EPJ Data Science

https://doi.org/10.1140/epjds/s13688-026-00648-z

Article in Press

Dual-input deep learning–based approach for propaganda narratives detection in a low-resource language: a case study in Lithuanian

Received: 13 October 2025

Accepted: 17 March 2026

Published online: 26 March 2026

Cite this article as: Rizgelienė, I., Marcinkevičius, V., Plikynas, D. Dual-input deep learning–based approach for propaganda narratives detection in a low-resource language: a case study in Lithuanian. *EPJ Data Science* (2026). https://doi.org/10.1140/epjds/s13688-026-00648-z

Ieva Rizgelienė, Virginijus Marcinkevičius, Darius Plikynas

Dual-Input Deep Learning–Based Approach for Propaganda Narratives Detection in a Low-Resource Language: A Case Study in Lithuanian

Ieva Rizgelienė[1*†], Virginijus Marcinkevičius[1] and Darius Plikynas[1]

[1*]Institute of Data Science and Digital Technologies, Vilnius University, Akademijos g. 4, Vilnius, 08412, Lithuania.

*Corresponding author(s). E-mail(s): ieva.rizgeliene@mif.stud.vu.lt, paulaviciute.ieva@gmail.com;
Contributing authors: virginijus.marcinkevicius@mif.vu.lt; darius.plikynas@mif.vu.lt;
†The main author of this paper is Ieva Rizgelienė

Abstract

Propaganda narratives play a central role in disseminating disinformation and are tailored to each targeted country's context and native language. This highlights the pressing need for automated systems capable of detecting and analyzing such narratives, particularly in low-resource languages, where natural language processing tools and data resources remain limited. In this study, we introduce the first supervised machine learning system for identifying pro-Kremlin propaganda narratives in Lithuanian, a low-resource language spoken in a country targeted by Kremlin disinformation. To our knowledge, this represents the first such approach not only for Lithuanian but also for other languages in Russia's broader neighborhood. Our method employs a novel dual-input architecture that significantly outperforms single-input baselines across all analyzed narratives and even surpasses the performance of ChatGPT-5. Although our study focuses on Lithuanian, our findings and methodology are applicable to other low-resource languages, offering practical guidelines for extending propaganda-narrative detection globally.

Keywords: transformers, hybrid approach, low-resource language, propaganda, deep-learning, narratives detection

1 Introduction

The deliberate spread of malign propaganda and disinformation has become a critical security concern in recent years. Drawing on insights from more than 900 experts, the Global Risks Report 2025 identifies misinformation and disinformation among the most severe short-term risks [1]. It is well established that Russia disseminates propaganda globally and is a pioneer in the use of disinformation, notably employing troll farms and bot networks [2–4]. Russian propaganda campaigns are characterized by a sophisticated, multi-pronged ecosystem of information channels [3, 4]. This multifaceted approach enables the widespread dissemination of disinformation, with coordinated influence campaigns targeting Western democracies, neighboring countries, and entire regions [4, 5].

Propaganda narratives are central to Russia's hybrid-warfare strategy [4–6]. Russia promotes recurring narratives that depict the West as hostile, present Russian-speaking communities as part of a unified civilizational space, question the legitimacy of post-Soviet borders, reinterpret history to strengthen national identity, and frame Russia as a defender of traditional values against Western influence [7].

For example, in the Baltic states, Kremlin propaganda exploits historical and demographic sensitivities, alleging pervasive "Russophobia" and systemic discrimination against Russian-speaking minorities, denying the Soviet occupation, and portraying the EU and NATO members Lithuania, Latvia, and Estonia as "failed states" [8]. Beyond the Baltic states, similar narratives are tailored to other countries' local contexts: Ukraine is cast as a Western-run "failed state" and Russian aggression is framed as self-defence [9]; Belarus is portrayed as besieged by Western "colour revolutions" [10]; and Moldova is depicted as imperilled by European integration [11]. In the South Caucasus, Western alignment is portrayed as a threat to national identity in Georgia, Russian influence is presented as a stabilising force in Azerbaijan, and Armenia's security dependence on Moscow is emphasised [12]. In Central Asia, Soviet nostalgia and appeals to "Eurasian unity" are invoked as responses to contemporary challenges [13].

The Role of Native Languages

To maximize its reach and impact, the Kremlin deliberately produces content in local languages, using local websites, television, and social media, to seamlessly integrate into domestic information environments and engage audiences beyond Russian-speaking populations [4]. By adapting its messages to each country's specific context and employing local languages, Kremlin ensures its narratives appear organic. This approach makes propaganda more credible and appealing to target populations.

For example, in the Baltic states, Russian propaganda is disseminated in both Russian and the national languages. In Latvia, for instance, the Kremlin-funded outlet Sputnik operated separate Latvian-language and Russian-language versions, curating stories differently for Latvian and Russian speaking audiences and tailoring content to each group's perspective [14]. This demonstrates that Moscow invests in local-language media to influence not only ethnic Russians, but also the broader population in the Baltic states. Similarly, Lithuania has faced propaganda in Lithuanian via local portals and social media groups promoting pro-Kremlin narratives [15].

Beyond the Baltic states, similar strategies are adapted to local contexts. In Ukraine, disinformation is circulated through Russian- and Ukrainian-language channels to reach broad segments of society [16]. In Moldova, Kremlin-aligned narratives frequently masquerade as native Moldovan media and circulate in Romanian [11], while in Georgia they are routed through domestic Georgian-language outlets [17].

2 Related Work

Early studies of propaganda identification focused on article-level classification and used classical machine-learning models with handcrafted features (e.g., TF-IDF, n-grams) [18, 19]. With the advent of deep learning, the research shifted to transformer-based approaches [20] and extended to document-level classifiers beyond English (e.g., Hindi [21], Urdu [22], Arabic [23]). Building on this document-level line of work, SemEval-2020 Task 11 formalized fine-grained propaganda-technique detection [24], shifting the focus from document-level classification to identifying propagandistic spans and assigning technique labels. In parallel with article-level identification, technique-detection studies have expanded beyond English, including Arabic [25], Spanish [26], and Czech [27], using approaches that range from classical baselines to monolingual and multilingual transformer models.

While propaganda-technique detection and article-level labeling are predominantly supervised-learning problems, the identification of propaganda narratives has largely relied on unsupervised methods. An early, pioneering application was the use of Latent Dirichlet Allocation (LDA) to uncover the main themes in political discourse [28]. Since then, LDA has been widely adopted for narrative analysis across domains, including studies of terrorist propaganda [29], automated analyses of Islamist extremist propaganda [30], climate-change debates [31], and media coverage of the Russia-Ukraine war to reveal underlying propaganda narratives [32]. Beyond LDA, researchers used Latent Semantic Analysis (LSA) to profile fake-news narrative patterns during the 2016 U.S. election [33] and applied Non-negative Matrix Factorization (NMF) to extract ideological and narrative dimensions in Russia–Ukraine media [34]. More recently, embedding-based topic modeling has emerged. By integrating contextual embeddings from transformer encoders with clustering algorithms, researchers can identify narratives more accurately. This approach has already been applied to tasks such as analyzing Telegram channels and multilingual news in the Russia–Ukraine information space [35].

Unsupervised methods are attractive for propaganda narrative analysis because they do not require labeled data, enabling rapid discovery of latent themes directly from text [36, 37]. Because clustering methods fundamentally rely on word co-occurrence, they primarily capture topical similarity rather than manipulative intent or argumentative structure [38, 39]. As a result, their outputs require substantial human interpretation and validation [40, 41]. In practice, unsupervised methods are best for initial textual corpus exploration, especially when paired with expert review for follow up tasks [42].

Given the limitations of unsupervised methods, supervised learning has attracted increasing attention for detecting propaganda narratives, with most progress occurring in recent years. In the COVID-19 context, for example, researchers developed BERT-based classifiers for vaccine narratives [43, 44], also some research extent to multilingual systems, such as an approach trained on text in 24 languages - via machine translation into English, were used to track disinformation across countries over a three - year period [45]. Extending beyond COVID-19, SemEval 2025 Task 10 introduced a multilingual shared task on identifying narratives in online news for two different contexts: the Ukraine–Russia War and Climate Change covering Bulgarian, English, Hindi, Portuguese, and Russian [46].

Motivation

Despite recent advances, automated detection of propaganda narratives remains underexplored, particularly in low-resource languages and even in the local languages of countries that are primary targets of Kremlin propaganda. Progress is constrained by the scarcity of datasets and by limited NLP infrastructure for many of these languages. Methodologically, commonly used unsupervised methods mainly reveal topical similarity based on words that frequently appear together, rather than identifying the manipulative content used to disseminate specific propaganda narratives.

Focusing on the Baltic countries, the NATO Strategic Communications Center of Excellence conducted a study [47] on narrative identification in Lithuanian, Latvian, and Estonian, surveying the tools and datasets commonly available in academic settings. The study found that, across all three languages, narrative detection remains largely unexplored. It also identified a methodological gap: pipelines developed for English tend to underperform in these languages, and the authors recommend using multilingual embeddings tailored to the region, such as LitLatBERT. More broadly, the report notes that the languages of the Baltic countries are underrepresented in academic resources due to persistent data and tooling shortages. Limited NLP capacity, therefore, continues to impede progress, underscoring the need for targeted data creation, tool development, and model adaptation for narrative detection in the Baltics.

To address the gap, we propose the first approach for detecting hostile propaganda narratives in Lithuanian, a low-resource language spoken in a country that is a primary target of Russian disinformation. To the best of our knowledge, this is also the first model of its kind in the Baltic region. More broadly, we are not aware of any prior approach of this type across Russia's wider neighborhood. We base the system design on three main research questions:

- How do fine-tuning strategies affect performance in low-resource narrative detection?
- Does restricting inputs to sentences containing propaganda techniques improve performance compared to using full articles?
- Do dual-input hybrid models outperform single-encoder baselines?

3 Methodology

3.1 Task Formulation

Propaganda comprises deliberate communication strategies intended to shape public opinion through targeted topics, narratives, and techniques. A *topic* denotes the core issue under discussion, whereas a *narrative* is a particular framing or interpretation of that topic, typically articulated through manipulative content and reinforced by propaganda techniques. Because a single article may express multiple narratives, even expert analysts must consider the article's full textual context to determine which topics are covered and whether specific propaganda techniques are used to frame those topics and influence public opinion. Building on this distinction, we develop dual-input models, where different branches are used for processing different inputs, such as full textual content and only propaganda sentences. These branches have untied (non-shared) parameters, are trained jointly, and their representations are fused and passed to a classifier, letting the system combine global context with technique-focused cues for more reliable narrative detection.

We model each narrative with an independent binary classifier trained under a one-vs-rest scheme: articles labeled with a given narrative serve as positives and all remaining articles as negatives. Each classifier outputs a binary decision for a given article, where 1 indicates that the specific narrative is present and 0 indicates that it is absent. Since seven narratives are modeled independently, the final prediction for an article consists of a multi-label vector composed of seven binary outputs. This design addresses the severe class imbalance and the partial overlap among narratives in our corpus, reducing cross-narrative interference that would otherwise hinder joint training [48]. To respect the fixed maximum sequence length of transformer encoders, long articles are split into paragraph-level chunks. Because annotations are available only at the article level, each paragraph inherits its article's labels, yielding weak supervision: the paragraph label indicates that the source article expresses the narrative somewhere, not necessarily that the paragraph itself contains it.

As our approach must operate on any Lithuanian text without pre-existing annotations, we also include in our framework an auxiliary binary sentence-level propaganda detector. At inference time, this detector identifies candidate sentences exhibiting propaganda techniques within each textual unit; these sentences form the technique-focused input to the second block, enabling end-to-end operation without manual markup.

