Classifying multilingual information streams: a deep learning approach to propaganda detection with annotation guidelines and error analysis*

Andriy Lutskiv[1,*], Vasyl Yatsyshyn[1]

[1] *Ternopil Ivan Pului National Technical University, Ruska st 56 46001 Ternopil, Ukraine*

Abstract

This paper addresses classification of multilingual information streams (Ukrainian, Russian, English) to reduce overload and identify propaganda during Russian-Ukrainian war. We present an annotation framework achieving excellent inter-rater agreement (κ=0.850) on 1,066 texts (4-class scheme: NEWS, ANALYSIS, PROPAGANDA, NOISE). Evaluating TF-IDF and two multilingual transformers, mBERT fine-tuned for 5 epochs achieved best performance (84.58% accuracy, 0.805 Macro F1). An ablation study revealed that Focal Loss and oversampling degrade performance; class-weighted loss is sufficient. Error analysis (15.4% error rate) identified three core challenges: NEWS-ANALYSIS boundary ambiguity (42% of errors), propaganda grounded in factually accurate content (31% of PROPAGANDA errors), and length-dependent degradation (18.1% on texts >500 characters). These findings reveal fundamental semantic limits of shallow language models: detecting propaganda requires reasoning over causal claims, not just factual accuracy. We describe a cost-effective, cloud-agnostic system architecture for real-time classification at scale and identify reasoning-based approaches to manipulation detection as future work.

Keywords

information warfare, propaganda detection, multilingual NLP, transformer models, annotation guidelines, BERT, machine learning, Russian-Ukrainian war, information streams, MLOps, Big Data, Apache Spark, cost-effective deployment

1. Introduction

Information warfare has become a critical challenge in conflict zones. High-volume information streams from messaging platforms (Telegram, WhatsApp), social networks (Facebook, X), and news sources create an environment where distinguishing factual reporting from propaganda is computationally demanding.

This paper addresses one component of a broader intelligent information system: real-time classification of multilingual information streams (Ukrainian, Russian, English) into four categories (NEWS, ANALYSIS, PROPAGANDA, NOISE) to enable downstream filtering and prioritization. While grounded in the Ukrainian-Russian war context, our methods are applicable to any language combination and information source. The core contributions (Research Questions) are:

RQ1: An annotation framework achieving excellent inter-rater agreement (κ=0.850), demonstrating that systematic guideline development is essential for conflict-context classification.

RQ2: Benchmarking multilingual transformers, identifying mBERT as optimal (84.58% accuracy).

RQ3: Error analysis revealing three fundamental challenges: NEWS-ANALYSIS boundary ambiguity, propaganda grounded in factually accurate content, and length-dependent degradation.

We describe a cost-effective, cloud-agnostic system architecture for real-time classification at scale and identify reasoning-based approaches to manipulation detection as future work. The

* [1] CMIS-2026: The Ninth International Workshop on Computer Modeling and Intelligent Systems, May 5, 2026, Zaporizhzhia, Ukraine.

l.andriy@gmail.com (A. Lutskiv); yacyshyn@tntu.edu.ua (V. Yatsyshyn)

0000-0002-9250-4075 (A. Lutskiv); 0000-0002-5517-6359 (V. Yatsyshyn)

MANIPULATION class was excluded from the main task due to severe class imbalance (n=14, 1.3%) but is addressed as a secondary detection layer in future research.

2. Related work

Corpus linguistics enables quantitative analysis of large text collections [1, 2]. Prior work [3] demonstrated Big Data approaches to corpus construction — integrating NLP preprocessing, feature extraction (TF-IDF, LSA), and distributed processing (Apache Spark) — supporting multilingual analysis of information flows. These principles underlie modern approaches to real-time information stream filtering and form the methodological basis for large-scale multilingual corpus construction. Propaganda detection in NLP has evolved from lexical-syntactic approaches [4] to deep learning methods using transformers. Foundational challenges, identified in prior work [5], include:

1. Propaganda expressed through factually accurate statements with emotional framing rather than overt falsification.
2. Fine-grained classification requiring understanding of how facts are framed relative to causal claims and value judgments.
3. Source-specific rhetorical patterns in multilingual contexts, where Russian-language propaganda in Ukrainian conflict contexts employs distinct lexical choices and argumentative structures [8].

Recent work has extended propaganda detection to multilingual and conflict-specific settings. The research [6] proposed a linguistics-based approach for detecting pro-Kremlin propaganda in Ukrainian and Romanian news and Telegram posts, combining transformer classifiers with hand-crafted keyword features. In [7] introduced the EUvsDisinfo dataset (18249 articles, 42 languages), showing that mBERT and XLM-RoBERTa achieve strong performance on cross-lingual propaganda classification. In [19] authors applied machine learning classifiers to war-related Twitter content following the 2022 full-scale invasion, finding that SVM and XGBoost outperform LSTM-based approaches on short social media texts. However, these studies focus primarily on binary (propaganda/not-propaganda) or source-level classification; none provides a multi-class annotation framework for real-time stream classification with formal inter-rater agreement validation across Ukrainian, Russian, and English simultaneously. Transformer-based models are standard for multilingual text classification. mBERT [9] and XLM-RoBERTa [10] are widely used for cross-lingual transfer. Comparative studies show:

1. mBERT is more effective for well-represented languages (English, German, French) and multilingual datasets simultaneously [11].
2. XLM-R generalizes better to unseen language pairs but requires more training data for fine-grained tasks [12].
3. Fine-tuning duration and hyperparameter selection are critical — longer training risks overfitting on small datasets [13].

