INFORMATICA, 2026, Vol. 0, No. 0, 1–30

DOI: https://doi.org/10.15388/26-INFOR633

Transformer-Based Detection of Propaganda Techniques in a Low-Resource Language: A Case Study in Lithuanian

Ieva RIZGELIENĖ*, Paulius ZARANKA, Gražina KORVEL, Virginijus MARCINKEVIČIUS

[1] *Vilnius University, Faculty of Mathematics and Informatics, Institute of Data Science and Digital Technologies, Akademijos st. 4, Vilnius, LT-08412, Lithuania*
e-mail: ieva.rizgeliene@mif.vu.lt, paulius.zaranka@mif.vu.lt, grazina.korvel@mif.vu.lt, virginijus.marcinkevicius@mif.vu.lt

Received: April 2026; accepted: June 2026

Abstract. Propaganda techniques are a key tool for creating misleading content, often disseminated in native languages to increase their impact. Therefore, it is increasingly important to develop detection models not only for high-resource languages but also for low-resource languages, which still face significant limitations in propaganda detection. This study presents the first approach to automated propaganda technique detection in Lithuanian using the HALT-PROP corpus. We adapt the standard framework to account for frequent overlap between techniques. Experiments with the Lithuanian transformer LT-MLKM-modernBERT show that BILOU tagging improves span identification, while sentence classification based on span-level information enhances technique detection for most techniques. The results also indicate that training separate binary classifiers is more effective than multi-label classification in this setting. Overall, the proposed approach outperforms GPT-5.3 on most techniques and provides a strong baseline for propaganda technique detection in Lithuanian.

Key words: propaganda technique detection, low-resource language, transformers.

1. Introduction

Over time, propaganda has increasingly shifted toward digital dissemination through social media and online news platforms. As propaganda often relies on rhetorical manipulation, sometimes incorporating factual elements while still misleading audiences through specific persuasive techniques, it is important not only to identify whether an article contains propagandistic content, but also to detect the specific techniques employed. Moreover, recent experimental research demonstrates that labelling content as propaganda and explicitly highlighting the rhetorical techniques used can significantly reduce users' intentions to share such content online (Jose *et al.*, 2025).

*Corresponding author.

With the advancement of machine learning technologies, particularly in natural language processing, substantial efforts have been devoted to automated propaganda detection. This is reflected in a series of shared tasks and studies focusing on propaganda and persuasion technique detection (Rashkin *et al.*, 2017; Barrón-Cedeño *et al.*, 2019; Da San Martino *et al.*, 2019, 2020; Dimitrov *et al.*, 2021; Piskorski *et al.*, 2023; Dimitrov *et al.*, 2024; Alam *et al.*, 2022; Hasanain *et al.*, 2024; Moral *et al.*, 2023, 2024; Horák *et al.*, 2024). Despite this progress, significant challenges remain, particularly for low-resource languages. A key limitation is the scarcity of essential resources, particularly annotated corpora tailored to these languages.

To address this gap, particularly for the Lithuanian language, the first human-annotated corpus, HALT-PROP, was released in 2025 (Rizgelienė *et al.*, 2025). The corpus contains Lithuanian texts annotated by five experts for propaganda techniques and narratives using a cross-annotation methodology, allowing multiple techniques and narratives to be assigned to the same text. As this is a newly released resource, there is currently no research on propaganda technique detection based on this corpus, nor are there any dedicated studies on propaganda technique detection for the Lithuanian language.

The main goal of this research is to develop the first approach for propaganda techniques detection in Lithuanian. To achieve this, the study addresses the following research questions:

1. Does incorporating span boundary information and increasing the maximum input sequence length improve propaganda span identification performance in Lithuanian?
2. How does incorporating span-level information, indicating where propaganda techniques occur, affect the performance of propaganda techniques classification?
3. Does modelling propaganda techniques as independent binary classification tasks improve detection performance compared to multi-class classification in scenarios with high label overlap?

In this study, a *span* refers to an uninterrupted fragment of text annotated as containing at least one propaganda technique.

To address these questions, we first perform an exploratory analysis of the HALT-PROP corpus. Based on these insights, we propose a methodology for detecting propaganda techniques in Lithuanian news articles that reflects the corpus characteristics. We then leverage the monolingual Lithuanian transformer model LT-MLKM-modernBERT to conduct experiments on both span identification and techniques classification. We explore different training setups, including the use of span-level information, as well as binary and multi-class classification strategies. Finally, we evaluate the approach under multiple conditions, including gold and predicted spans, and compare the results with zero-shot and few-shot performance of large language model.

2. Related Work

Until 2019, propaganda detection research primarily focused on document-level classification (Rashkin *et al.*, 2017; Barrón-Cedeño *et al.*, 2019). However, because propaganda often contain both propagandistic and non-propagandistic content, assigning a single label to

an entire document is overly coarse and has motivated a shift toward fine-grained analysis. One of the early approaches to propaganda technique detection focused on sentence-level classification and defined two tasks based on an expert-annotated English dataset covering 18 propaganda techniques: (i) Sentence-Level Classification (SLC) and (ii) Fragment-Level Classification (FLC) (Da San Martino *et al.*, 2019). Subsequently, the SemEval-2020 Task 11 shared task was introduced (Da San Martino *et al.*, 2020), which extended propaganda technique detection from sentence-level classification to span-level analysis. It defined two subtasks: (i) Span Identification (SI), a binary sequence labelling task aimed at identifying text fragments containing at least one propaganda technique, and (ii) Technique Classification (TC), a multi-class classification task that assigns a specific propaganda technique label to the identified spans. Later SemEval tasks further expanded the scope to include propaganda technique detection in multimodal content (text and images) (Dimitrov *et al.*, 2021), multilingual analysis (Piskorski *et al.*, 2023), and multilingual meme analysis (Dimitrov *et al.*, 2024). Beyond SemEval, several initiatives have also explored non-English and multilingual settings, including Arabic shared tasks on multi-label technique classification and persuasion detection in tweets (Alam *et al.*, 2022; Hasanain *et al.*, 2024), as well as the DIPROMATS shared tasks, which address propaganda identification, characterization, technique classification, and narrative detection in English and Spanish (Moral *et al.*, 2023, 2024).