3.2 Data

We use HALT-PROP [49], the first human-annotated Lithuanian corpus of propaganda narratives and techniques. The dataset consists of 1,000 news articles from multiple Lithuanian media outlets (2018–2024), selected to capture pro-Russian propaganda targeting Lithuania.

The 1,000 articles were processed using a two-level annotation procedure. First, a larger pool of articles was screened and labeled at the article level to determine

whether an article contained propagandistic content. Second, the articles identified as propagandistic were selected for detailed analysis.

For the selected articles, narratives were labeled at the article level, meaning that annotators assigned one or more narrative categories to the entire article without marking specific text spans. In contrast, propaganda techniques were annotated at the span level, where specific text fragments were marked and assigned technique labels.

Five trained annotators followed a cross-annotation protocol: each article was independently labeled and annotated by two annotators, and disagreements were resolved through pairwise discussion to produce the final gold-standard annotation.

Label overlap was allowed for both narratives and techniques. An article may contain more than one narrative label, and annotated spans corresponding to propaganda techniques may overlap with one another. Figure 1 presents an example of an annotated article. On the left, the labels indicate whether the article contains propaganda and which narratives are assigned at the article level, while the main text is annotated with span-level labels for propaganda techniques.

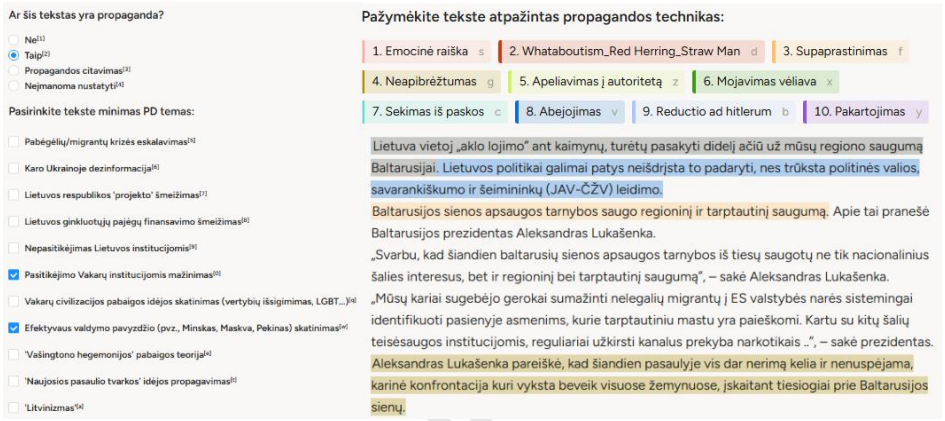

Fig. 1 Annotated article example [49]

3.2.1 Narratives Distribution

The corpus covers eleven narratives in total:

1. Disinformation about the war in Ukraine (i.e. spreading false narratives to justify Russia's aggression and delegitimize Ukrainian resistance);
2. Delegitimization of the Lithuanian State (i.e. slandering the Republic of Lithuania as a failed or artificial "project," questioning its sovereignty and historical foundations);
3. Undermining the Lithuanian Armed Forces (i.e. attacks on military funding, modernization efforts, and NATO deployments, aiming to portray Lithuania as militaristic or provocatively anti-Russian);
4. Erosion of Trust in Lithuanian Institutions (i.e. promoting narratives that depict state authorities as cor rupt, incompetent, or unrepresentative);

5. Attacks on Western Institutions and Alliances (i.e. discrediting the EU, NATO, and other multilateral bod ies, framing them as exploitative, ineffective, or morally bankrupt);
6. Decline of Western Civilization (i.e. spreading claims about Western moral decay, often emphasizing themes like gender ideology, LGBT rights, or secularism, to contrast it with "traditional values");
7. Authoritarian Model Promotion (i.e. highlighting regimes like Moscow, Minsk, or Beijing as examples of stability, efficiency, and sovereign governance in contrast to Western "chaos");
8. Narratives of US Decline and "Washington Hegemony" (i.e. framing the United States as a waning imperial power and suggesting the emergence of a multipolar or "New World Order" led by Russia and China);
9. Geopolitical Reordering and the "New World Order" (i.e. promoting conspiracy-laden ideas about a global realignment that replaces liberal democratic systems with alternative authoritarian alliances);
10. Weaponization of Migration and Refugees (i.e. amplifying fears around migrant flows and depicting them as tools of hybrid warfare or existential threats to national identity and security);
11. Revival of "Litvinism" (i.e. exploiting historical revisionism to push the narrative that parts of Lithuania historically belonged to Belarus, undermining the Lithuanian national identity and territorial integrity).

Figure 2 presents the distribution and overlap of propaganda narratives across the 1,000 annotated articles. Panel (A) shows the overall distribution of narratives based on article counts, distinguishing between single-narrative and multi-narrative occurrences, and indicating how many articles containing a specific narrative also overlap with other narratives. Panel (B) provides a more detailed analysis of pairwise narrative overlap, illustrating with which specific narratives each narrative co-occurs. It is important to note that in Panel (B), each row represents the number of articles containing a given narrative that also include another specific narrative. Consequently, row or column totals do not equal the total number of articles for that narrative, as a single article may overlap with multiple narratives. For example, in the case of *Revival of "Litvinism"*, there are three articles in total that contain this narrative. Of these, two also overlap with the narrative *Erosion of Trust in Lithuanian Institutions*, one overlaps with *Attacks on Western Institutions and Alliances*, and so forth.

In general, Figure 2 clearly demonstrates that the dataset is highly imbalanced from the perspective of narrative distribution. One dominant narrative, *Erosion of Trust in Lithuanian Institutions*, appears in 514 out of 1,000 articles (approximately half of the dataset) and is the only narrative that can be considered relatively balanced. The second most frequent narrative, *Attacks on Western Institutions and Alliances*, has a substantially smaller share, representing 17.2% of the dataset. All remaining narratives occur even less frequently.

The results also show that narratives tend to overlap extensively. The lowest overlap rate is observed for *Erosion of Trust in Lithuanian Institutions*, where approximately 42% of the articles containing this narrative also include at least one additional narrative. For the remaining narratives, overlap rates are even higher, ranging from

67% to 100%. These findings indicate a very high degree of co-occurrence among narratives.

The narrative overlap matrix further illustrates which narratives tend to co-occur. For example, articles containing *Erosion of Trust in Lithuanian Institutions* frequently also express *Delegitimization of the Lithuanian State*, *Decline of Western Civilization*, and *Attacks on Western Institutions and Alliances*. Similarly, articles that include *Disinformation about the War in Ukraine* often co-occur with *Attacks on Western Institutions and Alliances* and *Erosion of Trust in Lithuanian Institutions*, among others.

Given the highly imbalanced narrative distribution and the fact that several narratives occur too rarely for reliable modeling, we restrict supervised narrative detection to the seven most prevalent categories highlighted in Figure 2. Although the binary classification framework was introduced earlier, the distributional and overlap analysis presented here further supports this modeling choice. In a multiclass setting, the strong overlap between narratives would violate the assumption of mutual exclusivity and could limit the model's ability to learn narrative-specific features effectively. These findings therefore provide additional empirical support for modeling each narrative as an independent binary classification task.

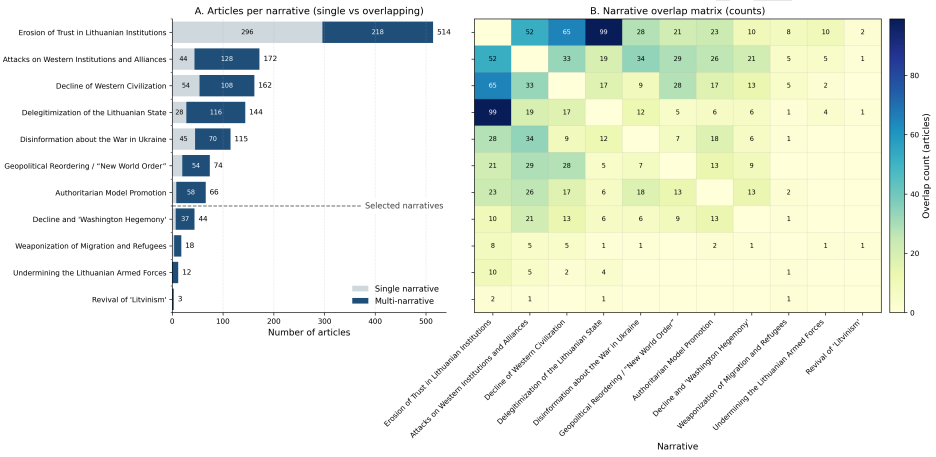

Fig. 2 Prevelance and Pairwise Overlap of Propaganda Narratives. A) Article-level distribution of narratives, distinguishing single-narrative and multi-narrative occurrences. B) Pairwise co-occurrence matrix showing the number of articles in which narratives appear together.

Sentence-Level Distribution

Additionally, we examined the sentence-level distribution of the dataset, as a separate classifier was developed to detect whether individual sentences contain propaganda techniques. Using the same 1,000 annotated articles, texts were segmented into sentences and labeled as propaganda (1) if they included any annotated technique span, and non-propaganda (0) otherwise. This process resulted in 32,078 sentences, with a relatively balanced distribution of 55.3% positive and 44.7% negative instances.

3.2.2 Data Preparation

Because transformer encoder models have a fixed maximum input length, articles that exceeded this limit were split into paragraph-level chunks that fit within the model's token budget. We define two input regimes (see Figure 3 for the article-level split): (1) *prop.*, where only sentences containing at least one annotated technique are retained (if any part of a sentence is tagged, the entire sentence is kept); and (2) *all*, which uses the full text. Paragraph boundaries are identical across both regimes: the *all* condition preserves every sentence, whereas the *prop.* condition masks non-propaganda sentences within the same paragraphs.

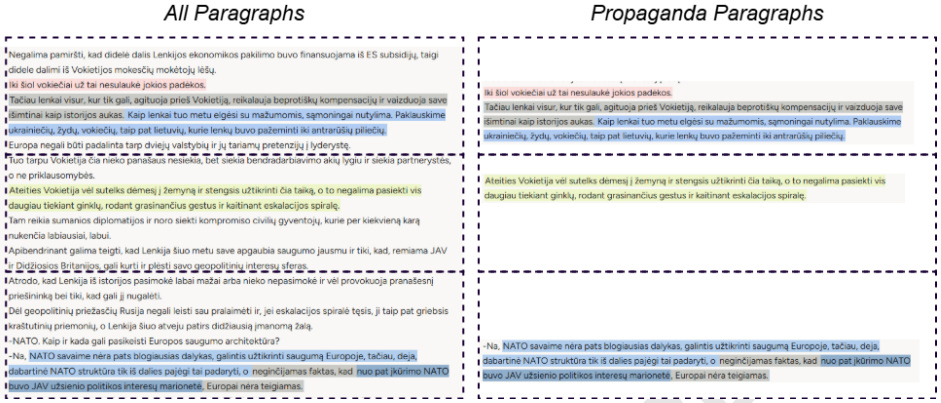

Fig. 3 All vs. propaganda-only paragraphs within an article.

In addition to preparing inputs for the two regimes, we convert each narrative label to a binary indicator and propagate the article-level labels to the paragraph level (weak supervision). For the auxiliary sentence-level propaganda-detection task, we split each article into sentences and assign a binary label to every sentence using expert annotations of propaganda techniques. The full data-preparation pipeline is shown in Figure 4.