Prior work [14] demonstrates that sophisticated loss functions such as Focal Loss do not universally improve performance on class-imbalanced datasets; class-weighted cross-entropy loss often provides a sufficient and more stable baseline. Robust annotation guidelines are essential for text classification in subjective domains like propaganda detection [15, 16]. Cohen's kappa (κ) [17] measures inter-rater agreement; $\kappa > 0.8$ is considered excellent agreement. Research [3, 18] demonstrated that iterative refinement of annotation procedures improves agreement by revealing ambiguous cases and motivating more precise class definitions.

Despite these advances, several gaps persist. No prior study has developed a formal four-class annotation taxonomy (NEWS, ANALYSIS, PROPAGANDA, NOISE) for conflict-context

multilingual streams, with guidelines validated through iterative inter-rater agreement measurement. Existing systems either operate on binary propaganda labels or do not address the NEWS–ANALYSIS semantic boundary that arises when factual reporting and interpretive framing co-occur. Furthermore, prior work does not address scalable, cloud-agnostic architectures for real-time classification of multilingual information streams at the message level.

3. Dataset and annotation methodology

3.1. Dataset composition

The dataset comprises 1,080 texts collected from 7 Telegram information channels over the period September 2025 to February 2026. Channels were selected to represent Ukrainian and Russian information sources, encompassing news, analysis, and propaganda-focused content. Channel-specific collection periods vary: the longest spanning approximately 5 months (September–February), while others span 2–3 weeks in January–February 2026. Channel composition: 2 Ukrainian news channels (300 texts); 2 Russian news channels (319 texts); 2 analysis channels (300 texts); 1 propaganda-focused channel (161 texts).

3.2. Annotation scheme and class definitions

Four primary classes were defined based on linguistic and pragmatic criteria: NEWS - factual reporting of events, statements, or occurrences. Texts classified as NEWS present information about observable events or documented facts without substantial interpretation or ideological framing. Emotional or harsh language does not disqualify a text from NEWS classification if the underlying facts are accurately conveyed.

ANALYSIS - factual information combined with interpretation, strategic reasoning, or causal explanation. ANALYSIS texts present facts or reported events and then contextualize them within broader narratives, explain their implications, or draw conclusions about causes and effects. ANALYSIS differs from NEWS in that it explicitly offers interpretive frameworks.

PROPAGANDA - factually inaccurate or distorted claims presented with emotional framing, loaded language, or implicit value judgments designed to influence opinion. PROPAGANDA may also include factually accurate information presented with false causal reasoning or misleading contextualization that obscures truth. The defining characteristic is the presence of factual distortion or misleading causal reasoning.

NOISE - spam, off-topic messages, system notifications, or content unintelligible due to corruption or poor quality. NOISE includes empty messages, link-only posts, non-linguistic content, and messages in languages other than Ukrainian, Russian, or English.

MANIPULATION - this fifth classification class was initially defined as accurate facts combined with distorted causal reasoning across unrelated topics. However, due to severe class imbalance (n=14, 1.3% of dataset), MANIPULATION was excluded from the main classification task. This class is discussed in Section 9 as a target for secondary analysis using reasoning-based approaches.

3.3. Annotation procedure

Two annotators independently labeled all 1080 texts [15, 16]. Annotators received detailed guidelines (described in Section 4) and participated in a training phase using 50 exemplar texts. After independent annotation, disagreements were identified and discussed. For texts with disagreement, annotators reviewed the guidelines, examined the specific text in context, and reached consensus through discussion. This consensus label was recorded as the final annotation. The dataset was then divided into three rounds of annotation refinement:

Version 1 (Initial): Full dataset annotated with initial guidelines. Inter-rater agreement $\kappa=0.686$ (substantial agreement) [17].

Version 2 (First Refinement): Following analysis of 50 disagreement cases from v1, guidelines were refined and clarified. A stratified sample of 150 texts (14% of dataset, balanced across classes) was re-annotated with revised guidelines. Agreement on this subsample: κ=0.762.

Version 3 (Final Refinement): Following analysis of remaining disagreements, guidelines were refined a second time. The same 150 texts were re-annotated. Agreement on this subsample: κ=0.850 (excellent agreement).

After three rounds of annotation and consensus-building final dataset composition with the follow class distribution (4-class scheme): NEWS - 478 texts (44.8%), ANALYSIS - 379 texts (35.6%), PROPAGANDA - 77 texts (7.2%), NOISE - 117 texts (11.0%) .