However, despite several initiatives extending propaganda technique detection beyond the English language, research in this area still faces significant limitations, particularly for low-resource languages. One of the few dedicated efforts is the work of Horák *et al.* (2024), which introduces an annotated corpus for the Czech language and proposes an initial approach to propaganda technique detection. This approach combines stylometric features with representations from pretrained transformer models. However, it focuses on document-level classification, identifying whether an article contains specific techniques. While this represents an improvement over binary propaganda detection, it still lacks fine-grained explainability, as it does not indicate where in the text the techniques occur.

In addition, some studies have also explored technique classification for non-English languages, such as Arabic (Alam *et al.*, 2022), which is generally considered a high-resource language, as well as multilingual approaches (Piskorski *et al.*, 2023) covering languages such as Italian, Russian, French, German, and Polish. However, none of the existing resources include the Lithuanian language, nor do they cover languages spoken in the Baltic countries. More broadly, languages spoken in countries neighbouring Russia, as well as in former Eastern Bloc countries, remain severely underrepresented, with entire regions lacking datasets and models for propaganda identification, despite being primary targets of information warfare. This work addresses this gap by proposing a method for propaganda technique detection in the Lithuanian language, representing the first such approach in the Baltic region, as well as one of the first approaches in general for a low-resource language spoken in the Russian neighbourhood, and one of the few in European languages.

3. Data

In this study, we used the first Lithuanian corpus for propaganda narratives and techniques (Rizgelienė *et al.*, 2025). The corpus consists of two complementary datasets: (1) 2 870 news articles manually labelled by five experts at the article level to identify the presence of propaganda; and (2) a subset of 1 000 articles annotated for specific propaganda techniques and narratives using a cross-annotation approach, in which each article was independently annotated by two of the five experts, and the final annotation was confirmed through pairwise discussion. In this study, we focus only on propaganda techniques annotations. Our selected corpus is annotated for the following ten propaganda techniques:

1. *Emotional Expression*. Intentionally uses emotionally charged language (e.g. fear, anger, pride, sympathy) to provoke strong feelings and influence audience beliefs or actions. Often avoids logical reasoning and relies on exaggeration, personal attacks, or vague but positive terms to shape perception.
2. *Whataboutism/Red Herring/Straw Man*. Distracts from the main issue by shifting blame or criticism to others (Whataboutism), introducing irrelevant information or arguments (Red Herring), or misrepresenting an opponent's view by exaggerating, distorting, or oversimplifying it to attack a weaker version (Straw Man). These strategies serve to divert or deflect attention from the core argument.
3. *Simplification*. Deliberately reduces complex issues to overly simple explanations by attributing blame or responsibility to a single cause or group, framing problems as having only two opposing options, using stereotyped phrases that shut down deeper thinking, and relying on short, catchy slogans that appeal more to emotions than logic. These strategies limit critical analysis and obscure the true complexity of issues.
4. *Intentional Vagueness* (*Obfuscation*). Uses ambiguous or imprecise language to obscure meaning, allowing multiple interpretations and helping avoid accountability or direct scrutiny.
5. *Appeal to Authority*. Refers to perceived authoritative figures or institutions to legitimize a claim, implying it is true based solely on the authority's status, often without supporting evidence.
6. *Flag-Waving*. Promotes a position by invoking patriotism or national pride, suggesting the idea serves the country's interests, even in the absence of a clear rationale or evidence.
7. *Bandwagon*. Encourages alignment with a belief or action by implying widespread acceptance. Leverages social pressure and the desire to conform to persuade individuals to adopt the majority view.
8. *Doubt/Smears*. Seeks to undermine credibility by casting suspicion or attacking character – either subtly (Doubt) or directly through baseless accusations or insinuations (Smears) – without presenting concrete evidence.
9. *Reduction ad Hitlerum/Stalinum*. Discredits a person, idea, or group by associating them with historically vilified figures (e.g. Hitler, Stalin), appealing to emotion rather than addressing the actual argument.

10. *Repetition*. Reinforces a message through frequent repetition. Over time, repeated statements may appear more familiar and thus more believable, even in the absence of evidence, a psychological effect known as the "illusion of truth."

3.1. *Dataset Analysis*

3.1.1. *Techniques Distribution*

We conducted an exploratory analysis of the HALT-PROP dataset to examine the distribution of propaganda techniques and the density and coverage of annotated spans (see Fig. 1). The results show that the dataset is imbalanced, with a small number of techniques dominating the corpus in both frequency and coverage. Additionally, techniques differ substantially in typical span length and annotation density, reflecting their different manipulation strategies: some are expressed through short lexical cues, while others appear as longer discourse-level segments. Below, we summarize the main findings for the techniques.

- **Emotional Expression**, **Simplification**, and **Doubt/Smears** are the three most frequent propaganda techniques in the corpus. Emotional Expression appears in 820 articles, Simplification in 689 articles, and Doubt/Smears in 649 articles. These techniques also have the highest text coverage: on average, 39.5% of the text is annotated with Emotional Expression, 23.2% with Simplification, and 22.8% with Doubt/Smears in articles where they appear. They also show relatively high span densities, with Emotional Expression appearing in 5–6 spans per article on average, Simplification in 4 spans, and Doubt/Smears in 3–4 spans. All three techniques have moderate-to-long span lengths compared to other techniques in the corpus.