3.3 Models

Multilingual Transformers

For context-sensitive tasks such as propaganda detection, long-range dependencies and subtle semantic relations matter. Transformer encoders are effective here because multi-head self-attention builds rich *contextual* token representations. We fine-tune two RoBERTa-base–style multilingual encoders as binary classifiers: XLM-RoBERTa and LitLatBERT. XLM-RoBERTa uses a 250k SentencePiece vocabulary and is pre-trained with masked language modeling on CC-100 ($\approx$ 2.5 TB across 100 languages); the Lithuanian portion is $\approx$ 13.7 GiB ($\approx$ 1.84 B tokens) [50–52]. LitLatBERT is a trilingual encoder (Lithuanian, Latvian, English) released by EMBEDDIA with an 84,201-token SentencePiece vocabulary and $\approx$ 4.06 B pre-training tokens in total (1.21 B LT, 0.53 B LV, 2.32 B EN) [53, 54]. Both models use the base configuration (12

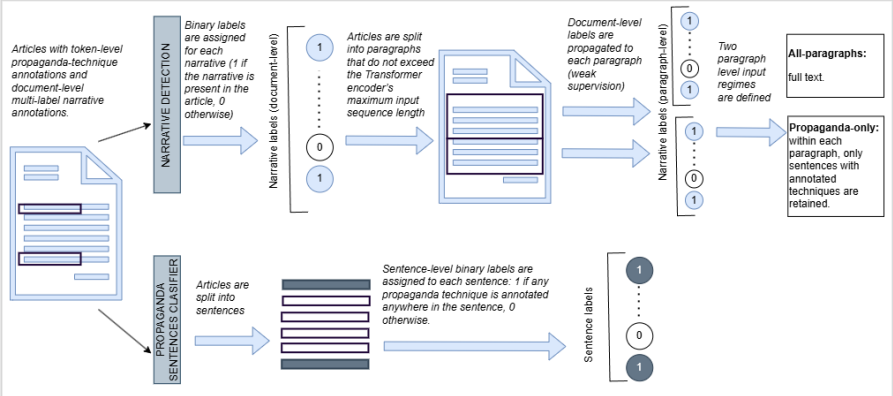

Fig. 4 Data-Preparation Pipeline

layers, hidden size 768, 12 attention heads, intermediate size 3072); XLM-RoBERTa has ≈ 270M parameters and LitLatBERT ≈ 150.8M due to the smaller vocabulary [50, 53]. Together they contrast a broad multilingual encoder with a Baltic-focused alternative highlighted for regional narrative tasks [47].

Single-Input Case

Figure 5 illustrates the single-input encoder-only architecture for binary classification. The Lithuanian input is tokenized into subwords and wrapped with special tokens: `<s>` at position 0 and `</s>` at the end. If the tokenized sequence is shorter than 512 tokens, `<pad>` tokens are appended. At each position t, the input vector is the sum of a token embedding and a learned positional embedding. The sequence is encoded by 12 transformer layers. Each layer applies 12-head self-attention followed by a position-wise feed-forward network (intermediate size 3072); both sublayers use residual connections, dropout, and layer normalization, and an attention mask blocks interactions with padding tokens. From the final layer we take the hidden state of the first special token (`<s>`, position 0) as a fixed-length sequence representation. A standard RoBERTa classification head maps this representation to logits as follows: (i) apply dropout to the `<s>` representation; (ii) apply a linear projection to an intermediate representation; (iii) apply a tanh nonlinearity; (iv) apply dropout again; and (v) apply a second linear projection to produce two logits. A softmax over these logits yields class probabilities (0 = no narrative, 1 = narrative).

Dual-Input Case

Figure 6 illustrates our proposed hybrid dual-input architecture, which takes two inputs: the full text and the set of propaganda sentences. We evaluate three variants: XLM-RoBERTa Dual, LitLatBERT Dual, and XLM - LitLat Multi-Embedding. The XLM-RoBERTa Dual and LitLatBERT Dual models, therefore, have two branches, whereas the XLM–LitLat Multi-Embedding model has four. In every branch, the contextual embedding is the embedding of the encoder final-layer hidden state at position 0, i.e., the model's special classification token (`<s>`). (See Figure 5)

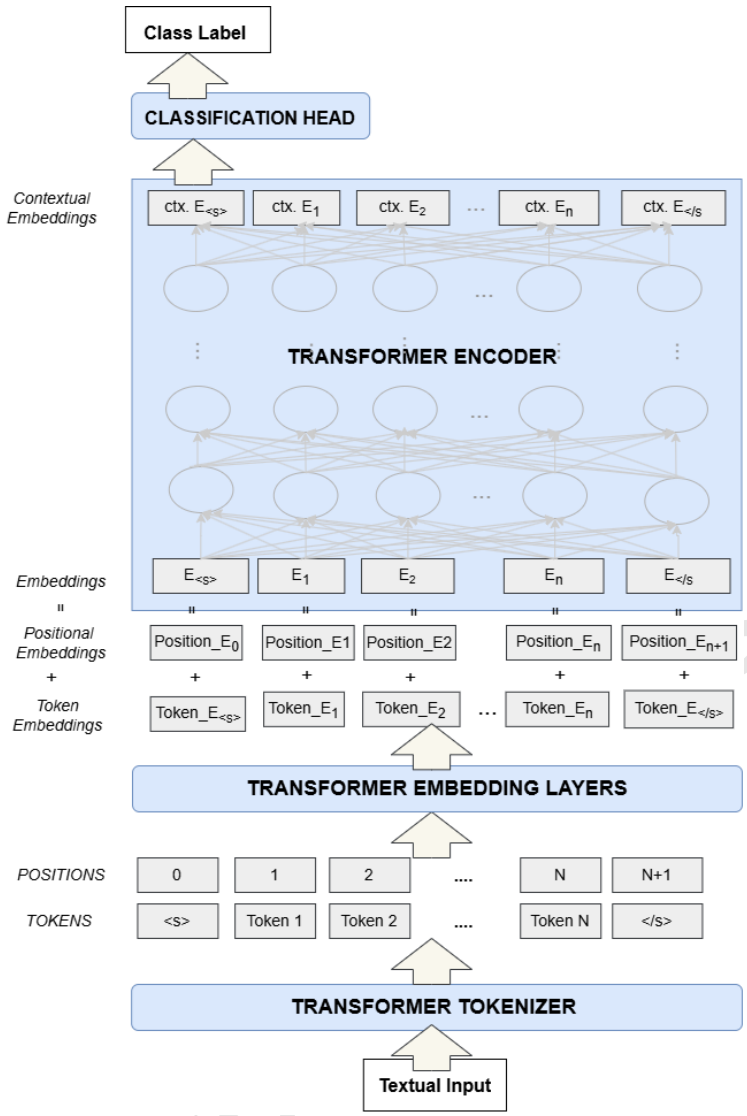

Fig. 5 Transformer encoder architecture for single-input classification

Let's define models XLM-RoBERTa and LitLatBERT as XLM and LitLat respectively, and the input regimes full text and propaganda sentences all and prop. Then the embedding sequence produced by each variant can be defined as:

$$\begin{aligned}
\text{XLM-RoBERTa Dual:} \quad & [\, x^{\text{XLM}}_{\text{all}}, x^{\text{XLM}}_{\text{prop}} \,], \\
\text{LitLatBERT Dual:} \quad & [\, x^{\text{LitLat}}_{\text{all}}, x^{\text{LitLat}}_{\text{prop}} \,], \\
\text{XLM–LitLat Multi-Embedding:} \quad & [\, x^{\text{XLM}}_{\text{all}}, x^{\text{XLM}}_{\text{prop}}, x^{\text{LitLat}}_{\text{all}}, x^{\text{LitLat}}_{\text{prop}} \,].
\end{aligned}$$

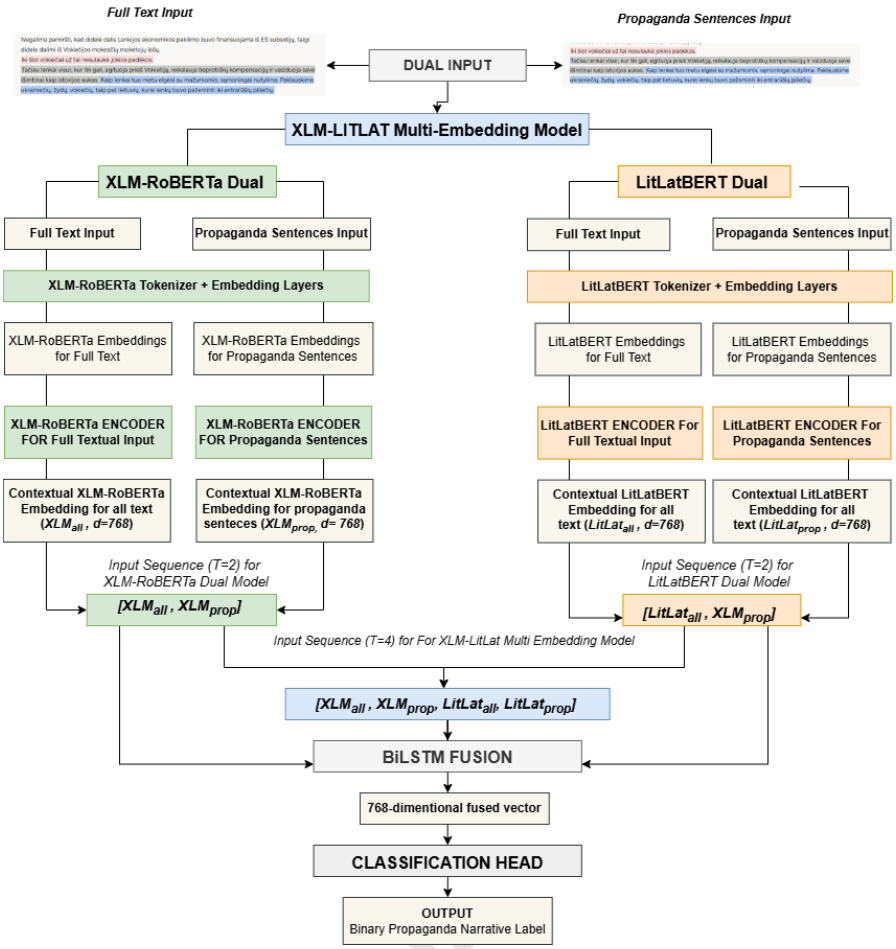

Fig. 6 Dual-input hybrid model architecture

More generally, write

$$X = [x_1, \ldots, x_T] \in \mathbb{R}^{T \times d}, \qquad T \in \{2, 4\},$$

where each $x_t \in \mathbb{R}^d$ is one of the contextual embeddings above. The sequence X is fed to a bidirectional LSTM (BiLSTM) with hidden size h per direction. At time step $t \in \{1, \ldots, T\}$ with input x_t, the (single-direction) LSTM computes

$$i_t = \sigma\big(W_i[h_{t-1} \,\|\, x_t] + b_i\big), \quad f_t = \sigma\big(W_f[h_{t-1} \,\|\, x_t] + b_f\big), \quad \tilde{c}_t = \tanh\big(W_c[h_{t-1} \,\|\, x_t] + b_c\big),$$
$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t, \qquad o_t = \sigma\big(W_o[h_{t-1} \,\|\, x_t] + b_o\big), \quad h_t = o_t \odot \tanh(c_t),$$

where:

- $h_t, c_t \in \mathbb{R}^h$ are the hidden and cell states (initialize $h_0 = c_0 = \mathbf{0}$ for the forward LSTM and $h_{T+1} = c_{T+1} = \mathbf{0}$ for the backward LSTM);
- $i_t, f_t, o_t, \tilde{c}_t \in \mathbb{R}^h$ are the input, forget, output, and candidate gates;
- $[u \,\|\, v]$ denotes feature-wise concatenation;
- $W_g \in \mathbb{R}^{h\times(h+d)}$ and $b_g \in \mathbb{R}^h$ are the weights and biases for $g \in \{i, f, o, c\}$;
- $\sigma(z) = \frac{1}{1+e^{-z}}$ is the logistic sigmoid, $\tanh(\cdot)$ is the hyperbolic tangent, and $\odot$ is element-wise multiplication.