Thus, total texts number is 1,066. Due to exclusion of MANIPULATION class (n=14) which retained for reference in future work. Per-class agreement (κ) in v3 final: NEWS – 93.2%, NOISE – 90.5%, ANALYSIS – 88.3%, PROPAGANDA – 87.5%.

The final 1066-text dataset was used for all subsequent experiments. A stratified 80/20 train/test split was applied: training set n=864, test set n=216.

4. Annotation guidelines and validation

4.1. Detailed annotation guidelines

This section describes the refined guidelines achieving κ=0.850 inter-rater agreement. Guidelines evolved through iterative rounds of annotation and disagreement analysis.

NEWS: Texts reporting facts, events, or statements with minimal interpretation or ideological framing. Emotional or harsh language does not disqualify NEWS classification if underlying facts are accurate. Key features: direct reporting of events with attribution of claims to sources (e.g., "officials stated"), minimal causal explanation, and absence of misleading causality claims.

Boundary case (NEWS vs. PROPAGANDA): A Russian-language text accurately reporting Ukrainian military losses but framed with nationalist rhetoric remains NEWS if facts are documented and accurate. However, if the text falsely claims Ukrainian forces "intentionally sacrificed soldiers for Western propaganda," it becomes PROPAGANDA (factually inaccurate causal claim).

ANALYSIS: Texts combining factual information with explicit interpretation, causal reasoning, or strategic implications. ANALYSIS presupposes factual accuracy but adds explanatory frameworks. Key features: presentation of facts, explicit explanation of causes/effects/strategic implications, conditional reasoning, and absence of factual distortion.

PROPAGANDA: Texts containing factually inaccurate claims, distorted causality, or misleading contextualization designed to influence opinion through deception. Key features: factually inaccurate claims, accurate facts combined with false causal reasoning, conspiracy narratives without evidentiary support, and emotional framing combined with factual distortion. Critical distinction: PROPAGANDA requires both factual elements and misleading causal reasoning. Emotional language or harsh tone alone do not constitute propaganda. A Russian-language source accurately reporting Ukrainian losses with nationalist rhetoric remains NEWS/ANALYSIS if facts are accurate. However, if the same report falsely implies these losses prove Ukrainian military collapse or that Ukraine is "doomed," it becomes PROPAGANDA (misleading causality). The ТСК (ukr. - “ТЦК”) case exemplifies this: documented casualty figures + harsh Russian-language commentary = ANALYSIS or NEWS, not PROPAGANDA. This distinction is critical for conflict contexts where emotional language is common but factual reporting is valued.

NOISE: Texts containing no substantive information about conflict or politics; spam, metadata, or corrupted content; or languages other than Ukrainian, Russian, or English. Key features: system notifications, media-only posts, spam, off-topic chat, unintelligible content. Boundary case (NOISE vs. NEWS): a post with only a link is NOISE; a post with a link plus headline or summary is NEWS.

4.2. Iterative refinement and agreement progression

Inter-rater agreement improved substantially across three annotation rounds.

Round v1 (Initial): κ=0.686 (substantial). Disagreements ~31%; primary sources: NEWS-ANALYSIS boundary ambiguity; PROPAGANDA definition too broad (including opinion without factual distortion).

Round v2 (First Refinement): κ=0.762 (substantial). Disagreements ~24%; changes: clarified PROPAGANDA definition (must include factual distortion or misleading causality), provided NEWS-ANALYSIS boundary examples. Results: 11% improvement in κ, largest gains on PROPAGANDA (66% ⟶ 79% per-class agreement).

Round v3 (Final Refinement): κ=0.850 (excellent). Disagreements ~15%; changes: added detailed boundary cases (harsh language + accurate facts = NEWS, not PROPAGANDA), refined NOISE definition, included language-specific examples. Results: 11% improvement in κ from v2; excellent agreement on NEWS (93.2%), NOISE (90.5%), ANALYSIS (88.3%), PROPAGANDA (87.5%).

4.3. Hard cases and lessons learned

Two core challenges emerged in remaining disagreements (15 out of 150 in v3 final; ~10%):

NEWS-ANALYSIS boundary: Factual descriptions of military equipment or tactics without explicit strategic interpretation (e.g., "Ukraine is using HIMARS to strike Russian ammunition depots") sometimes triggered disagreement about whether factual mention of equipment with location implies strategic interpretation. Resolution: clarified that equipment descriptions without explicit causal claims remain NEWS.

PROPAGANDA vs. ANALYSIS: Texts combining accurate facts with harsh nationalist rhetoric (especially Russian sources reporting Ukrainian losses with triumphal framing) were sometimes classified as PROPAGANDA (due to emotional tone) and sometimes as NEWS/ANALYSIS (due to factual accuracy). Resolution: introduced the "multitruth" principle—texts are PROPAGANDA only if they distort facts or causality, not if they express accurate information with emotional framing.