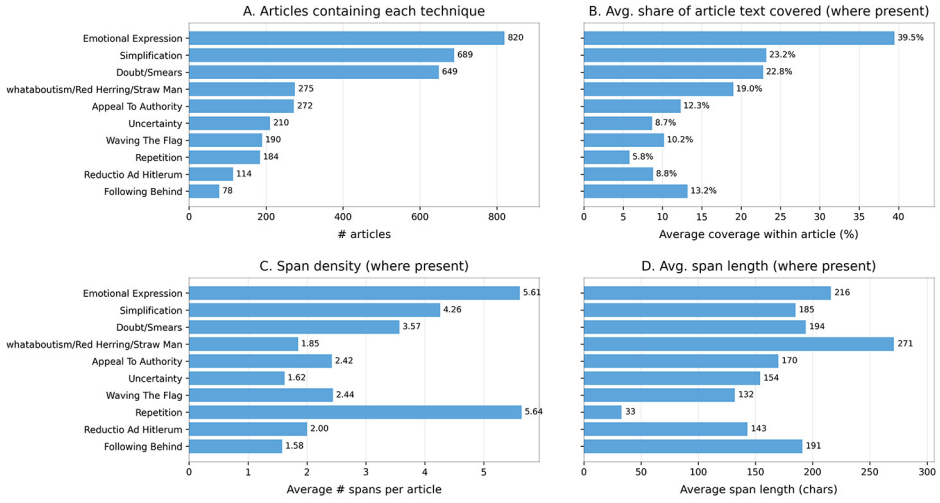

Fig. 1. Statistics of propaganda techniques in the HALT-PROP corpus. (A) Number of articles containing each technique. (B) Average share of article text covered by the technique. (C) Average number of spans per article. (D) Average span length in characters.

- **Whataboutism/Red Herring/Straw Man**, **Appeal to Authority**, and **Uncertainty** appear with medium to lower frequency in the corpus. Whataboutism/Red Herring/Straw Man occurs in 275 articles, Appeal to Authority in 272 articles, and Uncertainty in 210 articles. Among these techniques, Whataboutism/Red Herring/Straw Man has the highest text coverage, covering approximately 19% of the article text on average, while Appeal to Authority covers about 12.3% and Uncertainty 8.7%. The span density is similar for Whataboutism/Red Herring/Straw Man and Uncertainty, which, on average, appear in about 1–2 spans per article, whereas Appeal to Authority has a slightly higher span density, appearing in approximately 2–3 spans per article. In terms of span length, Whataboutism/Red Herring/Straw Man has the longest spans, averaging about 271 characters. This is likely because the goal of this technique is to redirect attention from the main issue or topic, which often requires longer text segments to shift the context.
- **Waving the Flag** and **Repetition** occur with relatively lower frequency, appearing in 190 and 184 articles, respectively. These techniques show distinct annotation patterns. In particular, Repetition stands out among all techniques, with the smallest text coverage (5.84%), the highest span density (approximately 5–6 spans per article), and the shortest spans (an average length of only 33 characters). This reflects the main characteristic of this technique: repeating the same message, phrase, or word multiple times within an article rather than expressing a complete argument or cue in a single span. Waving the Flag shows a pattern more similar to other techniques, with relatively low coverage (10.2%), a moderate span density (approximately 2–3 spans per article), and medium-length spans (about 143 characters on average).
- **Reductio ad Hitlerum** and **Following Behind** are the least frequent techniques in the corpus. Reductio ad Hitlerum appears in 114 articles, while Following Behind appears in only 78 articles. Following Behind has higher text coverage (13.2%) compared to Reductio ad Hitlerum (8.8%). The two techniques have similar span densities, typically appearing in about 1–2 spans per article. However, Following Behind generally has longer spans (about 191 characters on average), whereas Reductio ad Hitlerum has moderately long spans (about 143 characters on average).

3.1.2. *Span Overlap Analysis*

To better understand how propaganda techniques interact, we conducted a span-overlap analysis. Since a single phrase can express multiple manipulation strategies, overlap analysis helps identify which techniques commonly co-occur and which appear more independently. Figure 2 shows overall and pairwise overlap between techniques at the character level. The results indicate that overlap is common: for most techniques, more than 50% of spans overlap with at least one other technique, although the strength of overlap varies. Emotional Expression shows moderate overall overlap (46.7%), but pairwise analysis reveals that many techniques most frequently co-occur with it, which is expected given that it is the most common technique in the HALT-PROP corpus. The other frequent techniques, Simplification and Doubt/Smears, show slightly higher overall overlap (53.7% and 52.9%, respectively), but their pairwise overlaps are generally weaker than those involving Emotional Expression.

