gain over Random Forest (H3), integration of the three modules increased Recall_sys by 10.4 percentage points with acceptable latency growth (H4),

and the comprehensive Responsible AI framework achieved a Fairness Gap of 4.1% and a USS of 4.5/5 (H5), demonstrating full alignment with the stated objectives.

The system's key components include methods for context-sensitive sentiment analysis, polarity inversion detection, thematic classification and Responsible AI metrics (Fairness, Transparency, Accountability, Ethics). The novelty lies in the hybridisation of linguistic and machine learning models, as well as in the integration of quantitative metrics with formalised documentation (Model Cards), ensuring a balance between accuracy and accountability in critical information scenarios.

The system can be deployed in media monitoring analytical centres, cybersecurity structures (CERT, OSINT) and analytical units of government institutions. With a throughput of over 17 documents per second and a latency below 60 ms, it can be scaled for streaming analysis. The presence of an audit trail, Model Cards, and a Confidence Thresholding mechanism enables its integration into ecosystems that comply with the EU AI Act or ISO/IEC 42001 requirements.

While the current implementation focuses on Ukrainian- and Russian-language content, the platform's modular architecture allows for future adaptation to other languages and contexts, provided that appropriate datasets and model retraining procedures are available.

Further research is expected to expand the corpus to include multilingual and multimodal sources, automate model retraining using self-training, develop neural explainability modules (e.g., with SHAP or counterfactual generation), and investigate the dynamics of changes in information narratives over time using drift detection and streaming adaptation models.

Contribution of authors: Hennadii Bohuta, **Lesia Bilovus** and **Khrystyna Yurkiv** compiled a multilingual corpus of news and social-media posts on information threats related to Ukraine for 2022–2025. **Khrystyna Lipianina-Honcharenko**, **Hennadii Bohuta**, **Ihor Ihnatiev** implemented a sentiment analysis module with polarity inversion to detect covert hostility and sarcasm. **Khrystyna Lipianina-Honcharenko**, **Ihor Ihnatiev** developed a hybrid topic classification approach combining dictionary-based methods and machine-learning ensembles. **Oleg Illiashenko, Myroslav Komar**, **Ihor Ihnatiev** constructed the Responsible AI Evaluation (RAIE) framework with indicators for Fairness Gap, Model Cards, and User Satisfaction. **Ihor Ihnatiev** integrated all modules into a unified pipeline and performed quality (macro-F1, MAE) and performance (latency, throughput) evaluation. **Khrystyna Lipianina-Honcharenko, Ihor Ihnatiev** conducted the experimental validation of hypotheses H1–H5 and synthesised the resulting evidence. **Oleg Illiashenko** provided the general review and editing of the research results and the manuscript itself.

Project information: This study depicts the results of an interdisciplinary research project between West Ukrainian National University (Department of Information Computer System and Control, and Cyber Security Department), Leeds Beckett University (School of Built Environment, Engineering and Computing), UK, and National Aerospace University “Kharkiv Aviation Institute”, Ukraine, dedicated to the development, testing, and implementation of an integrated text-mining technology for real-time monitoring of information threats in Ukraine. Its goal is to design and experimentally validate an end-to-end system that automates context-sensitive sentiment analysis, polarity-inversion detection and thematic classification within a Responsible AI framework. The corpus comprises 10,350 unique documents collected from major Ukrainian news RSS feeds, social media APIs, and relevant blogs during January-December 2024, with stratified 5-fold cross-validation for evaluation. The system is realised as a modular client–server stack (Python + Streamlit + Transformers + Docker) suitable for secure, scalable deployment and integration into compliance-oriented environments (e.g., EU AI Act/ISO/IEC 42001). The authors note that the study was conducted without external financial support.