These lessons demonstrate the difficulty of distinguishing propaganda from harsh but factual reporting in polarized information environments.

4.4. Validation and reliability

The κ progression (0.686 ⟶ 0.850) demonstrates the effectiveness of iterative guideline refinement for conflict-context text classification. The 24% improvement from v1 to v3 reflects the semantic complexity of distinguishing propaganda from harsh but factual reporting.

Per-class agreement in v3 final shows varying reliability: HIGH (NEWS 93.2%, NOISE 90.5%) with clear boundaries; MODERATE (ANALYSIS 88.3%, PROPAGANDA 87.5%) with fuzzy semantic boundaries at NEWS-ANALYSIS and PROPAGANDA-ANALYSIS transitions. This variation reflects genuine difficulty in these boundary regions rather than annotation inconsistency.

5. Classification methods

As a baseline classification method was chosen Term Frequency-Inverse Document Frequency (TF-IDF) feature extraction combined with Logistic Regression classification [3, 18]. TF-IDF captures word importance within documents and across the corpus, weighting common words down and rare but informative words up. For each text, a TF-IDF vector is computed using unigrams and bigrams. Stop words in Ukrainian, Russian, and English are removed based on standard lists. Logistic Regression (L2 regularization, max_iter=1000) is trained on the training set. Class weights are set inversely proportional to class frequencies to address class imbalance. TF-IDF is computationally efficient and provides interpretable results (high-weight features indicate important words for each class). It serves as a reasonable baseline for text classification tasks. Obtained results are: Accuracy - 78.80%, Macro F1 - 0.5808.

Another approach which was analyzed and tested based on transformer models mBERT (Multilingual BERT) and XLM-RoBERTa (XLM-R). These models (BERT variants) have become the standard for multilingual text classification.

mBERT (Multilingual BERT) [9] model pretrained on 104 languages including Ukrainian, Russian, and English. mBERT uses subword tokenization (WordPiece) and 12 encoder layers with 768-dimensional hidden states. Key strengths of this model is a strong performance on well-represented languages, and high efficiency for languages with small fine-tuning datasets. But, it may underperform on morphologically complex or low-resource languages.

XLM-RoBERTa (XLM-R) [10] model pretrained on more than 100 languages with a larger pretraining corpus than mBERT. Uses SentencePiece subword tokenization and 24 encoder layers with 1024-dimensional hidden states. The model has a stronger cross-lingual transfer, better performance on low-resource language pairs, but it is a larger model and slower in an inference.

Both models are fine-tuned for 4-class classification (NEWS, ANALYSIS, PROPAGANDA, NOISE) on the training set (n=864) with the follow hyperparameters:

- learning rate: 2e-5 (adaptive, with linear warmup);
- batch size: 32 (limited by GPU memory; Nvidia RTX 3060, 12 GB VRAM);
- maximum sequence length: 512 tokens (truncate if longer; rare in dataset);
- number of epochs: 3, 5, 7, 10 (varied in ablation study);
- optimizer: AdamW (PyTorch library);
- loss function: Cross-entropy with class weights (inversely proportional to class frequencies);
- early stopping: Monitored validation F1; stopped if no improvement for 3 epochs.

Texts are tokenized using the model's native tokenizer. Sequences longer than 512 tokens are truncated (fewer than 2% of dataset affected). Special tokens [CLS] and [SEP] are added as per BERT conventions. Texts are padded to batch size. The training loop computes loss on the training set, backpropagates, and updates model weights. Validation occurs every 100 steps using the training set (80% split of full training data; 10% held for validation during training). The test set (20% of 1,066 texts; n=216) is never used during training or validation.

The dataset exhibits class imbalance (NEWS 44.8%, ANALYSIS 35.6%, NOISE 11.0%, PROPAGANDA 7.2%). To address this, three strategies were evaluated:

1. Class-weighted cross-entropy loss which is a standard approach. Loss is weighted by class frequency weights:

$$w_c = \left(\frac{1}{frequency_c}\right) \quad (1)$$

This penalizes misclassification of rare classes (PROPAGANDA, NOISE) more heavily.

2. Focal Loss - a loss function that focuses on hard examples and de-emphasizes easy examples [14]. Focal loss is computed as

$$FL(p_t) = -(1-p_t)^{\gamma} * \log(p_t) \quad (2)$$

where p_t is the model's probability for the true class and γ (gamma) is a focusing parameter (set to 2.0). Focal Loss is expected to improve classification of rare classes by emphasizing difficult predictions.

3. Synthetic oversampling of rare classes (PROPAGANDA, NOISE) to balance the training set. Random oversampling replicates minority class examples until all classes have equal frequency.

At inference time, a fine-tuned model processes an input text:

1. Tokenization: text is tokenized using the model's tokenizer.
2. Forward pass: tokens are processed through the model, producing a hidden state for the [CLS] token (class representation).
3. Classification head: a linear layer (768 ⟶ 4 classes for mBERT; 1024 ⟶ 4 for XLM-R) is applied to the [CLS] state, producing logits.
4. Softmax: logits are normalized to class probabilities [0, 1].
5. Prediction: the class with highest probability is selected. A confidence score is computed as the maximum probability.