Two such LSTMs operate in parallel: the forward one processes $x_1 \to x_T$, and the backward one processes $x_T \to x_1$. We concatenate their final hidden states to form the fused representation

$$h^* = [\, \overrightarrow{h}_T \,\|\, \overleftarrow{h}_1 \,] \in \mathbb{R}^{2h}.$$

With $h = 384$, h^* is 768-dimensional. This fused representation is fed to a linear layer that maps it to a logit,

$$z = W_{\text{cls}}\, h^* + b_{\text{cls}}, \qquad W_{\text{cls}} \in \mathbb{R}^{1\times 2h},\ b_{\text{cls}} \in \mathbb{R},$$

and the predicted probability is $p = \sigma(z)$.

This recurrent fusion allows the model to learn interactions between global discourse and localized persuasive cues. Empirically, encoder–sequence combinations have outperformed purely transformer-based or purely recurrent baselines on propaganda and misinformation detection [55–59].

3.4 Hyperparameter Search

For each transformer model (XLM-RoBERTa, LitLatBERT), narrative, and input regime (full text vs. propaganda sentences), we carried out hyperparameter optimisation using leave-one-group-out cross-validation (LOGO-CV). We chose LOGO-CV because it better captures the evolving, context-dependent nature of propaganda and more faithfully assesses out-of-domain generalisation by evaluating performance on previously unseen outlets. We implemented LOGO-CV with four predefined folds at the outlet (domain) level: in each fold, all articles from a single outlet were held out for validation, while articles from the remaining outlets formed the training set. We defined the folds based on the distribution of narratives across sources.

In total, the articles came from seven different sources, but only four could be used as held-out validation domains. Two sources had very limited coverage overall (and very few instances for some narratives), and one source accounted for too large a share of articles; removing it would have produced unrepresentative training data for certain narratives. The resulting folds are summarised in Table 1. The held-out outlet alternates across the four folds (domains 4, 2, 6, and 5, respectively), while domains 1, 3, and 7 remain in training in every fold. The validation share varies from 10.5% to 20.8%, reflecting differences in outlet size. Fold 1 has the largest validation split (20.8%), whereas Folds 3–4 are the smallest (about 10–11%).

For each narrative, we ran grid searches over two regularization hyperparameters: dropout rate and the number of frozen transformer layers, while holding the learning

Table 1 LOGO-CV folds by held-out outlet and split proportions

			Split proportion	
Fold	Training domains	Held-out	Train %	Val %
1	1,2,3,5,6,7	4	79.2%	20.8%
2	1,3,4,5,6,7	2	86.7%	13.3%
3	1,2,3,4,5,7	6	89.5%	10.5%
4	1,2,3,4,6,7	5	89.2%	10.8%

rate (5×10^{-6}), batch size (8), and weight decay (0.01) fixed. Optimization used `AdamW` (Adam with decoupled weight decay) [60].

Freeze-layers

The freeze-layers hyperparameter specifies how many of the lowest transformer encoder layers are kept fixed during fine-tuning. Freezing preserves general-purpose lexical, syntactic, and positional representations learned during pretraining while allowing the upper layers to adapt to the target task. Attention heads that specialize in proper nouns are concentrated in the upper (last) layers; those targeting determiners appear primarily in the first four layers; and the middle layers tend to emphasize positional patterns rather than rich contextual semantics. This depth specialization suggests that deeper layers encode higher-level, task-relevant semantics, whereas the early and middle layers capture more surface-level or structural information [61]. Based on these assumptions, we experimented with varying the number of frozen layers during hyperparameter fine-tuning. For our 12-layer encoders, we evaluated four settings that freeze the bottom N layers:

- $N = 0$: no layers frozen (full fine-tuning);
- $N = 4$: freeze layers 1–4;
- $N = 6$: freeze layers 1–6;
- $N = 8$: freeze layers 1–8.

These configurations test whether retaining lower- and mid-level representations while allowing the top layers to adapt improves generalization.

Dropout

Dropout is a regularization technique used to reduce overfitting and improve generalization. During training, a fixed fraction of activations is randomly set to zero; at inference, dropout is disabled. In transformer architectures, dropout can be applied to several components: token/position embeddings, attention probabilities (in the self-attention mechanism), the hidden states within each self-attention and feed-forward (MLP) block, and the task-specific classification head. For attention and MLP blocks, a dropout rate around 0.1 is the standard default in pretrained models (e.g., BERT/RoBERTa). Lower rates (0–0.1) provide less regularization and can speed convergence when overfitting is mild, whereas higher rates ($\approx$ 0.2–0.4) add stronger regularization for small or imbalanced datasets but risk underfitting if set too high. In our setting, although the dataset contains roughly 1,000 annotated articles, narrative

labels are highly imbalanced. To mitigate overfitting under this imbalance, we emphasized stronger regularization and evaluated the following dropout configurations for each narrative: no dropout, 0.2, and 0.4.

3.5 Experimental Setup

We fix the data split once and reuse it in all subsequent experiments. Concretely, we split the 1,000 articles as follows: **60%** (600 articles) for training, **10%**(100 articles) for validation, and **30%** (300 articles) for testing. All splits are stratified by the narratives labels. These sets are frozen before any training begins and are used unchanged across all stages.

Across all experiments, we set training for 10 epochs with early stopping (patience = 5), with a batch size 8 and AdamW optimizer with learning rate 5×10^{-6} and weight decay 0.01. To mitigate label imbalance, we optimize *class-weighted cross-entropy*, which increases the loss contribution of under-represented classes so the model does not default to the majority class.

Our main regularization controls are the dropout rate and the number of frozen transformer layers, with grids dropout $\in \{0, 0.2, 0.4\}$ and frozen layers $\in \{0, 4, 6, 8\}$. Because our architectures can consume different textual inputs, we treat the *input regime* as part of the model specification: full textual content (`all`), extracted propaganda sentences (`prop`), or both (`dual`).

Performance is reported and selected using *macro-F1*, the unweighted average of the per-class F1 scores, which gives each class equal importance and is therefore more reliable than accuracy under class imbalance. In the binary case, with positive (+) and negative (−) classes, macro-F1 is defined as follows.

Let $\mathrm{TP}, \mathrm{FP}, \mathrm{FN}, \mathrm{TN}$ be the counts in the confusion matrix with respect to the positive class (y=1). Define per-class precision and recall as

$$\mathrm{P}_+ = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}, \quad \mathrm{R}_+ = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}, \qquad \mathrm{P}_- = \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FN}}, \quad \mathrm{R}_- = \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}}.$$

The per-class F1 scores are

$$\mathrm{F1}_+ = \frac{2\,\mathrm{P}_+\mathrm{R}_+}{\mathrm{P}_+ + \mathrm{R}_+} = \frac{2\,\mathrm{TP}}{2\,\mathrm{TP}+\mathrm{FP}+\mathrm{FN}}, \qquad \mathrm{F1}_- = \frac{2\,\mathrm{P}_-\mathrm{R}_-}{\mathrm{P}_- + \mathrm{R}_-} = \frac{2\,\mathrm{TN}}{2\,\mathrm{TN}+\mathrm{FP}+\mathrm{FN}}.$$

The binary macro-F1 is the average of these two:

$$\mathrm{F1}_{\mathrm{macro}} = \frac{1}{2}\Big(\mathrm{F1}_+ + \mathrm{F1}_-\Big) = \frac{1}{2}\left(\frac{2\,\mathrm{TP}}{2\,\mathrm{TP}+\mathrm{FP}+\mathrm{FN}} + \frac{2\,\mathrm{TN}}{2\,\mathrm{TN}+\mathrm{FP}+\mathrm{FN}}\right).$$

We use macro-F1 as the primary metric because it balances precision and recall for both the positive and negative classes equally, making it more robust than accuracy under class imbalance.

Stage 1: Hyperparameter search

For each narrative, we conduct a grid search with LOGO-CV using the predefined folds described in Section 3.3. This stage evaluates only the `all` and `prop` single-input regimes. Each transformer is fine-tuned as an independent binary classifier, consisting of 12 encoder layers with a classification head (Figure 5), with training repeated separately for each fold. Hyperparameters are selected by averaging validation performance across folds.

Stage 2: Fine-tuning transformers: single-input case

Using the best hyperparameters identified in stage 1 for each narrative, model, and input regime, we fine-tune transformer models (single-input) on the training set, with early stopping based on validation performance. Final evaluation is carried out on the test set, and these results serve as the baseline for the hybrid models.

Stage 3: Dual-input models training: dual-input case

We build separate dual-input hybrid models for each narrative, all following the same general architecture (Figure 6). As these models operate on dual inputs by combining the `all` and `prop` regimes, we adopt the best hyperparameters identified in stage 1 for each encoder and input type. For each narrative, three hybrid variants are fine-tuned with these hyperparameters: (i) a Dual XLM-RoBERTa model, which uses two XLM-RoBERTa encoders for the `all` and `prop` regimes fused with a BiLSTM; (ii) a Dual LitLatBERT model, which uses two LitLatBERT encoders for the `all` and `prop` regimes fused with a BiLSTM; and (iii) a XLM-LitLat Multi-Embedding model, which uses four encoders in total, namely two XLM-RoBERTa and two LitLatBERT encoders for the `all` and `prop` regimes, fused with a BiLSTM. In all cases, the dual inputs are processed separately by each encoder and subsequently fused. Training is performed on the same training set with early stopping based on validation performance, and final evaluation is conducted on the test set.

Stage 4: Propaganda sentence classifier

We additionally train an auxiliary binary sentence-level classifier for propaganda techniques. Both XLM-RoBERTa and LitLatBERT transformers are fine-tuned independently with a classification head, using the same hyperparameter search space described in stage 1 to select the best configuration for each model. Each transformer is trained on our sentence-level dataset (section 3.2). To avoid data leakage, training and validation sets are constructed by selecting sentences from the same articles used in the training and validation samples of earlier stages, while ensuring that no sentence from any article in the test set is included. As in other stages, training is performed with early stopping based on validation performance. Finally, the best-performing transformer on the validation set is selected as the propaganda-sentence classifier.

Stage 5: Evaluation

We evaluate two model families on the test set: (i) single-input models, consisting of fine-tuned transformer encoders (XLM-RoBERTa and LitLatBERT) with the best

hyperparameters selected for each narrative and input regime; and (ii) dual-input models consist of hybrid architectures that process both input regimes, namely full text and propaganda sentences, through separate encoders and fuse their representations using a BiLSTM layer. All models are evaluated at both the paragraph and article levels. For article-level evaluation, an article is labeled as containing a narrative if any of its paragraphs is assigned that narrative. This rule mirrors the annotation protocol, in which narrative labels were defined at the article level.

For the `prop` input (where only propaganda sentences are used) regime, we evaluate two settings: (i) using gold-standard, human-annotated propaganda sentences from the dataset, and (ii) using sentences predicted by our propaganda-sentence classifier. The classifier is applied only in the final, article-level evaluation to simulate a real-world scenario in which users submit raw text without sentence-level propaganda annotations.