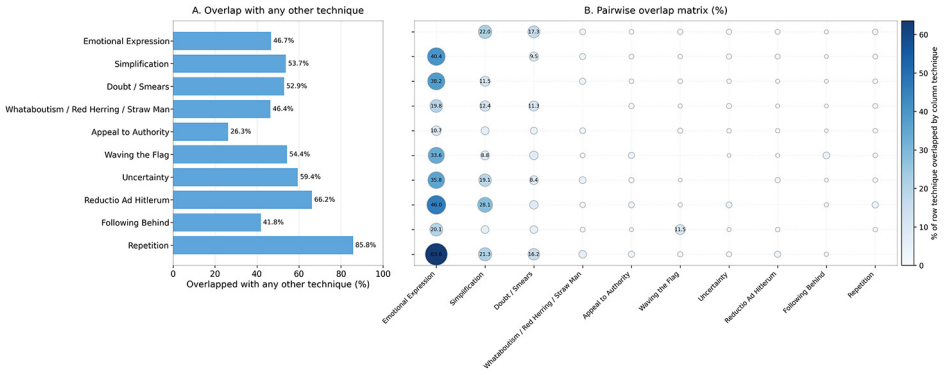

Fig. 2. Overlap between propaganda techniques at the character-span level. (A) Percentage of annotated characters for each technique that overlap with at least one other technique. (B) Pairwise overlap matrix showing the percentage of characters of the row technique that overlap with the column technique. Bubble size and colour indicate the magnitude of overlap.

Two notable exceptions emerge. Repetition has extremely high overlap (85.8%), indicating that its spans almost always co-occur with other techniques, most often with Emotional Expression (63.6%). In contrast, Appeal to Authority is relatively independent, with only 26.3% of its spans overlapping with other techniques.

Overall, Emotional Expression is the most common overlapping technique, followed by Simplification and Doubt/Smears. An additional pattern appears for Following Behind, which frequently overlaps with Waving the Flag, likely reflecting their shared appeals to collective identity or patriotism. Taken together, the analysis shows that multi-technique spans are common in the dataset and that overlap is a regular characteristic of the annotations.

3.1.3. *Sentence-Level Analysis*

As the previous analysis showed that technique spans are generally annotated as longer segments rather than short phrases marking only a few terms (with the exception of Repetition), we also conducted a sentence-level analysis to examine annotation coverage at the sentence level. To assess how much of each sentence is annotated, we measured the percentage of characters covered by technique spans within annotated sentences (see Fig. 3). The results show that the average coverage exceeds 78% for all techniques except Repetition, and for the most frequent techniques: Emotional Expression, Simplification, and Doubt/Smears – it exceeds 90%. This indicates that annotations are typically applied at the sentence level, capturing broader rhetorical context rather than isolated phrases or individual words. The main exception is Repetition, which often consists of short repeated words or phrases and therefore covers a smaller portion of the sentence.

We also analysed the distribution of techniques at the sentence level by measuring the proportion of sentences expressing each technique across the corpus and within sentences containing annotated propaganda spans. In the latter case, only sentences containing at least one propaganda technique were considered, excluding non-propagandistic sentences.

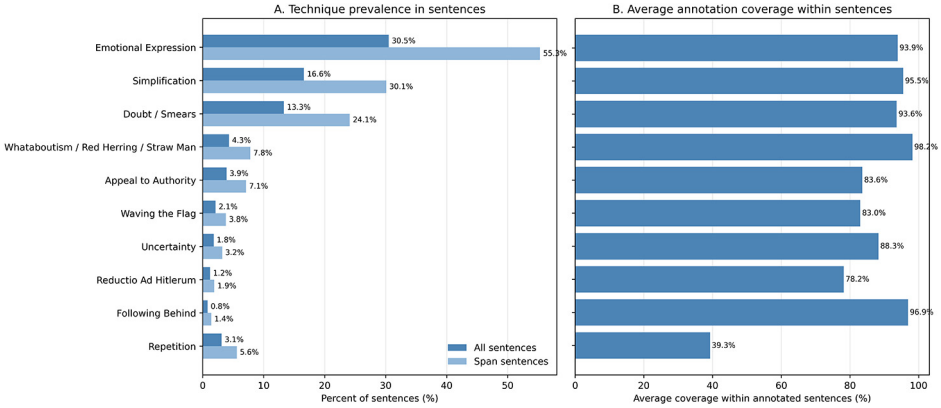

Fig. 3. Distribution and coverage of propaganda techniques at the sentence level. (A) Percentage of sentences containing each technique, shown for all sentences and for sentences that include annotated spans. (B) Average proportion of sentence text covered by the annotated technique within sentences where it appears.

The results again confirm the imbalance of the dataset. Restricting the analysis to span sentences slightly improves the balance. This effect is most evident for Emotional Expression, which appears in 30.5% of all sentences in the corpus but rises to 55.3% when only span sentences are considered. For rarely annotated techniques such as Uncertainty, Reductio ad Hitlerum, and Following Behind, the difference between the all and span sentences is small, and each still accounts for less than 5% of sentences.

3.1.4. *Key Findings from the Dataset Analysis*

After a detailed analysis of the HALT-PROP corpus, several main insights can be drawn:

- The corpus is generally imbalanced, with **Emotional Expression**, **Simplification**, and **Doubt/Smears** as the dominant techniques. These techniques are not only the most frequent, but also have the highest text coverage. Among them, Emotional Expression has the highest share in the corpus, and other techniques often overlap with it.
- Overlap between techniques is very common in the annotations, meaning that a single annotated span often contains multiple propaganda techniques at the same time.
- In general, propaganda techniques are annotated as longer fragments rather than short phrases or individual words. With the exception of *Repetition*, most techniques are annotated at the sentence level, with annotated spans covering more than 78% of the sentence characters on average, indicating that they are typically expressed across most of the sentence rather than through isolated lexical markers.
- *Repetition* shows a distinct annotation pattern compared to the other techniques: it has the shortest spans, the lowest text coverage, and the highest span density, reflecting its nature as a recurring word- or phrase-level phenomenon.

Based on the results of the data analysis, we decided to exclude the *Repetition* technique. This decision is motivated by the fact that *Repetition* exhibits a different usage pattern compared to the other techniques. Unlike most techniques, repetition is not defined

solely by individual annotated spans, but also by relations between multiple occurrences of the same word, phrase, or message across a document. Consequently, detecting repetition would require a different training strategy that models links between repeated elements throughout the document rather than relying only on local contextual features.