Confidence scores are used for filtering high-confidence predictions for downstream analysis (e.g., MANIPULATION detection).

6. Experimental results

The TF-IDF with Logistic Regression baseline achieved accuracy – 78.80%, macro F1 – 0.5808, per-class F1 scores: NEWS - 0.83, ANALYSIS - 0.81, PROPAGANDA - 0.47, NOISE – 0.80. The baseline establishes a reasonable starting point. Performance on PROPAGANDA is notably lower (0.47 F1), indicating that lexical features alone are insufficient for propaganda detection in this dataset. The relatively high NEWS and ANALYSIS scores (0.80+) suggest that these classes have distinct vocabulary patterns.

Based on the experiments conducted, the key findings are:

1. Best Model: mBERT fine-tuned for 5 epochs with standard class-weighted cross-entropy loss achieves the best overall performance (Accuracy: 84.58%, Macro F1: 0.8053). This model is selected for use in the system (Section 7).
2. Epoch Selection: Performance improves from 3 to 5 epochs, then degrades with further training (7, 10 epochs). At 10 epochs, PROPAGANDA F1 drops to 0.32, indicating severe overfitting on the training set. The 5-epoch checkpoint represents the optimal balance between underfitting and overfitting for this dataset.
3. Focal Loss Ineffective: Across all configurations, Focal Loss reduces performance compared to standard class-weighted loss. mBERT with Focal Loss (74.77% accuracy) underperforms the baseline (84.58%). This suggests that Focal Loss's focus on hard examples is unnecessary for this task; class weighting alone is sufficient. This finding aligns with recent work showing that loss function sophistication does not universally improve performance on imbalanced datasets [14].
4. Oversampling Ineffective: Synthetic oversampling of rare classes (PROPAGANDA, NOISE) slightly decreases performance for mBERT (80.84% vs 84.58% baseline) and provides minimal improvement for XLM-R. Standard class weighting is superior.
5. mBERT > XLM-RoBERTa: Despite XLM-R's larger model size and stronger cross-lingual pretraining, mBERT achieves better performance on this task. The best XLM-R configuration (82.71% accuracy, 7 epochs + Focal Loss + oversampling) underperforms the best mBERT configuration (84.58%). This suggests that mBERT's pretraining is more aligned with the linguistic patterns in this dataset, possibly due to stronger pretraining on Slavic languages (Ukrainian, Russian) [11].
6. PROPAGANDA Challenge: Even the best model achieves PROPAGANDA F1 of 0.63, significantly lower than NEWS (0.87) and ANALYSIS (0.86). PROPAGANDA is consistently the most difficult class to classify. This reflects the semantic challenge identified in error

analysis (Section 8): propaganda often shares factual elements with news and analysis, differing primarily in distorted causal reasoning or misleading contextualization – subtle distinctions not fully captured by shallow language models. Comparison mBERT (best configuration) with baseline approach shows substantial improvement over TF-IDF baseline: accuracy +5.78 percentage points (78.80% ⟶ 84.58%), macro F1 +0.2245 (0.5808 ⟶ 0.8053), PROPAGANDA F1 +0.16 (0.47 ⟶ 0.63, +34%).

Table 1
Ablation Study Results

Model	Epochs	Method	Accuracy, %	Macro F1	NEWS	ANALYSIS	PROPAGANDA	NOISE
mBERT	5	Baseline	84.58	0.8053*	0.87	0.86	0.63*	0.86
XLM-R	7	Focal + Oversampling	82.71	0.7777	0.87	0.84	0.52	0.88
XLM-R	7	Baseline	81.78	0.759	0.87	0.8	0.51	0.85
mBERT	5	Focal + Oversampling	81.78	0.7482	0.84	0.86	0.42	0.87
mBERT	10	Baseline	80.84	0.7137	0.84	0.82	0.32	0.88
mBERT	3	Baseline	78.04	0.7249	0.84	0.81	0.44	0.81
mBERT	5	Focal Loss	74.77	0.7018	0.77	0.84	0.36	0.84
mBERT	5	Oversampling	80.84	0.7148	0.84	0.83	0.31	0.88
XLM-R	7	Focal Loss	71.96	0.6945	0.77	0.82	0.37	0.82

1 summarizes results which are presented in descending order by Macro F1 score for transformer models mBERT and XLM-RoBERTa across 9 configurations (ablation study), varying: number of fine-tuning epochs (3, 5, 7, 10); loss function (standard class-weighted cross-entropy, Focal Loss); class imbalance strategy (baseline, oversampling, Focal Loss + oversampling); '*' - best performance.