After selecting the best-performing model for each narrative, we compared its outputs with those produced by a large language model baseline. Specifically, we used the GPT-5.1 Pro model in agent mode and configured it as an article-level labeling agent. The model was prompted with a task-specific instruction describing the target propaganda narrative, including its formal definition from the HALT-PROP corpus [49]. The model was instructed to determine whether the narrative was present in each article and to assign a binary label (1 = narrative present, 0 = narrative absent).

The output format was constrained to a structured Excel file containing article identifiers and separate binary labels for each narrative. This setup was designed to mirror the experimental framework used for the fine-tuned models, where each model was trained independently for a single narrative. Accordingly, GPT-5.1 was queried separately for each narrative to ensure methodological consistency and comparability across approaches. The comparison was conducted using the ChatGPT interface, and model parameters were not manually configured and therefore remained at the platform's default settings.

The objectives of this evaluation stage are to:

- investigate how hyperparameters such as frozen layers and dropout influence narrative-detection performance in a low-resource language setting;
- analyze the effect of two input regimes, namely full articles and propaganda sentences, to test whether propaganda-focused content improves narrative detection.
- examine whether dual-input models outperform single-input models;
- identify the best-performing configuration for each narrative;
- compare the selected best-performing models for each narrative with a strong LLM baseline (ChatGPT 5).

4 Results

4.1 Hyperparameter Search

Figure 7 reports macro-F1 across dropout rates and numbers of frozen layers for both encoders and both input regimes. Overall, partial freezing is beneficial: peak scores cluster at 4–8 frozen layers rather than with all layers trainable (0 frozen). A consistent

Fig. 7 avg. F1 for naratives, models and regimes

pattern holds across encoders. In the propaganda-only input regime ("prop."), higher freezing tends to work best; in the full-text regime ("all"), lower freezing (i.e., more trainable layers) is preferred, suggesting that the model must adapt more when the full context is provided. Model-wise, XLM-RoBERTa generally prefers more freezing ($\approx$ 4–6 layers) than LitLatBERT (0–6, typically fewer), implying that LitLatBERT benefits from greater task-specific adaptation. There are exceptions. For the narrative *Delegitimization of the Lithuanian State*, XLM-RoBERTa achieves its best results with no freezing, whereas for *Attacks on Western Institutions and Alliances*, LitLatBERT performs best without freezing. In general, this suggests that while XLM-RoBERTa often benefits from greater freezing, it requires full adaptation for narratives that are

highly specific to the Lithuanian context; conversely, LitLatBERT tends to require more adaptation when dealing with more broadly Western-oriented content.

These trends align with how fine-tuning interacts with transformer encoders: lower layers capture broadly useful linguistic features that can be safely frozen, while higher layers adapt to task and domain. When the domain shift is larger (e.g., full-text inputs or narratives less aligned with a model's pretraining), training more layers is advantageous. Regarding dropout, the general trend is positive: the best results are usually obtained with dropout in the 0.2–0.4 range. This complements freezing: freezing preserves broadly useful pretrained features, while dropout regularizes the remaining trainable layers. For each narrative, input regime, and encoder, we select the dropout–freeze combination that yields the highest fold-averaged macro-F1; the selected hyperparameters are highlighted in Figure 7.

Table 2 Single-input models results across narratives and input regimes

			Training		Validation		Test	
Narrative	Input Regime	Model	Acc.	F1	Acc.	F1	Acc.	F1
Erosion of Trust in Lithuanian Institutions	All	XLM-RoBERTa	80.36%	80.22%	85.08%	85.04%	73.62%	73.50%
		LitLatBERT	79.64%	79.43%	85.48%	85.41%	75.57%	75.54%
	Prop.	XLM-RoBERTa	79.86%	79.84%	85.48%	85.13%	76.38%	76.33%
		LitLatBERT	88.99%	88.98%	84.68%	84.56%	77.99%	**77.13%**
Attacks on Western Institutions and Alliances	All	XLM-RoBERTa	88.87%	84.31%	83.03%	71.40%	78.32%	64.68%
		LitLatBERT	84.08%	78.45%	80.73%	68.46%	73.79%	61.20%
	Prop.	**XLM-RoBERTa**	74.40%	67.25%	87.94%	75.78%	81.12%	**65.85%**
		LitLatBERT	87.20%	83.30%	87.44%	75.13%	78.81%	61.67%
Decline of Western Civilization	**All**	XLM-RoBERTa	99.37%	98.95%	84.26%	73.40%	80.74%	66.98%
		LitLatBERT	97.19%	95.49%	84.26%	74.41%	80.26%	**67.84%**
	Prop.	XLM-RoBERTa	80.24%	60.98%	81.48%	68.07%	77.99%	63.92%
		LitLatBERT	92.62%	87.51%	81.94%	67.86%	79.77%	64.55%
Delegitimization of Lithuanian State	All	XLM-RoBERTa	88.01%	81.25%	82.46%	66.25%	82.36%	59.80%
		LitLatBERT	85.25%	76.51%	84.65%	68.80%	83.62%	**66.87%**
	Prop.	**XLM-RoBERTa**	79.79%	71.16%	79.19%	65.53%	80.15%	**66.96%**
		LitLatBERT	81.71%	72.86%	79.70%	67.39%	78.23%	64.60%
New world order	All	XLM-RoBERTa	84.01%	60.98%	87.16%	71.32%	88.03%	64.25%
		LitLatBERT	85.92%	63.16%	86.24%	67.25%	87.38%	64.53%
	Prop.	XLM-RoBERTa	86.62%	71.27%	92.20%	74.90%	91.42%	67.23%
		LitLatBERT	87.32%	71.37%	88.53%	69.03%	89.97%	**68.55%**
Authoritarian Model Promotion	All	XLM-RoBERTa	91.05%	70.64%	94.35%	74.42%	88.51%	56.15%
		LitLatBERT	89.06%	63.83%	90.87%	67.53%	86.08%	56.37%
	Prop.	XLM-RoBERTa	94.60%	81.87%	91.30%	68.24%	89.97%	54.28%
		LitLatBERT	91.04%	76.49%	90.04%	78.28%	85.46%	**58.35%**
Disinformation about War in Ukraine	All	XLM-RoBERTa	94.03%	85.66%	90.04%	78.28%	91.10%	76.20%
		LitLatBERT	93.67%	84.68%	88.31%	75.83%	89.64%	74.85%
	Prop.	XLM-RoBERTa	94.69%	87.97%	87.44%	74.66%	90.37%	76.10%
		LitLatBERT	96.86%	92.79%	88.89%	76.60%	91.33%	**76.94%**

4.2 Paragraph - Level Results

Tables 2 and 3 report paragraph-level results on the test set for the single-input and dual-input models, respectively, with the best score for each narrative highlighted. Overall, dual-input models consistently outperform single-input models. Within the single-input family, the *prop* regime (using only propaganda sentences) is beneficial: it yields the best result in 6/7 narratives in Table 2. When both regimes are combined in a dual-input architecture, performance improves further, and dual-input models achieve the top results across all narratives (Table 3).

These results also highlight the strength of the LitLatBERT encoder. Among single-input models, LitLatBERT attains the best scores in 5/7 narratives; in the dual-input setting, the LitLatBERT-only dual model (*LitLat-Dual*) achieves the highest performance in 4/7 narratives. In general, *LitLat-Dual* tends to lead precisely where the single-input LitLatBERT was already strongest. A similar pattern holds for XLM-RoBERTa; for example, on *Attacks on Western Institutions and Alliances*, *XLM-Dual* improves from 65.85% (*XLM-RoBERTa-Prop*) to 67.06%.

Table 3 Hybrid results across narratives and splits

		Training		Validation		Testing	
Narrative	Model	Acc.	F1	Acc.	F1	Acc.	F1
Erosion Of Trust in Lithuanian Institutions	**XLM Dual**	82.52%	82.44%	85.48%	85.39%	79.45%	**79.41%**
	LitLat Dual	81.73%	81.63%	85.48%	85.32%	78.96%	78.95%
	XLM–LitLat	79.42%	79.35%	85.08%	84.93%	78.64%	78.58%
Attacks on Western Institutions and Alliances	**XLM Dual**	78.03%	70.92%	83.03%	70.21%	80.56%	**67.06%**
	LitLat Dual	87.54%	83.21%	75.23%	64.96%	78.80%	63.65%
	XLM–LitLat	88.31%	83.75%	85.78%	73.31%	81.55%	66.00%
Decline of Western Civilization	XLM Dual	94.16%	91.07%	77.78%	66.85%	80.74%	70.08%
	LitLat Dual	99.23%	98.72%	85.19%	75.92%	83.82%	**71.95%**
	XLM–LitLat	88.47%	82.64%	77.78%	69.20%	79.61%	70.61%
Delegitimization of Lithuanian State	XLM Dual	95.46%	91.80%	81.58%	68.16%	80.26%	63.01%
	LitLat Dual	85.04%	77.47%	80.26%	67.46%	83.50%	**68.01%**
	XLM–LitLat	90.64%	84.80%	84.65%	66.88%	78.64%	67.48%
New World Order	XLM Dual	86.55%	73.47%	86.70%	67.80%	87.70%	66.42%
	LitLat Dual	90.85%	79.12%	89.45%	71.50%	88.83%	67.43%
	XLM–LitLat	89.65%	77.11%	88.99%	74.64%	89.81%	**69.32%**
Authoritarian Model Promotion	XLM Dual	92.26%	78.18%	91.74%	68.99%	89.81%	59.32%
	LitLat Dual	90.02%	75.13%	88.70%	65.94%	86.89%	**60.55%**
	XLM–LitLat	91.62%	76.60%	90.43%	66.85%	87.38%	57.57%
Disinformation about War in Ukraine	XLM Dual	85.93%	72.18%	85.28%	72.65%	89.48%	75.48%
	LitLat Dual	89.48%	76.76%	88.74%	76.42%	91.59%	**77.34%**
	XLM–LitLat	83.30%	70.14%	88.74%	77.02%	90.29%	76.97%

4.3 Article - Level Results

4.3.1 Auxiliary Model: Sentence Classifier

Figure 8 reports the validation performance of our auxiliary sentence-level propaganda detector, which predicts whether a sentence contains any propaganda technique. Across configurations, macro F1 ranges from 69.32% to 71.81%. LitLatBERT generally outperforms XLM-RoBERTa across most combinations of frozen layers and dropout; the sole exception is the setting with eight frozen layers and no dropout, where XLM-RoBERTa is marginally higher. In all other cases, LitLatBERT leads by approximately 1-2 percentage points, confirming its advantage for Lithuanian. The best result is obtained with LitLatBERT fine-tuned using six frozen layers and dropout = 0.2, accordingly, for the article-level evaluation we used the model trained on annotated sentences for the prop regime.

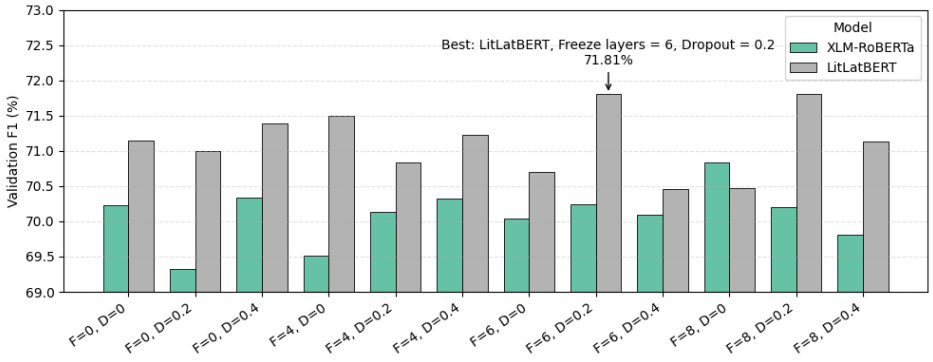

Fig. 8 Validation F1 results for sentence classifier

4.3.2 Final Evaluation

We report results for all models using a maximum rule across paragraphs: if an article is split into paragraphs and any paragraph receives the narrative label, the entire article is labeled positive for that narrative.