The technology’s effectiveness and reliability hinge on accurate detection of covert manipulations (polarity inversion), fair/transparent classification across multilingual, multi-topic streams and low-latency throughput for operational use (CERT, OSINT, government analytics). Empirically, the sentiment module attains macro-F1 = 0.85; the inversion algorithm reduces polarity MAE by 18.2% with minimal latency overhead; and the hybrid thematic classifier achieves macro-F1 = 0.83 at ≈55 ms/doc and ≈18 docs/s. Integrated end-to-end, the pipeline raises overall recall by +10.4 percentage points while keeping latency growth within ~10% and the RAIE framework ensures $\Delta F1 \leq 5\%$ with an expert User Satisfaction Score of 4.14/5.

The results presented in the paper contribute to addressing critical challenges in information security by providing novel algorithms for context-sensitive sentiment analysis, inversion-aware classification, and responsible evaluation through fairness, accountability, transparency, and user satisfaction indicators.

Conflict of Interest

The authors declare that they have no conflict of interest about this research, whether financial, personal, authorship or otherwise, that could affect the research and its results presented in this paper.

ISSN 1814-4225 (print)
ISSN 2663-2012 (online)

Financing

This study was conducted without financial support.

Data Availability

The work is accompanied by associated data in the data repository:

https://figshare.com/articles/dataset/News_dataset_about_Ukraine/29020670?file=54415925

Use of Artificial Intelligence

The authors have used artificial intelligence technologies within acceptable limits to provide their own verified data, as described in the research methodology section.

All authors have read and approved the published version of the manuscript.

References

1. Robertson, S., & Zaragoza, H. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 2009, vol. 3, iss. 4, pp. 333–389. DOI: 10.1561/1500000019.

2. Xiong, L., Xiong, C., Li, Y., Tang, K.F., Liu, J., Bennett, P., Ahmed, J., & Overwijk, A. Approximate nearest neighbor negative contrastive learning for dense text retrieval. *Proceedings of the International Conference on Learning Representations*, ICLR, 2020. Available at: https://arxiv.org/abs/2007.00808 (accessed 12.05.2025).

3. Stanovsky, G., Michael, J., Zettlemoyer, L., & Dagan, I. Supervised open information extraction. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, New Orleans, Louisiana, USA, ACL, 2018, pp. 885–895. DOI: 10.18653/v1/N18-1081.

4. Hamborg, F., Donnay, K. & Gipp, B. Automated identification of media bias in news articles: an interdisciplinary literature review. *International Journal on Digital Libraries*, 2019, vol. 20, pp. 391–415. DOI: 10.1007/s00799-018-0261-y.

5. Shu, K., Bhattacharjee, A., Alatawi, F., Nazer, T.H., Ding, K., Karami, M., & Liu, H. Combating disinformation in a social media age. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2020, vol. 10, iss. 6, article no. e1385. DOI: 10.1002/widm.1385.

6. Doddapaneni, S., Khan, M. S. U. R., Venkatesh, D., Dabre, R., Kunchukuttan, A., & Khapra, M. M. Cross-lingual auto evaluation for assessing multilingual LLMs. *Proceedings of the 64rd Annual Meeting of the Association for Computational Linguistics (volume 1: Long Papers)*, Vienna, Austria, ACL, 2025, pp. 29297–29329. DOI: 10.18653/v1/2025.acl-long.1419.

7. Pires, T., Schlinger, E., & Garrette, D. How multilingual is Multilingual BERT? *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy, ACL, 2019, pp. 4996–5001. DOI: 10.18653/v1/P19-1493.

8. Chen, D., Li, Z., Gu, B., & Chen, Z. Multimodal named entity recognition with image attributes and image knowledge. *Database Systems for Advanced Applications.* DASFAA, 2021, *Lecture Notes in Computer Science*, vol. 12682, pp. 183–198. DOI:10.1007/978-3-030-73197-7_12.

9. Satvat, K., Gjomemo, R., & Venkatakrishnan, V.N. Extractor: Extracting attack behavior from threat reports. *Proceedings of the 2021 IEEE European Symposium on Security and Privacy*, EuroS&P, Vienna, Austria, IEEE, 2021, pp. 414–429. DOI:10.1109/EuroSP51992.2021.00046.