4. Methodology

4.1. *Task Formulation*

The main goal of this research is to develop the first approach for propaganda technique detection in Lithuanian. Propaganda technique detection is typically formulated following the SemEval framework (Da San Martino *et al.*, 2020), which splits the task into two subtasks: (i) span identification and (ii) technique classification. In this pipeline, the algorithm first identifies text fragments containing propaganda and then assigns a specific propaganda technique label to each detected span using multi-class classification.

However, the HALT-PROP corpus differs substantially from the datasets used in SemEval-style propaganda techniques detection frameworks. In particular, our analysis shows that propaganda techniques in HALT-PROP frequently overlap and are typically annotated as longer textual fragments, often covering large parts of sentences. In contrast, in the PTC-SemEval20 corpus, overlapping annotations are relatively rare: only about 1.8% of spans are associated with multiple techniques (Da San Martino *et al.*, 2020). Due to this small proportion of overlapping cases, the SemEval task formulation simplified the problem by treating technique classification as a single-label multi-class task. When a span was annotated with multiple techniques, the dataset included duplicate instances of the same span, each associated with one label, effectively avoiding a multi-label formulation.

Another important difference concerns the length of annotated spans. In the PTC-SemEval20 corpus, most techniques are typically expressed through short lexical cues and therefore correspond to relatively short spans. In contrast, the HALT-PROP corpus contains a much higher degree of overlap between techniques, with more than 50% of spans overlapping with other techniques, and annotations often covering larger textual units such as full clauses or sentences. Because of this high overlap and broader span coverage, applying the standard SemEval pipeline directly would be problematic. In particular, reducing the problem to a single-label classification task would fail to capture the multi-technique nature of many annotated fragments. Therefore, the detection approach must be adapted to account for the multi-label and highly overlapping structure of propaganda annotations in the HALT-PROP corpus.

Based on the insights discussed earlier, we adopt the standard two-subtask framework consisting of *span identification* and *technique classification*, while incorporating adaptations that account for the nuances of the HALT-PROP corpus. For the span identification task, we follow the same standard formulation used in prior approaches. However, the second subtask, technique classification, is modified to address the high degree of overlap

between techniques and the longer annotated fragments present in our corpus. As a result, our approach consists of the following two subtasks:

1. ***Span Identification***. Given a text, the task is to identify spans that contain at least one propaganda technique. This task is formulated as a token-level sequence tagging problem.
2. ***Technique Classification***. Given spans identified as containing propaganda techniques, the task is to determine which specific technique is expressed in each span. Due to the high overlap between techniques and the presence of long annotated fragments that often cover entire sentences in our corpus, we fine-tune a separate model for each technique. Consequently, the task is formulated as a binary sentence classification problem, where a sentence is assigned label 1 if it contains the target technique and 0 otherwise.

In addition to developing models for the selected subtasks, we also investigate several task-specific research questions and modelling decisions. In particular, we examine different sequence tagging approaches for span identification and explore alternative formulations of the technique classification task, including whether sentence classification should be performed on annotated spans or on all sentences, as well as whether a binary or multi-class formulation is more appropriate. Furthermore, we define the evaluation metrics used to assess the performance of the proposed approaches. The details of these investigations and modelling choices are described in the subsections dedicated to each task.

4.2. *Span Identification Task*

In the span identification task, the objective is to detect text fragments that contain any propaganda technique, without distinguishing between specific technique types. A span is defined as a continuous segment of text corresponding to the annotated region of a propaganda technique in the corpus. Figure 4 illustrates an example of spans in an article.

4.2.1. *Tagging Schemes*

Span identification task is formulated as a sequence tagging problem at the token level, where each token in the document is assigned a label indicating whether it belongs to a propaganda span. Specifically, the model predicts a label for every token in the sequence,

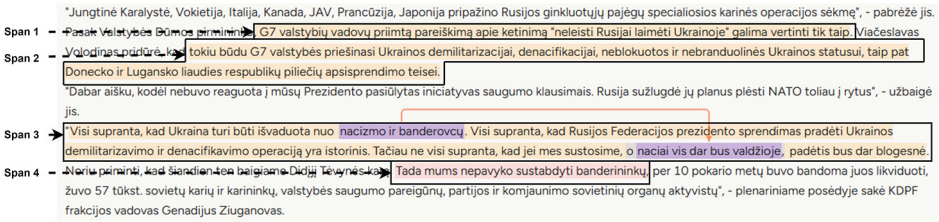

Fig. 4. Example of propaganda spans in an article. The English translation of the example is provided in Appendix A.

Table 1
Example of token-level binary and BILOU tagging for the propaganda span "*Lietuvą masiškai palieka jauni žmonės*".

Representation	Content
Original fragment	**Lietuvą masiškai palieka jauni žmonės**, dalis jų užsienyje jau įsitvirtino, ten sukūrė šeimas ir pasiliko gyventi. *Translation:* ***Young people are leaving Lithuania en masse****, some of them have already settled abroad, started families there, and stayed to live.*
Tokenized fragment	**Lietuvą, mas, iškai, palieka, jauni, žmonės**, „, dalis, jų, už, sienyje, jau, įsitvirtin, o, ten, sukūrė, šeima, s, ir, pasilik, o, gyventi, .
Binary tagging	**1 1 1 1 1 1** 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
BILOU tagging	**B I I I I L** O O O O O O O O O O O O O O O O

allowing contiguous tokens with propaganda labels to form the boundaries of propaganda fragments. We investigate two tagging schemes for the span identification task: *binary tagging* and *BILOU tagging* (Ratinov and Roth, 2009).