For the models (single-input models using prop regime and all dual-input hybrids) we compute results in two ways: (i) using gold-standard sentence labels derived from article annotations, and (ii) using sentences predicted by our propaganda-sentence classifier (Section 4.3.1). For the final model selection, we adopt the latter (classifier-assigned sentences) because it reflects the real-world setting in which inputs lack human annotations.

Table 4 reports article-level results for all models. Overall, one more time as it was already noticed in paragraph level results, the hybrid dual-input models consistently outperform single input models across all narratives. A notable exception is *Disinformation about the War in Ukraine*: XLM-RoBERTa in the `prop` regime attains the best score when provided with gold, expert-annotated propaganda sentences

Table 4 Models results on article level with actual and predicted propaganda sentences (test set)

Model	Prop. Sent.	Distrust of Lithuanian Institutions		Attacks on West. Institutions and Alliances		Decline of Western Civilization		Deligitimization of the Lithuania State		New World Order		Authoritarian Model Promotion		Disinformation about the War in Ukraine	
		ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1
XLM-ALL	–	73.62%	73.50%	71.33%	63.92%	78.33%	66.36%	80.67%	62.09%	88.00%	68.52%	85.67%	52.22%	90.33%	79.73%
LitLat-ALL	–	75.57%	75.54%	77.33%	61.04%	80.33%	70.53%	77.33%	66.23%	86.33%	68.76%	79.33%	51.07%	90.00%	74.46%
XLM-PROP	Pred.	74.27%	74.27%	73.67%	63.73%	77.00%	67.03%	80.00%	64.56%	90.61%	65.94%	85.00%	51.78%	88.33%	76.01%
	Actual	76.38%	76.33%	78.00%	67.95%	78.33%	67.54%	82.33%	66.86%	91.42%	67.23%	86.67%	56.36%	92.33%	**82.87%**
LitLat-PROP	Pred.	77.99%	77.71%	76.33%	63.25%	80.33%	69.83%	69.67%	60.08%	87.22%	63.81%	82.67%	56.88%	90.33%	79.73%
	Actual	77.99%	77.93%	81.33%	67.43%	80.67%	68.59%	74.00%	63.66%	89.97%	68.55%	81.67%	53.80%	91.00%	81.50%
XLM Dual	Pred.	79.67%	**79.16%**	74.33%	65.31%	80.00%	70.86%	76.33%	62.79%	82.33%	65.04%	85.33%	60.43%	87.33%	76.44%
	Actual	83.00%	**82.64%**	77.33%	68.93%	79.67%	70.21%	77.33%	62.68%	85.00%	67.89%	87.33%	62.58%	87.67%	76.86%
LitLat Dual	Pred.	76.00%	74.78%	73.00%	62.45%	82.33%	**71.86%**	81.00%	**68.05%**	90.67%	69.45%	82.00%	59.32%	90.33%	80.51%
	Actual	82.33%	81.67%	76.33%	65.71%	84.33%	**74.38%**	81.67%	**68.25%**	91.33%	71.64%	84.00%	62.12%	91.67%	82.16%
XLM–LitLat	Pred.	77.67%	76.70%	76.33%	**65.33%**	78.67%	71.03%	71.00%	61.16%	87.67%	**70.45%**	84.33%	**60.52%**	90.67%	**81.36%**
	Actual	82.67%	82.17%	80.67%	**69.79%**	79.67%	72.00%	74.00%	64.02%	89.33%	**72.76%**	85.33%	61.52%	89.00%	80.02%

(82.87%), but its performance drops markedly when those sentences are supplied by the sentence-level predictor instead (76.01%). This illustrates that models trained on propaganda-only inputs can achieve strong results when given gold-standard (human-annotated) propaganda sentences, yet become highly sensitive when those sentences are provided by an automatic sentence-level predictor. By contrast, dual-input hybrid models that condition on both the full article and propaganda sentences are substantially more robust to sentence-selection noise and deliver the best article-level performance. Below we give the evaluation for each narrative.

- **Erosion of Trust in Lithuanian Institutions.** For this narrative hybrid XLM-dual model shows the best results, it has the highest scores with actual propaganda sentences (82.64%), as well as with predicted (79.16%), this follows the achieved results for paragraph level, where the XLM-dual model also demonstrated the best results for this narrative.
- **Attacks on Western Institutions and Alliances, New World Order**. For both narratives, the two-encoder hybrid XLM-LitLat model achieves the highest performance. For the *New World Order* narrative, it already leads at the paragraph level. For *Attacks on Western Institutions and Alliances*, its superiority appears only at the article level, with the best scores for "actual" (69.7%) and "predicted" (65.33%) sentences. This differs slightly from the paragraph-level case, where XLM-Dual was the clear leader; however, at the article level XLM-Dual's performance (65.31%) on predicted sentences is very close to that of XLM-LitLat (65.33%).
- **Decline of Western Civilization, Delegitimization of the Lithuanian State** LitLatBERT encoder-based models are superior across all the experiments for these narratives with a LitLat-Dual showing best results on both paragaph and article levels.
- **Authoritarian Model Promotion.** This narrative had the lowest share among all selected narratives (6.6%), so the results were sensitive to positive-class predictions: small differences in precision or recall could lead to sizable changes in the final score. At the article level, using predicted propaganda sentences, *XLM-LitLat Dual* achieved the highest score (60.52%), with *XLM-Dual* very close behind (60.43%). With actual annotated sentences, *XLM-Dual* performed best at the article level, whereas at the paragraph level the best results were achieved by *LitLat-Dual*. Overall, considering article-level results with predicted sentences for dual-input models (which significantly outperform single-input models), scores fall in the range of ∼59–61%. As noted above, we select the best model based on article-level performance with predicted sentences; for this narrative, that is *XLM-LitLat Dual*.
- **Disinformation about the War in Ukraine.** At the article level, the best results with predicted sentences were achieved by *XLM-LitLat Dual* (81.36%). With actual sentences, the best result came from the single-input *XLM-RoBERTa* model using propaganda sentences (82.87%). However, as noted above, this single-input model is very sensitive to errors from the sentence predictor: when using predicted sentences, its performance dropped substantially (76.01%).

We compare the results for each narrative against a ChatGPT-5 baseline. This setup reflects our goal of evaluating the system in a real-world scenario in which raw

Table 5 Article-level results: hybrid models vs. GPT-5 baseline.

		Hybrid Model		GPT-5	
Narrative	Best model	Acc.	F1	Acc.	F1
Erosion of Trust in Lithuanian Institutions	XLM–Dual	79.67%	**79.16%**	69.50%	68.93%
Attacks on Western Institutions and Alliances	XLM– LitLat	76.33%	**65.33%**	73.49%	51.36%
Decline of Western Civilization	LitLat– Dual	82.33%	**71.86%**	75.84%	56.04%
Delegitimization of the Lithuanian State	LitLat–Dual	81.00%	**68.05%**	86.24%	60.23%
New World Order	XLM–LitLat	87.67%	**70.45%**	85.57%	63.85%
Authoritarian Model Promotion	XLM– LitLat	84.33%	**60.52%**	91.95%	57.88%
Disinformation about the War in Ukraine	XLM– LitLat	90.67%	**81.36%**	92.62%	78.32%

text, without any sentence annotations, is submitted to the model. The results in Table 5 shows the article level results with the selected best performing model for each narrative and ChatGPT-5 results. In general our hybrid models outperform ChatGPT-5 across all narratives. Both systems achieve their best results on *Disinformation about the War in Ukraine* (our hybrid model: 81.36%; ChatGPT-5: 78.32%), consistent with this narrative's event-bounded nature, which yields a more separable context despite class imbalance. The second-best results for both systems is also achieved for the same narrative - *Erosion of Trust in Lithuanian Institutions*, as was already mentioned before this is the only one balanced narrative, however for this narrative our hybrid model substantially exceeds ChatGPT-5 (79.16% vs. 68.93%). Also, worth mentioning that despite the fact that *Attacks on Western Institutions and Alliances* narrative was the second largest narrative in the data, for both systems the performance for this narrative was one of the lowest compared to other narratives: our best hybrid model(*XLM-LitLat*) achieves 65.33% which is second lowest result after Authoritorian Model Promotion narrative, and ChatGPT-5 achieves the lowest performance for this narrative with 51.36%. We attribute this to contextual overlap with related narratives, particularly *Erosion of Trust in Lithuanian Institutions*, which shares institution-directed distrust, and *Decline of Western Civilization*, which shares West-oriented thematic content. Such overlap induces highly correlated lexical and topical features and reduces class separability in the representation space.

5 Conclusion and Discussion

This paper introduces the first supervised system for detecting pro-Kremlin propaganda narratives in Lithuanian, a low-resource language in a country heavily targeted

by Russian disinformation. To our knowledge, this is the first such approach in the Baltic region, and Russia's wider neighborhood.

We base the system design on three main research questions: (i) How do fine-tuning strategies affect performance in low-resource narrative detection? (ii) Does restricting inputs to sentences containing propaganda techniques improve performance compared to using full articles? (iii) Do dual-input hybrid models outperform single-encoder baselines? The key findings are:

- **Fine-tuning strategies.** Freezing the lower layers of a transformer encoder, together with a medium-to-high dropout rate, is beneficial for propaganda-narrative detection in low-resource languages. The optimal freezing depth and dropout rate depend on the narrative, the model, and the input regime. Training on propaganda-bearing sentences usually requires less model adaptation and tends to work best with stronger freezing, whereas full-text input generally benefits from training more layers and thus greater task-specific adaptation. The Lithuanian-tailored LitLatBERT encoder tends to benefit from stronger freezing on local narratives, whereas multilingual XLM-RoBERTa typically requires less adaptation for Western- or globally oriented narratives. These findings underscore that the fine-tuning strategy should be guided by the narrative's cultural and linguistic context.
- **Input granularity.** Training on sentences flagged for propaganda techniques improves narrative detection compared with using full articles; for some narratives, it can even surpass dual-input models at the article level. However, propaganda sentence-only training is sensitive to errors when human annotations are unavailable and sentence-level labels must be predicted automatically. A dual-input hybrid approach that processes both the full article and the subset of propaganda sentences retains these gains while remaining robust to noise in automatically predicted sentence labels.
- **Model architecture.** Hybrid dual-input models consistently outperform single-encoder, fine-tuned transformer baselines across all seven narratives and also outperform ChatGPT-5 under identical evaluation conditions. In our article-level evaluation using the sentence predictor, the XLM–LitLat multi-embedding model achieved the best performance on 4 of the 7 narratives, the LitLat-dual model on 2, and the XLM-dual model on 1. These results suggest that hybrid architectures that combine representations from different encoders should be considered alongside single-encoder systems, as they deliver strong performance across diverse narrative contexts.

Beyond these answers, our results suggest practical guidelines for extending narrative detection to other low-resource languages:

1. **Context matters.** Choose encoders and fine-tuning strategies according to the narrative's context. Language-specific encoder are particularly suited to local narratives, while well-tuned multilingual encoders can match or even surpass them when dealing with narratives in a more global context.
2. **Freezing is an important hyper-parameter.** By preserving lower-layer representations and adapting the upper layers, pre-learned lexical, syntactic, and

positional knowledge is leveraged. The optimal freezing depth varies according to the narrative's context and the choice of model.