10. Kapan, S., & Sora Gunal, E. Improved phishing attack detection with machine learning: A comprehensive evaluation of classifiers and features. *Applied Sciences*, 2023, vol. 13, iss. 24, pp. 1–13, DOI: 10.3390/app132413269.

11. Kondratyuk, D., & Straka, M. 75 languages, 1 model: Parsing Universal Dependencies universally. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing* EMNLP-IJCNLP, Hong Kong, China, ACL, 2019, pp. 2779–2795. DOI: 10.18653/v1/D19-1284.

12. Lu, Y., Nie, Z., Cheng, T., Gao, Y., & Wen, J. R. Name disambiguation using a web connection. *Proceedings of AAAI 2007 Workshop on Information Integration on the Web*, IIWeb, Vancouver, Canada, AAAI, 2007, pp. 57–61. Available at: https://cdn.aaai.org/Workshops/2007/WS-07-14/WS07-14-010.pdf (accessed 12.05.2025).

13. Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. A density-based algorithm for discovering clusters in large spatial databases with noise. *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining*, AAAI, Portland, Oregon, USA, 1996, pp. 226–231. DOI: 10.5555/3001460.3001507.

14. Reimers, N., & Gurevych, I. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Hong Kong, China, ACL, 2019, pp. 3982–3992. DOI: 10.18653/v1/D19-1410.

15. Gao, T., Yao, X., & Chen, D. SimCSE: Simple contrastive learning of sentence embeddings. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Punta Cana, Dominican Republic, EMNLP, 2021, pp. 6894–6910. DOI: 10.18653/v1/2021.emnlp-main.552.

16. Potthast, M., Kiesel, J., Reinartz, K., Bevendorff, J., & Stein, B. A stylometric inquiry into hyperpartisan and fake news. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, Melbourne, Australia, ACL, 2018, pp. 231–240. DOI: 10.18653/v1/P18-1022.

17. Corley, C., & Mihalcea, R. Measuring the semantic similarity of texts. *Proceedings of the ACL*

Workshop on Empirical Modeling of Semantic Equivalence and Entailment, Ann Arbor, USA, ACL, 2005, pp. 13–18. DOI: 10.3115/1641356.1641359.

18. Schubert, E., Sander, J., Ester, M., Kriegel, H.-P., & Xu, X. DBSCAN revisited, revisited: Why and how you should (still) use DBSCAN. *ACM Transactions on Database Systems*, TODS, 2017, vol. 42, iss. 3, pp. 1–21. DOI: 10.1145/3068335.

19. Gu, X., Angelov, P. P., Kangin, D., & Principe, J. C. A new type of distance metric and its use for clustering. *Evolving Systems*, 2017, vol. 8, iss. 3, pp. 167–177. DOI: 10.1007/s12530-017-9195-7.

20. Artetxe, M., & Schwenk, H. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 2019, vol. 7, pp. 597–610. DOI: 10.1162/tacl_a_00288.

21. Hutto, C J., & Gilbert, E. VADER: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the 8th International Conference on Weblogs and Social Media*, ICWSM, Ann Arbor, Michigan, USA, AAAI, 2014, vol. 8, no. 1, pp. 216–225. DOI: 10.1609/icwsm.v8i1.14550.

22. Medhat, W., Hassan, A., & Korashy, H. Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, 2014, vol. 5, no. 4, pp. 1093–1113. DOI: 10.1016/j.asej.2014.04.011.

23. Sun, C., Huang, L., & Qiu, X. Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentence. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT*, Minneapolis, Minnesota, USA, Association for Computational Linguistics, 2019, pp. 380–385. DOI: 10.18653/v1/N19-1035.

24. Joshi, A., Bhattacharyya, P., & Carman, M.J. Automatic sarcasm detection: A survey. *ACM Computing Surveys*, 2017, vol. 50, iss. 5, pp. 1–22. DOI: 10.1145/3124420.