In the binary tagging approach, each token is assigned one of two labels indicating whether it belongs to a propaganda span or not. Specifically, tokens that are part of a propaganda fragment receive the label 1, while all other tokens receive the label 0. This formulation focuses only on identifying whether a token is inside a propaganda span, without explicitly modelling span boundaries.

In contrast, the BILOU tagging scheme explicitly represents span boundaries using five labels: *B* (Beginning), *I* (Inside), *L* (Last), *O* (Outside), and *U* (Unit). Tokens labelled *B* mark the beginning of a multi-token span, *I* indicate tokens inside the span, and *L* mark the final token of the span. The label *U* is used for spans consisting of a single token, while *O* denotes tokens outside any propaganda span. Table 1 illustrates an example of span tagging using both tagging schemes. In this example, the BILOU tagging does not include the *U* label, since it is only used when a span consists of a single token, which occurs very rarely in our corpus.

4.2.2. *Evaluation Metrics for Span Identification*

For the evaluation of the span identification task, we use an *overlap-based span* F_1 *score* (Da San Martino *et al.*, 2020), which assigns partial credit to predictions that partially overlap with the gold annotations.

Let d denote a document, $T_d = \{t_1, \ldots, t_n\}$ the set of gold technique spans in document d, and $S_d = \{s_1, \ldots, s_m\}$ the set of predicted spans. Each span is represented as a contiguous sequence of character indices. Predicted spans may overlap and are merged prior to evaluation. Precision (P) and recall (R) are computed by measuring the character-level overlap between predicted and gold spans:

$$P = \frac{1}{|S|} \sum_{d \in D} \sum_{s \in S_d} \sum_{t \in T_d} \frac{|s \cap t|}{|s|}, \tag{1}$$

$$R = \frac{1}{|T|} \sum_{d \in D} \sum_{s \in S_d} \sum_{t \in T_d} \frac{|s \cap t|}{|t|}, \tag{2}$$

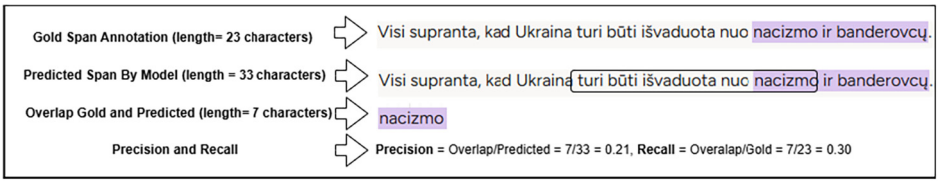

Fig. 5. Example illustrating the calculation of precision and recall for span identification based on character-level overlap between predicted and gold spans. The full sentence translates as: "Everyone understands that Ukraine must be liberated **from Nazism and Banderites**". The bold part corresponds to the gold span.

where D denotes the set of all documents in the dataset, $|S|$ and $|T|$ are the total numbers of predicted and gold spans, respectively. For a span s, $|s|$ denotes its length in characters, and $|s \cap t|$ denotes the number of overlapping characters between spans s and t.

Precision measures the proportion of predicted characters that are correctly assigned to a technique and therefore penalizes *over-tagging*. Recall measures the proportion of gold characters that are successfully recovered by the model and therefore penalizes *under-tagging*. Figure 5 illustrates an example showing how precision and recall are calculated, providing a clearer understanding of what the metric measures. The final evaluation score is computed as the harmonic mean of precision and recall:

$$F_1 = \frac{2PR}{P+R}. \tag{3}$$

4.3. *Technique Classification Task*

This task is formulated as a binary sentence classification problem. We investigate two settings: (i) a setting without span information, in which all sentences in an article are considered, and (ii) a setting with span information, in which only sentences containing annotated propaganda spans, i.e. text fragments labelled with at least one propaganda technique, are considered.

The data for this task is prepared as follows. First, documents are split into sentences. Then, binary labels are assigned for each propaganda technique separately. In the span-sentence setting, sentences that do not overlap with any annotated span are excluded; in other words, sentences that receive zero labels for all techniques are removed. An example of the data preparation process for this task is illustrated in Fig. 6.

In this task, we investigate several research questions to assess whether formulating technique detection as a sentence-level classification task is appropriate for our corpus. First, we evaluate the effect of span-based filtering by comparing models trained on all sentences with models trained only on sentences containing propaganda spans. Second, we examine whether a binary classification formulation is more suitable than a multi-class approach. To this end, we also fine-tune models in a multi-class setting, where each sentence may receive multiple technique labels.

Overall, for the technique classification task we fine-tune the same transformer-based model under several experimental configurations:

Annotated text for propaganda techniques

Kol ukrainiečiai didvyriškai gina savo Tėvynę ir mus, kol visa lietuvių tauta grumiasi informaciniame kare, leftistiniai išdavikai smogia Lietuvai į nugarą tautos branduolį – šeimą – griaunančiu Partnerystės įstatymu . Lyg to būtų maža, jie prigimtinės šeimos gynėjus tapatina su Kremliaus režimu ir asmeniškai – su Vladimiru Putinu.
Girdi, jis pasisako prieš genderizmą, tad ir visi, kas prieštarauja genderizmui, sėdi už vieno stalo su Maskvos agresoriumi.
Tikrovė – visiškai priešinga. Ukrainoje nėra partnerystės įstatymo. Kaip jo nėra ir mūsų strateginių partnerių šalyse – Lenkijoje, Latvijoje, Gruzijoje. Ukrainiečiai grumiasi ne už genderizmą ar globalizmą, bet už savo tautinę valstybę, už savo šeimas, už savo laisvę.