3. **Propaganda techniques improve narrative detection results but can introduce additional noise.** Propaganda techniques convey strong narrative signals and should be included, but models must remain effective when sentence annotations are noisy or absent. Hybrid architectures incorporating various textual features are less tied to propaganda techniques alone, help maintain balance, and make the model more versatile and less dependent on a single feature type.
4. **Follow human annotation logic.** Just as annotators first read the entire text to understand what it's about and then, to assess whether the text manipulates a certain topic, evaluate how it is talked about, dual-input models better recognize narratives when they evaluate both the entire textual content and the phrases where propaganda techniques are expressed; they can better detect the presence of the narrative in the text.
5. **Fine-grained narrative and event annotation.** Narratives reflecting only the context of a single specific event are easier to detect than those that constantly change and adapt to different contexts. Therefore, we recommend that, when creating datasets for narrative detection, one should not limit labeling only at the article level but also annotate exactly where in the text the narrative is expressed, and add annotations that reflect the specific event to which the narrative is assigned.

5.1 Limitations and Future work

Dependency on Sentence-Level Technique Detection

Although our experiments show that hybrid models are more robust to sentence-classifier errors than single-input models, the comparison between gold-standard and automatically predicted sentence annotations still reveals noticeable performance deterioration. This suggests that hybrid models remain sensitive to noise introduced by the sentence-level classifier used to identify propaganda-technique-bearing sentences. At the same time, the overall performance of dual-input models incorporating such sentences indicates that propaganda techniques provide valuable signals for narrative detection. However, the current pipeline relies on two separate models operating sequentially, which introduces error propagation between stages. To address this limitation, future work could explore the development of a unified model that jointly learns to detect propaganda techniques and narratives. Such a joint learning approach would allow the model to share internal representations between tasks and reduce reliance on a separate sentence-level classifier.

Narrative Overlap and Weak Supervision Noise

Another limitation of the proposed approach is related to the substantial overlap between narratives. The narrative overlap analysis conducted in this study shows that narratives frequently co-occur and are often expressed through interconnected thematic storylines. In the current modeling approach, narratives are treated as independent labels, which may overlook important relationships between them. When narratives appear together within the same article, treating them as fully independent categories may limit the model's ability to capture these underlying connections.

Future research could therefore explore approaches that explicitly model dependencies between narratives.

Narrative overlap also interacts with the weak supervision strategy used in this study. Due to the limited maximum sequence length of transformer models, longer articles are split into smaller paragraphs, and each paragraph inherits the article-level narrative labels. However, when an article expresses multiple narratives, different narratives may appear in different parts of the article, sometimes spanning entire paragraphs. As a result, some paragraph fragments may receive narrative labels even though they do not explicitly express the corresponding narrative, which introduces additional label noise and makes it more difficult for the model to learn narrative-specific features. To mitigate this issue, future narrative corpora could include annotations not only at the article or claim level but also at the span level, indicating where exactly narratives are expressed within the text. Such annotations would reduce label noise and allow models to learn more precise narrative representations.

From the modeling perspective, future work could also explore transformer models that support longer input sequences. However, this remains challenging for many low-resource languages, where fewer models are available with extended context windows. In addition, longer input sequences require significantly greater computational resources, which increases training costs.

Finally, when working with datasets labeled only at the article level and models with limited input sizes, future work could explore hybrid approaches that combine the outputs of large language models and supervised classifiers. For example, ensemble methods could integrate predictions from both systems, or LLMs could be used as a verification mechanism to assess whether the labels produced by supervised models are consistent with the narrative definitions.

Detection of Rare and Emerging Narratives

Due to data limitations this study focuses on the seven most frequent narratives in the dataset. However, propaganda discourse is dynamic, and new narratives continuously emerge or existing narratives evolve over time. This raises concerns about the model's ability to detect narratives that are either underrepresented in the training data or newly emerging. Several directions could address this issue. Although narratives change depending on context, propaganda messages are often constructed using recurring rhetorical techniques. For example, techniques such as emotional language rely on similar persuasive strategies designed to evoke strong emotional responses, even when applied to different topics or events. Therefore, approaches that model these rhetorical techniques alongside narratives may help improve the robustness of propaganda detection systems.

Future research could also explore hybrid systems that combine supervised and unsupervised approaches. While supervised models typically achieve higher accuracy, they depend on labeled data, whereas unsupervised methods can analyze unlabeled content. For example, documents that are not assigned a narrative label by supervised models could be analyzed using clustering methods to identify potential thematic similarities. A more advanced approach could integrate a supervised narrative detection model, a propaganda technique detection model, and an unsupervised clustering

component. In such a system, documents containing propaganda techniques but lacking a detected narrative could be grouped through unsupervised methods to identify emerging narrative patterns.

Abbreviations

ACC – Accuracy, an evaluation metric reflecting the proportion of correctly classified instances. BERT -Bidirectional Encoder Representations from Transformers, a transformer-based language model. BiLSTM - Bidirectional Long Short-Term Memory, a recurrent neural network architecture. CC-100 – Common Crawl corpus covering 100 languages (≈2.5 TB), used for multilingual pretraining. COVID-19 – Coronavirus Disease 2019. EMBEDDIA– European project providing cross-lingual word embeddings and resources. EN– English language. EU– European Union. FN – False Negative, an instance where a positive example is incorrectly classified as negative. FP– False Positive, an instance where a negative example is incorrectly classified as positive. GPT-5 – Fifth-generation Generative Pre-Trained Transformer model used as a baseline. HALT-PROP – Human-Annotated Lithuanian Propaganda corpus used in this study. LDA– Latent Dirichlet Allocation, a topic-modeling method. LGBT– Lesbian, Gay, Bisexual, and Transgender. LitLatBERT– A trilingual transformer model pretrained on Lithuanian, Latvian and English text. LOGO-CV– Leave-One-Group-Out cross-validation. LLM– Large Language Model. LSA– Latent Semantic Analysis. LSTM – Long Short-Term Memory, a type of recurrent neural network. LT– Lithuanian language. LV– Latvian language. MLP– Multi-Layer Perceptron, a feed-forward neural network. NATO– North Atlantic Treaty Organization. NMF– Non-Negative Matrix Factorization. NLP– Natural Language Processing. TF-IDF– Term Frequency–Inverse Document Frequency, a feature representation for text. TN– True Negative, a correctly classified negative instance. TP– True Positive, a correctly classified positive instance. XLM-RoBERTa– A multilingual RoBERTa-based transformer model. WWII– Second World War.

Declarations

Availability of data and materials

The HALT-PROP corpus used in this study is freely available at 10.18279/MIDAS.260057.

Competing Interests

The authors declare that they have no competing interests.

Funding

This research was supported by the Lithuanian Government Priority Research Program “Building Societal Resilience and Crisis Management in the Context of Contemporary Geopolitical Developments” (implemented through the Lithuania Research

Council) under grant number S-VIS-23-8. Project title: "Propaganda and Disinformation Research: Machine Learning-Based Automatic Detection, Impact and Societal Resilience."

Authors' contributions

IR designed and executed all experiments; implemented the models and pipelines; performed the analyses; curated the data; and wrote the entire manuscript. VM served as the primary advisor throughout, providing methodological guidance and consulting on model design; VM also contributed critical revisions. DP supervised the project and provided oversight feedback on the research and the manuscript. All authors read and approved the final manuscript.

Acknowledgements

The authors thank the anonymous reviewers for their valuable suggestions. The authors are also grateful to the annotators who created the HALT-PROP corpus.

Materials availability

Not applicable.

Code availability

The code supporting this study will be made available upon reasonable request.

References

[1] Elsner, M., Atkinson, G., Zahidi, S.: Global risks report 2025 (20th ed.). World Economic Forum (2025). Published January 2025

[2] Kelley, M.J.: Understanding Russian disinformation and how the joint force can address it. Parameters **54**(2), 39–52 (2024) https://doi.org/10.55540/0031-1723.3286

[3] Paul, C., Matthews, M.: The Russian "firehose of falsehood" propaganda model: Why it might work and options to counter it. Rand perspective pe-198-osd, RAND Corporation (2016). Retrieved July 30, 2025. https://www.rand.org/pubs/perspectives/PE198.html

[4] Karlsen, G.H.: Divide and rule: Ten lessons about Russian political influence activities in europe. Palgrave Communications **5**, 19 (2019) https://doi.org/10.1057/s41599-019-0227-8

[5] Garriaud Maylam, J.: The Russian war on truth: Defending allied and partner democracies against the Kremlin's disinformation campaigns. General report 014 cds 23 e rev. 2 fin., NATO Parliamentary Assembly (2023). Retrieved July 30, 2025. https://www.nato-pa.int/download-file?filename=/sites/default/

files/2023-10/014%20CDS%2023%20E%20rev.%202%20fin%20-%20RUSSIA%20DISINFORMATION%20-%20DEMEUSE%20REPORT.pdf

[6] Hiršs, M.: Echoes from Kremlin: New platforms, old narratives. Technical report, Civil Resilience Initiative: Disinformation Monitoring (July 2025). Retrieved from https://www.rgsl.edu.lv/data/pdf-files/echoes-from-kremlin-in-latvia.pdf. https://www.rgsl.edu.lv/data/pdf-files/echoes-from-kremlin-in-latvia.pdf

[7] Karpchuk, N.: The Russian federation propaganda narratives. Torun International Studies **1**(14), 19–30 (2021) https://doi.org/10.12775/TIS.2021.002

[8] Götz, E.: Near abroad: Russia's role in post-soviet eurasia. Europe-Asia Studies **74**, 1529–1550 (2022) https://doi.org/10.1080/09668136.2022.2093315

[9] Bryjka, F.: Russian disinformation regarding the attack on Ukraine. Pism spotlight no. 15/2022, Polish Institute of International Affairs (PISM) (February 2022). https://pism.pl/publications/russian-disinformation-regarding-the-attack-on-ukraine

[10] EUvsDisinfo: DISINFO: The West organises a coup in Belarus in the name of democracy. EUvsDisinfo (2021). https://euvsdisinfo.eu/report/the-west-organises-a-coup-in-belarus-in-the-name-of-democracy/

[11] University of Pennsylvania, Annenberg School for Communication: Foreign Voices, Familiar Faces: How (Pro)-Russian (Dis)information Undermines Trust in Moldova's Democracy. Milton Wolf Seminar on Media and Diplomacy. Retrieved August 20, 2025 (n.d.). https://www.asc.upenn.edu/research/centers/milton-wolf-seminar-media-and-diplomacy-19

[12] NATO Strategic Communications Centre of Excellence: Georgia's information environment through the lens of Russia's influence. Report, NATO StratCom COE, Riga (2021). https://stratcomcoe.org/publications/georgias-information-environment-through-the-lens-of-russias-influence/79

[13] Dadabaev, T.: Manipulating post-soviet nostalgia: contrasting political narratives and public recollections in Central Asia. International Journal of Asian Studies **18**(1), 61–81 (2021) https://doi.org/10.1017/S1479591420000443

[14] Hoyle, A., Wagnsson, C., Powell, T.E., Berg, H., Doosje, B.: Life through grey-tinted glasses: how do audiences in Latvia psychologically respond to sputnik latvia's destruction narratives of a failed latvia? Post-Soviet Affairs **40**(1), 1–18 (2024) https://doi.org/10.1080/1060586X.2023.2275507