The largest improvement is on PROPAGANDA, indicating that transformer-based contextual embeddings capture propaganda-specific patterns better than TF-IDF. Per-Class Performance comparison shows that mBERT with 5 epochs has the best results:

- NEWS: 0.87 (highest confidence class; clear linguistic markers for factual reporting);
- ANALYSIS: 0.86 (strong performance; interpretation is linguistically marked with causal connectives, conditional structures);
- PROPAGANDA: 0.63 (difficult; requires detecting subtle distortions in factual content);
- NOISE: 0.86 (spam and off-topic content are linguistically distinct).

The inverse relationship between class difficulty and F1 score reflects semantic properties: classes with distinct linguistic markers (NEWS, NOISE) are easier to classify, while classes distinguished by subtle semantic properties (PROPAGANDA vs NEWS/ANALYSIS boundary) are more difficult.

7. Information processing pipeline

The classification model (trained in Sections 3–6) is deployed within a comprehensive information processing pipeline designed for real-time, multilingual stream classification. The following figure shows the complete architecture, organized in five stages: ingestion, stream processing, classification, storage, and analytics.

On the Information Sources and Aggregation stage - information arrives from diverse sources (Telegram, Facebook, X, news websites) and is aggregated via native APIs. For Telegram, the TDLib library enables reliable message ingestion with metadata (timestamp, channel, language).

On the Stream Processing Pipeline stage incoming messages are queued in Apache Kafka (decoupling ingestion from processing) and processed via Apache Spark Streaming [3]. Processing steps include text normalization, metadata extraction, tokenization, and feature engineering (BERT embeddings).

The Content Classifier stage applies the fine-tuned mBERT model to each message, producing class predictions and confidence scores. Messages with confidence ≥0.85 pass directly to storage; lower-confidence predictions are flagged for human review.

On the Data Storage and Analytics stage results are stored in two complementary systems: MinIO Data Lake (persistent storage for raw data, classification results, and historical context) and ElasticSearch (indexed, searchable storage enabling real-time queries). A dashboard (Grafana/Kibana) consumes both layers to visualize stream composition, classification results, and detection outcomes.

Secondary Analysis Layer (Future Work): High-confidence classifications are passed to a reasoning-based MANIPULATION detection layer accessing the MinIO data lake for cross-document analysis. This offline layer detects cases where accurate facts are combined with distorted causal reasoning—a task requiring contextual analysis across multiple documents and topics.

The designed architecture is cost-efficient and cloud-agnostic. The system uses an entirely open-source stack (Apache Spark, Kafka, MinIO, ElasticSearch, Grafana) deployed via Docker and Kubernetes. This design eliminates licensing costs and vendor lock-in, enabling deployment on commodity hardware or modest cloud resources without code changes.

A minimal setup requires 1 master node (8 CPU, 32 GB RAM) and 2–4 worker nodes (8 CPU, 16 GB RAM each); production deployment scales horizontally. Optional GPU acceleration (RTX 3060, RTX 4090) accelerates BERT inference. Real-time processing achieves < 5 seconds ingestion latency, ~100–200 ms processing latency per message, and < 10 seconds end-to-end latency—acceptable for most analytical workflows.

The architecture separates primary and secondary analysis: primary classification operates on individual messages in real-time (< 500 ms latency), producing class predictions and confidence scores; secondary MANIPULATION detection operates offline on high-confidence classifications, accessing the data lake for historical context. This separation maintains rapid response times for primary classification while dedicating resources to complex reasoning tasks.

Thus, the architecture (1) integrates five stages: 1 - information ingestion from diverse sources via APIs (Telegram, TDLib); 2 - message queuing in Apache Kafka to decouple ingestion from processing; 3 - stream processing with Apache Spark Streaming for text normalization and feature extraction; 4 - classification with mBERT and confidence filtering; 5 - storage in MinIO Data Lake and ElasticSearch. A secondary MANIPULATION detection layer operates offline, accessing MinIO for cross-document analysis.