25. Kaushik, D., Hovy, E., & Lipton, Z.C. Learning the difference that makes a difference with counterfactually-augmented data. *Proceedings of the 8th International Conference on Learning Representations*, ICLR, 2020. Available at: https://arxiv.org/abs/1909.12434 (accessed 12.09.2025).

26. Volkova, S., & Jang, J.Y. Misleading or falsification: Inferring deceptive strategies and types in online news and social media. *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, Beijing, PRC, ACM, 2018, pp. 575–583. DOI: 10.1145/3184558.3188728.

27. Zhang, L., Wang, S., & Liu, B. Deep learning for sentiment analysis: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2018, vol. 8, no. 4, e1253. DOI: 10.1002/widm.1253.

28. Ghosh, D., & Veale, T. Fracking sarcasm using neural network. *Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, WASSA, San Diego, USA, ACL, 2016, pp. 161–169. DOI: 10.18653/v1/W16-0425.

29. Cakebread‑Andrews, O., Ha, L. A., Frommholz, I., & Can, B. Error analysis of NLP models and non‑native speakers of English identifying sarcasm in Reddit comments. *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*, Torino, Italy, ELRA and ICCL, 2024, pp. 6247–6256. Available at: https://aclanthology.org/2024.lrec-main.552/ (accessed 12.05.2025).

30. Breve, B., Caruccio, L., Cirillo, S., Deufemia, V., & Polese, G. Analyzing the worldwide perception of the Russia–Ukraine conflict through Twitter. *Journal of Big Data*, 2024, vol. 11, article no. 76, pp. 1–33. DOI: 10.1186/s40537‑024‑00921‑w.

31. Manning, C. D., Raghavan, P., & Schütze, H. *Introduction to Information Retrieval*. Cambridge, Cambridge University Press, 2008. 506 p.

32. Joachims, T. Text categorization with support vector machines: learning with many relevant features. *Proceedings of the 10th European Conference on Machine Learning*, Chemnitz, Germany, AAAI, 1998, pp. 137–142. DOI: 10.1007/BFb0026683.

33. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Minneapolis, Minnesota, USA, ACL, 2019, vol. 1, pp. 4171–4186. DOI: 10.18653/v1/N19-1423.

34. Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., & Hovy, E. Hierarchical attention networks for document classification. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, San Diego, California, USA, ACL, 2016, pp. 1480–1489. DOI: 10.18653/v1/N16-1174.

35. Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. Unsupervised cross-lingual representation learning at scale. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, ACL, 2020, pp. 8440–8451. DOI: 10.18653/v1/2020.acl-main.747.

36. Baccianella, S., Esuli, A., & Sebastiani, F. Senti WordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. *Proceedings of the Seventh International Conference on Language Resources and Evaluation*, Valletta, Malta, ELRA, 2010, pp. 2200–2204. Available at: http://lrec-conf.org/proceedings/lrec2010/pdf/769_Paper.pdf (accessed 13.09.2025).

37. Sanh, V., Debut, L., Chaumond, J., & Wolf, T. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *Proceedings of the 5th Workshop on Energy Efficient Machine Learning and Cognitive Computing - NeurIPS 2019*, Vancouver, BD, Canada, IEEE 2019. DOI: 10.48550/arXiv.1910.01108.

38. Jiao, X., Yin, Y., Shang, L., Jiang, X., Chen, X., Li, L., Wang, F., & Liu, Q. TinyBERT: Distilling BERT for natural language understanding. *Findings of the Association for Computational Linguistics*, 2020, pp. 4163–

ISSN 1814-4225 (print)
ISSN 2663-2012 (online)

4174. DOI: 10.18653/v1/2020.findings-emnlp.372.

39. Breiman, L. Random forests. *Machine Learning*. 2001, vol. 45, no. 1, pp. 5–32. DOI: 10.1023/A:1010933404324.

40. Prabowo, R., & Thelwall, M. Sentiment analysis: A combined approach. *Journal of Informetrics*, 2009, vol. 3, no. 2, pp. 143–157. DOI: 10.1016/j.joi.2009.01.003.