Sentence labels including all sentences

Sentence	Simplification	Emotional Expression	Flag-Waving
Kol ukrainiečiai didvyriškai gina savo Tėvynę ir mus, kol visa lietuvių tauta grumiasi informaciniame kare, leftistiniai išdavikai smogia Lietuvai į nugarą tautos branduolį – šeimą – griaunančiu Partnerystės įstatymu.	1	1	1
Lyg to būtų maža, jie prigimtinės šeimos gynėjus tapatina su Kremliaus režimu ir asmeniškai – su Vladimiru Putinu.	0	0	1
Girdi, jis pasisako prieš genderizmą, tad ir visi, kas prieštarauja genderizmui, sėdi už vieno stalo su Maskvos agresoriumi.	1	0	0
Tikrovė – visiškai priešinga.	1	0	0
Ukrainoje nėra partnerystės įstatymo.	0	0	0
Kaip jo nėra ir mūsų strateginių partnerių šalyse – Lenkijoje, Latvijoje, Gruzijoje.	0	0	0
Ukrainiečiai grumiasi ne už genderizmą ar globalizmą, bet už savo tautinę valstybę, už savo šeimas, už savo laisvę.	1	0	0

Sentence labels including span sentences

Sentence	Simplification	Emotional Expression	Flag-Waving
Kol ukrainiečiai didvyriškai gina savo Tėvynę ir mus, kol visa lietuvių tauta grumiasi informaciniame kare, leftistiniai išdavikai smogia Lietuvai į nugarą tautos branduolį – šeimą – griaunančiu Partnerystės įstatymu.	1	1	1
Lyg to būtų maža, jie prigimtinės šeimos gynėjus tapatina su Kremliaus režimu ir asmeniškai – su Vladimiru Putinu.	0	0	1
Girdi, jis pasisako prieš genderizmą, tad ir visi, kas prieštarauja genderizmui, sėdi už vieno stalo su Maskvos agresoriumi.	1	0	0
Tikrovė – visiškai priešinga.	1	0	0
Ukrainiečiai grumiasi ne už genderizmą ar globalizmą, bet už savo tautinę valstybę, už savo šeimas, už savo laisvę.	1	0	0

Fig. 6. Illustration of converting span-level annotations into sentence-level binary labels. Highlighted spans correspond to annotated techniques. The tables show the resulting sentence-level labels when using all sentences and when using only sentences containing annotated spans. The English translation of the example is provided in Appendix B.

- Fine-tuning the transformer model separately for each technique using all sentences, formulated as a binary classification task.
- Fine-tuning the transformer model separately for each technique using only sentences that contain propaganda spans (i.e. sentences containing at least one propaganda technique). When a span starts or ends in the middle of a sentence, the entire sentence is still used for classification.
- Fine-tuning the transformer model jointly for all techniques as a multi-label classification task using all sentences.
- Fine-tuning the transformer model jointly for all techniques as a multi-label classification task using only sentences containing propaganda spans.

4.3.1. *Evaluation Metrics for Techniques Classification on Sentence Level*

To compare the binary and multi-label approaches, predictions are evaluated separately for each propaganda technique as a binary classification task. We use *macro-F1* as the main evaluation metric. For each technique, macro-F1 is computed as the unweighted average of the F1 scores for the positive and negative classes:

$$\mathrm{F1}_{\mathrm{macro}} = \frac{1}{2}(\mathrm{F1}_{+} + \mathrm{F1}_{-}), \tag{4}$$

where $F1_+$ denotes the F1 score for the positive class, corresponding to sentences containing the target technique, and $F1_-$ denotes the F1 score for the negative class, corresponding to sentences not containing the target technique. This metric is more suitable than accuracy in the presence of class imbalance, since it assigns equal importance to both classes.

Accuracy is also reported as a supplementary metric to provide a general indication of the proportion of correctly classified instances.

4.3.2. *Multi-Class Case Evaluation*

Since we also fine-tune a propaganda technique detection model in a multi-class setting, we use multi-class evaluation metrics during training, specifically for monitoring model performance and selecting the best model. In particular, the performance of the multi-class model is monitored using the macro-F1 score computed over the positive classes, which is defined as:

$$F1_{\text{macro}} = \frac{1}{K} \sum_{k=1}^{K} F1_k, \tag{5}$$

where $F1_k$ is the F1-score computed for technique k, and K is the total number of techniques.

4.4. *Transformer Model*

In this study, we employ the *LT-MLKM-modernBERT* model (State Digital Solutions Agency, 2025). *LT-MLKM-modernBERT* is a monolingual, encoder-only transformer based on the *ModernBERT-base* architecture and specifically pretrained for the Lithuanian language. Pretraining was conducted on approximately 1.87 billion words (around 49 billion tokens) collected from diverse Lithuanian sources, including news, legal, academic, and public sector texts. The model comprises 22 Transformer encoder layers with 12 attention heads and a hidden representation size of 768 dimensions, resulting in approximately 149 million parameters. It employs a custom Lithuanian tokenizer with a vocabulary of 64 000 tokens and supports a maximum input sequence length of 8 192 tokens. We selected this model because it is currently the largest publicly available Lithuanian language model and supports the longest input sequence length among Lithuanian pretrained transformer models, making it particularly suitable for processing longer textual contexts.

To assess whether *LT-MLKM-modernBERT* indeed provides superior performance, we additionally evaluate two alternative transformer models specifically on the span identification task. These models are included solely for comparative purposes and are not used in other tasks within this study. In particular, we consider two multilingual models that support the Lithuanian language: XLM-RoBERTa (Conneau *et al.*, 2020) and LitLat-BERT (Ulčar and Robnik-Šikonja, 2020). This comparison allows us to determine whether the selected monolingual model offers a tangible advantage over widely used multilingual alternatives for span identification.