[15] Chkhartishvili, T.: Russian propaganda: Adapting to Russian channels' ban in Lithuania. Georgetown Security Studies Review. Accessed August 20, 2025 (2023). https://georgetownsecuritystudiesreview.org/2023/11/30/russian-propaganda-adapting-to-russian-channels-ban-in-lithuania/

[16] NATO Strategic Communications Centre of Excellence: Analysis of Russia's information campaign against Ukraine: Examining non-military aspects of the crisis in ukraine from a strategic communications perspective. Report, NATO Strategic Communications Centre of Excellence, Riga (2016). https://stratcomcoe.org/cuploads/pfiles/russian_information_campaign_public_12012016fin.pdf

[17] NATO Strategic Communications Centre of Excellence: Georgia's Information Environment Through the Lens of Russia's Influence. Technical report, NATO StratCom COE, Riga (2021). https://stratcomcoe.org/cuploads/pfiles/Georgias-information-environment-through-the-lens-of-Russias-infulence.pdf

[18] 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 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 2931–2937. Association for Computational Linguistics, Copenhagen, Denmark (2017). https://doi.org/10.18653/v1/D17-1317

[19] Barrón-Cedeño, A., Jaradat, I., Da San Martino, G., Nakov, P.: Proppy: Organizing the news based on their propagandistic content. Information Processing & Management **56**(5), 1849–1864 (2019) https://doi.org/10.1016/j.ipm.2019.04.003

[20] Malik, M.S.I., Imran, T., Mona Mamdouh, J.: How to detect propaganda from social media? exploitation of semantic and fine-tuned language models. PeerJ Computer Science **9**, 1248 (2023) https://doi.org/10.7717/peerj-cs.1248

[21] Chaudhari, D., Pawar, A.V.: Empowering propaganda detection in resource-restraint languages: A transformer-based framework for classifying hindi news articles. Big Data and Cognitive Computing **7**(4), 175 (2023) https://doi.org/10.3390/bdcc7040175

[22] Kausar, S., Tahir, B., Mehmood, M.A.: Prosoul: A framework to identify propaganda from online urdu content. IEEE Access **8**, 186039–186054 (2020) https://doi.org/10.1109/ACCESS.2020.3029907

[23] Al-Henaki, L., Al-Khalifa, H., Al-Salman, A.: Enhancing propaganda detection in arabic news context through multi-task learning. Applied Sciences **15**(15), 8160 (2025) https://doi.org/10.3390/app15158160

[24] 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), pp. 5635–5645. Association for Computational Linguistics, Stroudsburg, PA, USA (2019). https://doi.org/10.18653/v1/D19-1565

[25] Shohan, S., Hossain, M., Paran, A., Ahsan, S., Hossain, J., Hoque, M.M.: SemanticCuetSync at araieval shared task: Detecting propagandistic spans with

persuasion techniques identification using pre-trained transformers. In: Proceedings of the Second Arabic Natural Language Processing Conference, pp. 518–523. Association for Computational Linguistics, Copenhagen, Denmark (2024)

[26] García-Díaz, J.A., Valencia-García, R.: UMUTeam at Dipromats 2023: Propaganda detection in spanish and english combining linguistic features with contextual sentence embeddings. In: Proceedings of IberLEF@SEPLN 2023 (2023)

[27] Horák, A., Sabol, R., Herman, O., Baisa, V.: Recognition of propaganda techniques in newspaper texts: Fusion of content and style analysis. Expert Systems with Applications **251**, 124085 (2024) https://doi.org/10.1016/j.eswa.2024.124085

[28] Quinn, K.M., Monroe, B.L., Colaresi, M., Crespin, M.H., Radev, D.R.: How to analyze political attention with minimal assumptions and costs. American Journal of Political Science **54**(1), 209–228 (2010) https://doi.org/10.1111/j.1540-5907.2009.00427.x

[29] Kinney, A.B., Davis, A.P., Zhang, Y.: Theming for terror: Organizational adornment in terrorist propaganda. Poetics **69**, 27–40 (2018) https://doi.org/10.1016/j.poetic.2018.04.003

[30] Clever, L., Schatto-Eckrodt, T., Clever, N.C., Frischlich, L.: Behind blue skies: A multimodal automated content analysis of islamic extremist propaganda on instagram. Social Media + Society **9**(1) (2023) https://doi.org/10.1177/20563051221150404

[31] Ylä-Anttila, T., Eranti, V., Kukkonen, A.: Topic modeling for frame analysis: A study of media debates on climate change in India and USA. Global Media and Communication **18**(1), 91–112 (2022) https://doi.org/10.1177/17427665211056794

[32] Verbytska, A.: Topic modelling as a method for framing analysis of news coverage of the Russia-Ukraine war in 2022–2023. Language & Communication **99**, 174–193 (2024) https://doi.org/10.1016/j.langcom.2024.05.005

[33] Mayopu, R.G., Wang, Y.Y., Chen, L.S.: Analyzing online fake news using latent semantic analysis: Case of USA election campaign. Big Data and Cognitive Computing **7**(2), 81 (2023) https://doi.org/10.3390/bdcc7020081

[34] Roozenbeek, J.: A note about methodology. In: Propaganda and Ideology in the Russian–Ukrainian War. Contemporary Social Issues Series, pp. 120–122. Cambridge University Press, ??? (2024)

[35] Lai, C., Toriumi, F., Yoshida, M.: A multilingual analysis of pro-Russian misinformation on twitter during the russian invasion of Ukraine. Scientific Reports

14, 10155 (2024) https://doi.org/10.1038/s41598-024-60653-y

[36] Blei, D.M.: Surveying a suite of algorithms that offer a solution to managing large document archives. In: Computer Science, Princeton University, pp. 77–84 (2012). https://cs.princeton.edu

[37] Liu, L., Tang, L., Dong, W., Yao, S., Zhou, W.: An overview of topic modeling and its current applications in bioinformatics. SpringerPlus **5**(1), 1608 (2016) https://doi.org/10.1186/s40064-016-3252-8

[38] Grimmer, J., Stewart, B.M.: Text as data: The promise and pitfalls of automatic content analysis methods for political texts. Political Analysis **21**(3), 267–297 (2013) https://doi.org/10.1093/pan/mps028

[39] Rüdiger, M., Antons, D., Joshi, A.M., Salge, T.O.: Topic modeling revisited: New evidence on algorithm performance and quality metrics. PLoS ONE **17**(4), 0266325 (2022) https://doi.org/10.1371/journal.pone.0266325

[40] Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J., Blei, D.: Reading tea leaves: How humans interpret topic models. In: Advances in Neural Information Processing Systems, vol. 22. Curran Associates, Inc., ??? (2009)

[41] Ying, L., Montgomery, J.M., Stewart, B.M.: Topics, concepts, and measurement: A crowdsourced procedure for validating topics as measures. Political Analysis **30**(4), 570–589 (2022) https://doi.org/10.1017/pan.2021.33

[42] Roberts, M.E., Stewart, B.M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S.K., Albertson, B., Rand, D.G.: Structural topic models for open-ended survey responses. American Journal of Political Science **58**(4), 1064–1082 (2014) https://doi.org/10.1111/ajps.12103

[43] Li, Y., Scarton, C., Song, X., Bontcheva, K.: Classifying COVID-19 Vaccine Narratives (2022). https://arxiv.org/abs/2207.08522

[44] Song, X., Petrak, J., Jiang, Y., Singh, I., Maynard, D., Bontcheva, K.: Classification aware neural topic model for COVID-19 disinformation categorisation. PLoS ONE **16**(2), 0247086 (2021) https://doi.org/10.1371/journal.pone.0247086

[45] Kotseva, B., Vianini, I., Nikolaidis, N., Faggiani, N., Potapova, K., *et al.*: Trend analysis of COVID-19 mis/disinformation narratives – a 3-year study. PLOS ONE **18**(11), 0291423 (2023) https://doi.org/10.1371/journal.pone.0291423

[46] Piskorski, J., Mahmoud, T., Nikolaidis, N., Campos, R., Jorge, A.M., Dimitrov, D., Da San Martino, G., *et al.*: SemEval 2025 Task 10: Multilingual characterization and extraction of narratives from online news. In: Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025), pp. 2610–2643. Association for Computational Linguistics, Vienna, Austria (2025)

[47] NATO Strategic Communications Centre of Excellence: Narrative detection and topic modelling in the baltics. Technical report, NATO Strategic Communications Centre of Excellence, Riga (2023)

[48] Charte, F., Rivera, A.J., Jesus, M.J., Herrera, F.: Addressing imbalance in multilabel classification: Measures and random resampling algorithms. Neurocomputing **163**, 3–16 (2015) https://doi.org/10.1016/j.neucom.2014.08.091

[49] Rizgelienė, I., Zubaitienė, V., Maliukevičius, N., Marcinkevičius, V.: Haltprop: Human-annotated lithuanian textual corpus for propaganda narratives and techniques. Scientific Data **13**, 47 (2026) https://doi.org/10.1038/s41597-025-06367-w

[50] 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. arXiv preprint arXiv:1911.02116 (2019). v2, 2020

[51] Hu, J., Ruder, S., Siddhant, A., Neubig, G., Firat, O., Johnson, M.: Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization. In: Proceedings of the 37th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 119, pp. 4411–4421 (2020). https://proceedings.mlr.press/v119/hu20b.html

[52] Ulčar, M., Robnik-Šikonja, M.: Training dataset and dictionary sizes matter in BERT models: The case of baltic languages. arXiv preprint arXiv:2112.10553 (2021)

[53] Ulčar, M., Robnik-Šikonja, M., EMBEDDIA: LitLat BERT: Model Card. https://huggingface.co/EMBEDDIA/litlat-bert. XLM-RoBERTa-base configuration; 12 layers, 12 heads; vocab size 84,201 (2020)

[54] Ulčar, M.: LitLat BERT. CLARIN-LT Repository. https://clarin.vdu.lt/xmlui/handle/20.500.11821/42 (2020)

[55] Vlad, G., Tanaşe, M., Onose, C., Cercel, D.: Sentence-level propaganda detection in news articles with transfer learning and bert-bilstm-capsule model. In: Proceedings of the 2nd Workshop on NLP for Internet Freedom (NLP4IF) at EMNLP-IJCNLP, pp. 148–154 (2019)

[56] Khosla, S., Joshi, R., Dutt, R., Black, A.W., Tsvetkov, Y.: LTI@CMU at semeval-2020 task 11: Incorporating multi-level features for multi-granular propaganda span identification. In: Proceedings of SemEval 2020 (2020)

[57] Ahmad, P.N., Khan, K.: Propaganda fragment detection and auto-fact-check in bi-lingual corpus. EAI Endorsed Transactions on Smart Cities (2023)

[58] Rai, N., Kumar, D., Kaushik, N., Raj, C., Ali, A.: Fake news classification using

transformer-based enhanced LSTM and BERT. International Journal of Cognitive Computing in Engineering **3**, 98–105 (2022) https://doi.org/10.1016/j.ijcce.2022.03.003

[59] Roy, K.S., Bina, F.A.: Tweetguard: Combining transformer and bi-LSTM architectures for fake news detection in large-scale tweets. International Journal of Data Science and Analysis **11**(2), 23–45 (2025) https://doi.org/10.11648/j.ijdsa.20251102.12

[60] Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (2019). ICLR. https://openreview.net/forum?id=Bkg6RiCqY7

[61] Vig, J., Belinkov, Y.: Analyzing the structure of attention in a transformer language model. In: Linzen, T., Chrupała, G., Belinkov, Y., Hupkes, D. (eds.) Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, pp. 63–76. Association for Computational Linguistics, Florence, Italy (2019). https://doi.org/10.18653/v1/W19-4808 . https://aclanthology.org/W19-4808