41. Forman, G. An extensive empirical study of feature selection metrics for text classification. *Journal of Machine Learning Research*. 2003, vol. 3, pp. 1289–1305. Available at: https://www.jmlr.org/papers/volume3/forman03a/forman03a.pdf (accessed 14.09.2025).

42. Campos, R., Mangaravite, V., Pasquali, A., Jorge, A.M., Nunes, C., & Jatowt, A. YAKE! Keyword extraction from single documents using multiple local features. *Information Sciences*, 2020, vol. 509, pp. 257–289. DOI: 10.1016/j.ins.2019.09.013.

43. Mihalcea, R., & Tarau, P. TextRank: Bringing order into texts. *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, Barcelona, Spain, ACL, 2004, pp. 404–411. Available at: https://aclanthology.org/W04-3252 (accessed 14.05.2025).

44. Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 1990, vol. 41, no. 6, pp. 391–407. DOI: 10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9.

45. Blei, D. M., Ng, A. Y., & Jordan, M. I. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 2003, vol. 3, pp. 993–1022. Available at: https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf (accessed 14.05.2025).

46. Grootendorst, M. *KeyBERT: Minimal keyword extraction with BERT*. 2020. Available at: http://dx.doi.org/10.5281/zenodo.4461265 (accessed 14.05.2025).

47. Cohan, A., Feldman, S., Beltagy, I., Downey, D., & Weld, D. S. SPECTER: Document-level representation learning using citation-informed transformers. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. ACL, 2020, pp. 2270–2282. DOI: 10.18653/v1/2020.acl-main.207.

48. Wan, X., & Xiao, J. Single document keyphrase extraction using neighborhood knowledge. *Proceedings of the 23rd AAAI Conference on Artificial Intelligence*, Chicago, USA, AAAI, 2008, pp. 855–860. Available at: https://cdn.aaai.org/AAAI/2008/AAAI08-136.pdf (accessed 12.05.2025).

49. Papagiannopoulou, E., & Tsoumakas, G. A review of keyphrase extraction. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*. 2020, vol. 10, no. 2, article no. e1339. DOI: 10.1002/widm.1339.

50. Hardt, M., Price, E., & Srebro, N. Equality of opportunity in supervised learning. *NIPS'16: Proceedings of the 30th International Conference on Neural Information Processing Systems*, Barcelona, Spain, Curran Associates Inc., 2016, pp. 3323–3331. DOI: 10.5555/3157382.

51. Arnold, M., Bellamy, R.K.E., Hind, M., Houde, S., Mehta, S., Nair, R., & Nushi, B. Factsheets: Increasing trust in AI services through supplier's declarations of conformity. *IBM Journal of Research and Development*. 2019, vol. 63, no. 4/5, pp. 6:1–6:13. DOI: 10.1147/JRD.2019.2942288.

52. Dodge, J., Ilharco, G., Schwartz, R., Farhadi, A., Hajishirzi, H., & Smith, N. A. *Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping*. 2020. DOI: 10.48550/arXiv.2002.06305.

53. Fort, S., Ren, J., & Lakshminarayanan, B. Exploring the limits of out-of-distribution detection. *NIPS'21: Proceedings of the 35th International Conference on Neural Information Processing Systems*. NeurIPS, Virtual Conference, Curran Associates Inc., 2021, pp. 1–14. Available at: https://proceedings.neurips.cc/paper_files/paper/2021/file/3941c4358616274ac2436eacf67fae05-Paper.pdf (accessed 14.05.2025).

54. Ribeiro, M. T., Singh, S., & Guestrin, C. "Why should I trust you?": Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, USA, ACM, 2016, pp. 1135–1144. DOI: 10.1145/2939672.2939778.

55. Lundberg, S. M., & Lee, S.-I. *A* unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*. NeurIPS, Long Beach, California, USA, 2017, vol. 30. Available at: https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf (accessed 14.05.2025).