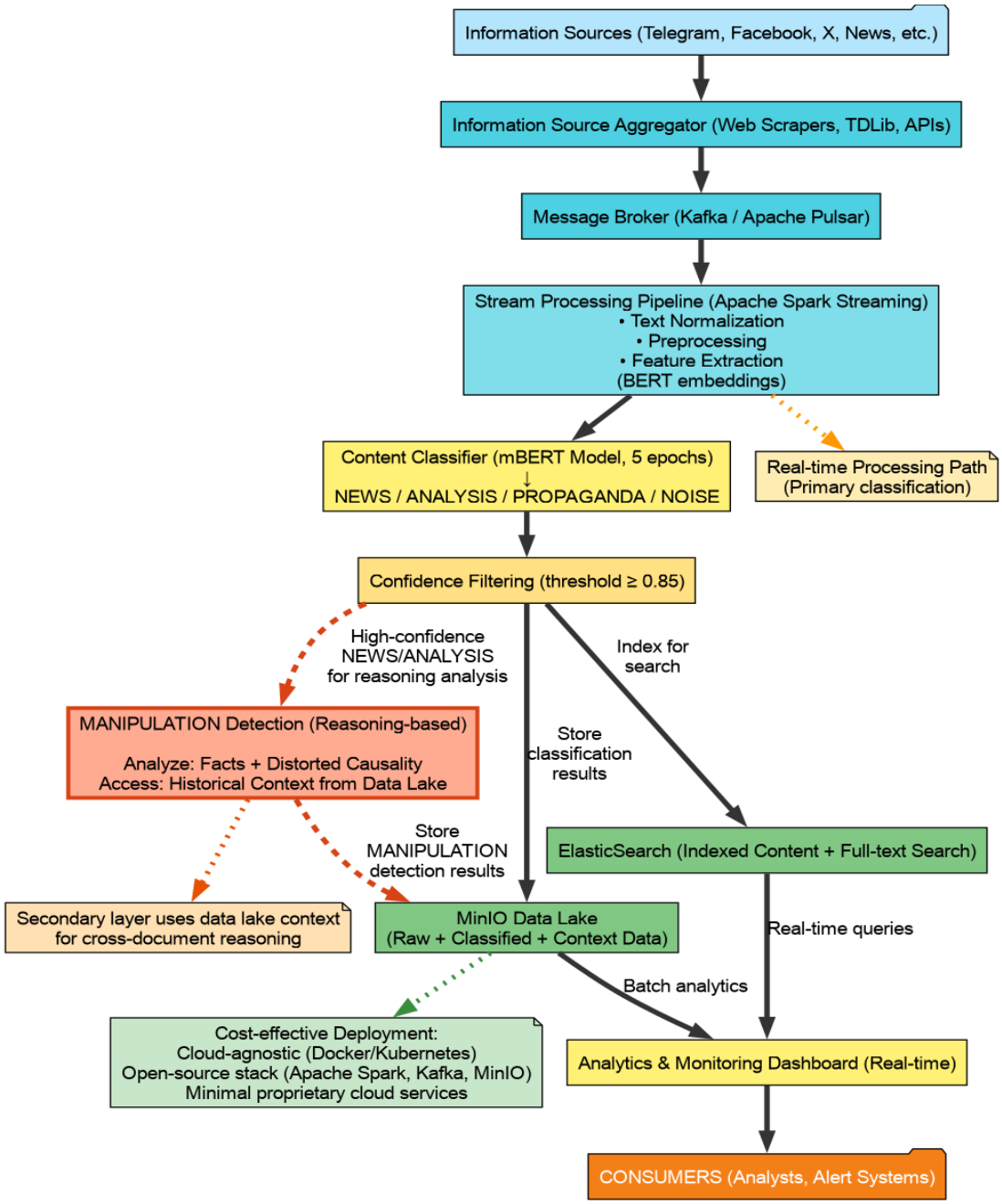

Figure 1: Information System Architecture for Multilingual Stream Classification

Continuous monitoring tracks classification metrics, system health, and model accuracy. As new labeled data accumulates, the model is automatically retrained (weekly or monthly) via a continuous integration pipeline that includes data collection, training, validation, A/B testing, and deployment.

The system integrates extensions seamlessly: additional language-specific models, new information sources, secondary layers, and custom analytics queries all integrate without major infrastructure changes. This design demonstrates that cost-effective, cloud-agnostic deployment of multilingual text classification at scale is achievable using open-source components while prioritizing robustness, maintainability, and extensibility.

8. Conclusions

This paper presented a four-class classification system (NEWS, ANALYSIS, PROPAGANDA, NOISE) for multilingual Ukrainian/Russian/English information streams in the Russian-Ukrainian war context.

RQ1 – Annotation framework. Three rounds of guideline refinement raised inter-rater agreement from κ=0.686 to κ=0.850 (+24%). Final per-class agreement: NEWS 93.2%, NOISE 90.5%, ANALYSIS 88.3%, PROPAGANDA 87.5%. Defining boundary cases between emotionally framed factual content and propaganda proved essential for conflict-context annotation.

RQ2 – Transformer benchmarking. mBERT fine-tuned for 5 epochs with class-weighted cross-entropy loss achieves 84.58% accuracy and 0.805 Macro F1, outperforming TF-IDF by 5.78 pp and 0.2245 Macro F1. XLM-RoBERTa, Focal Loss, and oversampling all underperform, confirming that class weighting alone is sufficient.

RQ3 – Error analysis. Three fundamental challenges underlie the 15.4% error rate: NEWS-ANALYSIS boundary ambiguity (42% of errors), propaganda grounded in factually accurate content (31% of PROPAGANDA errors), and length-dependent degradation (18.1% on texts >500 characters). PROPAGANDA F1 of 0.63 is insufficient for unsupervised deployment; a multi-stage pipeline with secondary reasoning-based validation is required.

The proposed open-source architecture (Kafka, Spark Streaming, MinIO, ElasticSearch) achieves <10 s end-to-end latency. The dataset (n=1,066, 7 Telegram channels, narrow temporal window) limits generalization. Future work targets reasoning-based MANIPULATION detection, language-specific models, and extended temporal evaluation.