56. Wachter, S., Mittelstadt, B., & Russell, C. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology*. 2017, vol. 31, no. 2, pp. 841–887. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3063289 (accessed 14.05.2025).

57. Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter*. 2017, vol. 19, no. 1, pp. 22–36. DOI: 10.1145/3137597.3137600.

58. Vidgen, B., Harris, A., Nguyen, D., Tromble, R., Hale, S., & Margetts, H. Challenges and frontiers in abusive content detection. *Proceedings of the Third Workshop on Abusive Language*, Florence, Italy, ACL, 2020, pp. 6025–6044. Available at: https://ora.ox.ac.uk/objects/uuid:3864e746-88c8-4f99-b912-52f4b4be289a/files/m25d33782b006cec944b22e5d744ed1b7 (accessed 14.05.2025).

59. Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. AI4People – An ethical framework for a good AI society. *Minds and Machines*. 2018, vol. 28, no. 4, pp. 689–707. DOI: 10.1007/s11023-018-9482-5.

60. Sachenko, A., Lendiuk, T., Lipianina‑Honcharenko, K., Dobrowolski, M., Boguta, G., & Bytsyura, L. Method of determining the text sentiment by thematic rubrics. *Proceedings of the 8th International Conference on Computational Linguistics and Intelligent Systems*, Lviv, Ukraine, CEUR, 2024, pp. 404–414. DOI: 10.31110/COLINS/2024-3/026.

61. Lipianina-Honcharenko, K., Soia, M., Yurkiv, K., & Ivasechko, A. Evaluation of the effectiveness of machine learning methods for detecting disinformation in Ukrainian text data. *Proceedings of the 5th International Conference on Computational Methods for Information Security*, Zaporizhzhia, Ukraine, CEUR, 2024, pp. 97–109. Available at: https://ceur-ws.org/Vol-3702/paper9.pdf (accessed 14.05.2025).

62. Lipianina-Honcharenko, K., Lendiuk, D., Melnyk, N., Komar, M., & Lendiuk, T. Evaluation of the keyword selection methods effectiveness for the fake news classification. *Proceedings of the 10th International Scientific Conference on Information Technology and Interactions*, Kyiv, Ukraine, CEUR, 2024, pp. 109–122. Available at: https://ceur-ws.org/Vol-3909/Paper_9.pdf. (accessed 12.05.2025).

63. Zhang, L. L., Han, S., Wei, J., Zheng, N., Cao, T., Yang, Y., & Liu, Y. NN‑Meter: Towards accurate latency prediction of deep‑learning model inference on diverse edge devices. *Proceedings of the 19th Annual International Conference on Mobile Systems, Applications, and Services*. MobiSys, Wisconsin, USA, ACM, 2021, pp. 81–93. DOI: 10.1145/3458864.3467882.

64. Reddi, V.J., Cheng, C., Kanter, D., Mattson, P., Schmuelling, G., Wu, C.J., Coleman, C., Diamos, G., Elibol, M., Hall, D., Hazelwood, K., Hsu, B., Idiculla, N., Kumar, D., Levenberg, J., Tang, H., Warden, P., & et. al. MLPerf inference benchmark. *Proceedings of the ISCA'20: Proceedings of the ACM/IEEE 47th Annual International Symposium on Computer Architecture*. ISCA, Valencia, Spain, IEEE, 2020, pp. 446–459. DOI: 10.1109/ISCA45697.2020.00045.

65. Memarian, B., & Doleck, T. Fairness, accountability, transparency, and ethics (FATE) in artificial intelligence (AI) and higher education: A systematic review. *Computers and Education: Artificial Intelligence*. 2023, vol. 5, article no. 100152. DOI: 10.1016/j.caeai.2023.100152.

66. Kharchenko, V., Fesenko, H., & Illiashenko, O. Basic model of non-functional characteristics for assessment of artificial intelligence quality. *Radioelectronic and Computer Systems*. 2022, no. 2, pp. 131–144. DOI: 10.32620/reks.2022.2.11

67. *Streamlit. Streamlit web application*. Available at: https://github.com/streamlit/streamlit (accessed 14.05.2025).

68. Dotsenko, S., Illiashenko, O., Kharchenko, V., & Morozova, O. Integrated information model of an enterprise and cybersecurity management system: From data to activity. *International Journal of Cyber Warfare and Terrorism*. 2022, vol. 12, no. 2, pp. 1–21. DOI: 10.4018/IJCWT.305860.

69. Mygal, V., Mygal, G., & Illiashenko, O. Intelligent decision support – cognitive aspects. In: *Digital Transformation, Cyber Security and Resilience of Modern Societies*. *Studies in Big Data*, vol. 84. Cham, Springer, 2021, pp. 395–411. DOI: 10.1007/978-3-030-79934-2_26.

70. Kharchenko, V., Fesenko, H., & Illiashenko, O. Quality Models for Artificial Intelligence Systems: Characteristic-Based Approach, Development and Application. *Sensors*, 2022, vol. 22, iss. 13, article no. 4865. DOI: 10.3390/s22134865.

Received 15.05.2025, Accepted 25.08.2025

ЗАГАЛЬНИЙ МЕТОД ВИЯВЛЕННЯ ІНФОРМАЦІЙНИХ ЗАГРОЗ У РЕЖИМІ РЕАЛЬНОГО ЧАСУ НА ПРИКЛАДІ УКРАЇНИ

Х. В. Ліп'яніна-Гончаренко, М. П. Комар, Г. М. Богута, І. В. Ігнатєв, Х. В. Юрків, О. О. Ілляшенко, Л. І. Біловус

Предметом дослідження є загальний набір методів і системна архітектура для текстової аналітики, що забезпечують виявлення та моніторинг інформаційних загроз у режимі реального часу, верифіковані на кейсі України. Запропонований набір методів інтегрує аналіз тональності, оброблення інверсії полярності та тематичну класифікацію на основі методів машинного навчання. Дослідження актуалізується в умовах гібридної війни, коли інформаційне середовище стає полем дезінформації, маніпулятивних кампаній та когнітивного впливу. **Метою** є розробка та експериментальна валідація комплексної інформаційної технології для автоматизованого виявлення загроз в інформаційному просторі України, яка базується на принципах відповідального штучного інтелекту (Responsible AI) та сучасних методах обробки природної мови. **Завдання**: формування багатомовного корпусу текстів новин та соціальних медіа, реалізація модуля аналізу тону з урахуванням інверсії полярності, розробка гібридного методу тематичної класифікації із залученням словників ключових слів та ансамблів моделей машинного навчання, побудова фреймворку оцінювання відповідального ШІ (RAIE) із показниками чесності, прозорості та користувацької задоволеності. Отримані **результати** підтверджують всі п'ять висунутих гіпотез: розроблений модуль аналізу тону досягає макро-F1 = 0.85 та знижує MAE на 18.2% порівняно з базовою моделлю; алгоритм виявлення інверсії дозволяє автоматично змінювати знак емоційного індексу при виявленні маніпулятивних повідомлень, покращуючи точність визначення ворожих наративів; гібридна тематична класифікація забезпечує макро-F1 = 0.83 при затримці 55 мс/документ та продуктивності 18 документів/секунда; інтеграція модулів у єдиний пайплайн підвищує повноту виявлення загроз на 10.4% без істотного зростання затримки; впровадження концептуальної моделі RAIE забезпечує ΔF1

ISSN 1814-4225 (print)
ISSN 2663-2012 (online)

≤ 5%, середній бал задоволеності експертів 4.14/5 та незначне збільшення затримки (<10%). Висновки свідчать про те, що запропонована система поєднує високу точність виявлення інформаційних загроз із принципами етичності, прозорості та користувацької довіри, що забезпечує її практичну цінність для державних центрів кіберзахисту, CERT та OSINT-платформ. **Висновки.** Наукова новизна полягає у розробці нових методів: контекстно-чутливого аналізу тону з урахуванням специфіки військової лексики; алгоритму інверсії полярності для виявлення прихованої ворожості; гібридної тематичної класифікації, що поєднує машинне навчання та експертні словники; інтегрованої архітектури інформаційної технології з продуктивністю >17 документів/секунда; моделі оцінювання відповідального ШІ з впровадженням Fairness Gap, Model Cards та User Satisfaction Score.