9. Limitations and future work

The dataset (n=1066, September 2025–February 2026) is modest by deep learning standards and collected over a narrow temporal window, limiting generalization claims. The MANIPULATION class (n=14, 1.3%) was excluded due to severe imbalance; detecting manipulation requires distinct reasoning-based methods combining factual accuracy analysis with causal reasoning, reserved for future work. Ground truth validation beyond inter-rater agreement would strengthen claims about class reliability. Future work includes: reasoning-based MANIPULATION detection integrating data lake context, language-specific models for improved per-language F1, and extended temporal evaluation to assess propaganda pattern evolution.

Declaration on Generative AI

This research was conducted with assistance from Claude (Anthropic), a large language model. The model was used for generating initial drafts, grammar and spelling check. All research design, data annotation, annotation guidelines, experimental methodology, model training, and interpretation of results were conducted by the authors.

References

[1] Anthony, L., A critical look at software tools in corpus linguistics, Linguistic Research 30(2) (2013) pp. 141–161.

[2] McEnery, T., Hardie, A., Corpus linguistics: Method, theory and practice, Cambridge University Press, 2012.

[3] Lutskiv, A., Popovych, N., Big data approach to developing adaptable corpus tools, in: Computational Linguistics and Intelligent Systems. Proceedings of the 4th International Conference COLINS 2020, volume 2604, CEUR-WS.org, 2020, pp. 374–395. URL: https://ceur-ws.org/Vol-2604/paper28.pdf

[4] Rashkin, H., Choi, E., Jang, J.Y., Volkova, S., Choi, Y., Truth of Varying Shades: Analyzing Language in Fake News and Political Fact-Checking, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP), ACL, 2017, pp. 2931–2937.
[5] Da San Martino, G., Yu, S., Barrón-Cedeño, A., Petrov, R., Nakov, P., Fine-Grained Analysis of Propaganda in News Article, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), ACL, 2019, pp. 5636–5646.
[6] Vysotska, V., Przystupa, K., Kulikov, Yu., Chyrun, S., Ushenko, Yu., Hu, Z., Uhryn, D., Recognizing Fakes, Propaganda and Disinformation in Ukrainian Content based on NLP and Machine-learning Technology, International Journal of Computer Network and Information Security (IJCNIS) 17(1) (2025) pp. 92–127.
[7] Leite, J.A., Razuvayevskaya, O., Bontcheva, K., Scarton, C., EUvsDisinfo: A Dataset for Multilingual Detection of Pro-Kremlin Disinformation in News Articles, in: Proceedings of the 33rd ACM International Conference on Information and Knowledge Management (CIKM), ACM, 2024. doi:10.1145/3627673.3679167.
[8] Darczewska, J., The anatomy of Russian information warfare: The Crimean operation, a case study, Point of View 52, Centre for Eastern Studies, Warsaw, 2014.
[9] Devlin, J., Chang, M.-W., Lee, K., Toutanova, K., BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) (NAACL-HLT), ACL, 2019, pp. 4171–4186.
[10] Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, É., Schwenk, H., Unsupervised Cross-lingual Representation Learning at Scale, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL), ACL, 2020, pp. 8440–8451.
[11] Wu, S., Dredze, M., Beto, Bentz, Becas: The Surprising Cross-Lingual Effectiveness of BERT, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), ACL, 2019, pp. 833–844.
[12] Hu, J., Ruder, S., Siddhant, A., Neubig, G., Firat, O., Hall, M., XTREME: A Massively Multilingual Multi-task Benchmark for Evaluating Cross-lingual Generalization, in: Proceedings of the 37th International Conference on Machine Learning (ICML), PMLR, Vol. 119, 2020. URL: https://proceedings.mlr.press/v119/hu20b.html
[13] Mosbach, M., Andriushchenko, M., Klakow, D., On the stability of fine-tuning BERT: Misconceptions, explanations, and strong baselines, in: Proceedings of the 11th Workshop on Representation Learning for NLP (RepL4NLP), 2020.
[14] Lin, T.-Y., Goyal, P., Girshick, R., He, K., Dollár, P., Focal loss for dense object detection, in: IEEE International Conference on Computer Vision (ICCV), 2017.
[15] Krippendorff, K., Content analysis: An introduction to its methodology, 2nd ed., SAGE Publications, 2004.
[16] Artstein, R., Poesio, M., Inter-coder agreement for computational linguistics, Computational Linguistics 34(4) (2008) 555–596.
[17] Cohen, J., A coefficient of agreement for nominal scales, Educational and Psychological Measurement 20(1) (1960) 37–46.
[18] Lutskiv, A., Lutsyshyn, R., Corpus-based translation automation in adaptable corpus translation module, in: Computational Linguistics and Intelligent Systems. Proceedings of the 5th International Conference COLINS 2021, volume 2870, CEUR-WS.org, 2021, pp. 511–527. URL: https://ceur-ws.org/Vol-2870/paper38.pdf