Ключові слова: текстовий майнінг; аналіз настроїв; виявлення інверсій; класифікація текстів; машинне навчання.

Ліп'яніна-Гончаренко Христина Володимирівна – д-р техн. наук, доц., Західноукраїнський національний університет, Тернопіль, Україна.

Комар Мирослав Петрович – д-р техн. наук, проф., Західноукраїнський національний університет, Тернопіль, Україна.

Богута Геннадій Миколайович – студент, Західноукраїнський національний університет, Тернопіль, Україна.

Ігнатєв Ігор Васильович – викладач, Західноукраїнський національний університет, Тернопіль, Україна.

Юрків Христина Володимирівна – студентка, Західноукраїнський національний університет, Тернопіль, Україна.

Ілляшенко Олег Олександрович – канд. техн. наук., Університет Лідс Беккет, Лідс, Велика Британія; доц., каф. комп'ютерних систем, мереж та кібербезпеки факультету радіоелектроніки, комп'ютерних систем та інфокомунікацій, Національний аерокосмічний університет «Харківський авіаційний інститут», Харків, Україна.

Біловус Леся Іванівна– д-р іст. наук, проф., Західноукраїнський національний університет, Тернопіль, Україна.

Khrystyna Lipianina-Honcharenko – Doctor of Technical Sciences, Associate Professor Department of Information Computer System and Control, West Ukrainian National University, Ternopil, Ukraine,
e-mail: xrustya.com@gmail.com, ORCID: 0000-0002-2441-6292, Scopus Author ID: 59548850400.

Myroslav Komar – Doctor of Technical Sciences, Professor Department of Information Computer System and Control, West Ukrainian National University, Ternopil, Ukraine,
e-mail: mko@wunu.edu.ua, ORCID: 0000-0001-6541-0359, Scopus Author ID: 35366491300.

Hennadii Bohuta – Student, Department of Information Computer System and Control, West Ukrainian National University, Ternopil, Ukraine,
e-mail: genaboguta7@gmail.com, ORCID: 0009-0000-9788-1753, Scopus Author ID: 59155725000.

Ihor Ihnatiev – Lecturer, Department, of Cyber Security, West Ukrainian National University, Ternopil, Ukraine,
e-mail: iiv@wunu.edu.ua, ORCID: 0000-0003-0729-9247, Scopus Author ID: 57191954223

Khrystyna Yurkiv – Student, Department of Information Computer System and Control, West Ukrainian National University, Ternopil, Ukraine,
e-mail: kh.yurkiv@wunu.edu.ua, ORCID:0009-0007-4917-3251, Scopus Author ID: 58776326000.

Oleg Illiashenko – PhD, School of Built Environment, Engineering and Computing, Leeds Beckett University, Leeds, United Kingdom; Associate Professor, Department of Computer Systems, Networks and Cybersecurity, Faculty of Radio Electronics, Computer Systems and Infocommunications, National Aerospace University "Kharkiv Aviation Institute", Kharkiv, Ukraine;
e-mail: o.illiashenko@leedsbeckett.ac.uk, o.illiashenko@khai.edu, ORCID: 0000-0002-4672-6400, Scopus Author ID: 55842633400.

Lesia Bilovus – Doctor of History, Professor of the Department of Information and Socio-Cultural Activity, West Ukrainian National University, Ternopil, Ukraine,
e-mail: l.bilovus@wunu.edu.ua, ORCID: 0000-0003-4882-4511, Scopus Author ID: 57219598554.