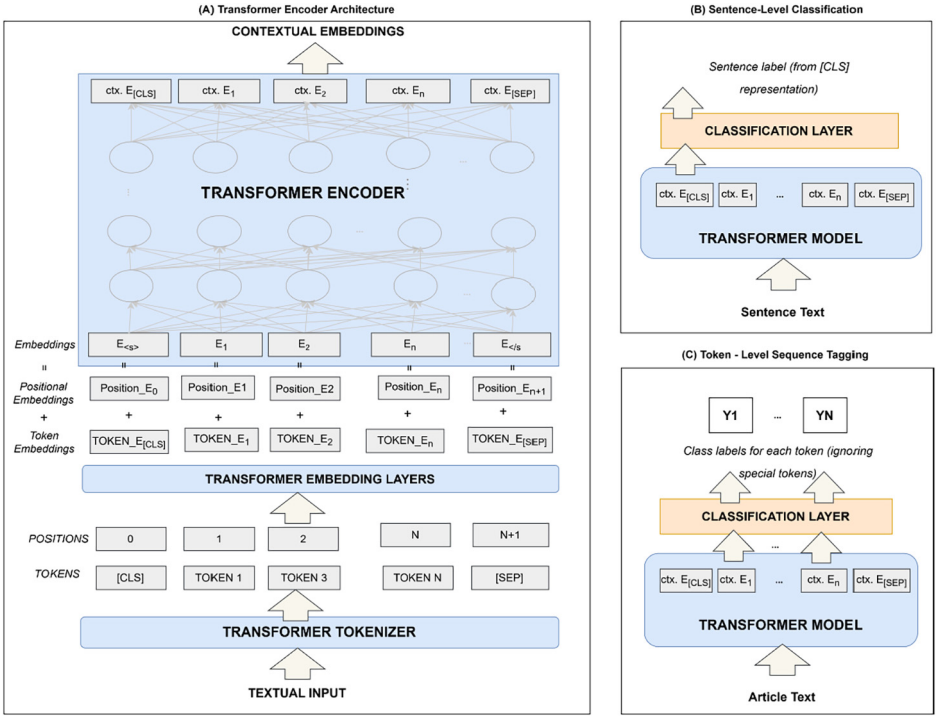

Fig. 7. Overview of a Transformer encoder architecture and its application. (A) Encoder producing contextual token representations. (B) Sentence classification using the [CLS] representation. (C) Token-level sequence tagging.

Figure 7 illustrates the general Transformer encoder architecture and its adaptations for the tasks analysed in this study: sequence tagging and sentence classification. In both tasks, the input data undergoes the same processing stages, including tokenization, embedding, encoding, and generation of contextual embeddings. The primary difference lies in the final stage, where task-specific classification layers are applied. In the following section, we describe the architecture in detail.

Transformer Encoder Architecture. A Transformer encoder maps an input token sequence $x = (x_0, x_1, \ldots, x_n, x_{n+1})$ into contextualized hidden representations $H = (h_0, h_1, \ldots, h_n, h_{n+1})$, where $x_0 = \text{[CLS]}$ and $x_{n+1} = \text{[SEP]}$ denote special tokens. First, the input text is tokenized and converted into token identifiers. Each token identifier is mapped to a trainable token embedding vector, which is combined with a positional embedding to encode word order information. The resulting sequence of embeddings is then passed through a stack of Transformer encoder layers.

Each encoder layer applies multi-head self-attention followed by a position-wise feed-forward network, together with residual connections and layer normalization. Self-attention allows each token representation to attend to all other tokens in the sequence, enabling the model to capture long-range dependencies and contextual interactions. As a

result, the final hidden state h_i corresponding to token x_i represents a contextual embedding that incorporates information from the entire input sequence.

Sentence-Level Classification. In the sentence classification setting, the model predicts a single label for the entire input sequence based on the contextual representation of the special classification token [CLS], i.e. the hidden state $h_{[\text{CLS}]}$. This vector serves as a fixed-dimensional representation of the whole sequence.

A classification layer maps this representation to a two-dimensional output space corresponding to binary classes. The resulting vector $z \in \mathbb{R}^2$ contains unnormalized scores for each class and is computed as:

$$z = W_{\text{sent}} h_{[\text{CLS}]} + b_{\text{sent}}, \tag{6}$$

where $W_{\text{sent}} \in \mathbb{R}^{2\times d}$, $b_{\text{sent}} \in \mathbb{R}^2$, and d is the hidden dimension of the encoder.

The predicted label distribution is obtained by applying the softmax function to these scores. During training, the model is optimized using cross-entropy loss with respect to the gold binary label.

Sequence Tagging. In the sequence tagging setting, the model predicts a label for each input token in the sequence, excluding special tokens such as [CLS] and [SEP]. Instead of using only the sequence-level representation $h_{[\text{CLS}]}$, the model uses the contextual representations of individual tokens, i.e. $h_1, h_2, \ldots, h_n$. A token-level classification layer is applied independently to each token representation:

$$z_i = W_{\text{tok}} h_i + b_{\text{tok}}, \tag{7}$$

where z_i denotes the vector of unnormalized scores for token i.

In this work, token-level labels are modelled using two different tagging formulations: a binary tagging scheme and the BILOU tagging scheme. In the binary tagging formulation, each token is assigned one of two labels indicating whether the token belongs to the target span or not. Formally, the label set is {0, 1}, where label 1 denotes that the token is part of a target span and label 0 indicates that the token does not belong to any labelled span. This formulation simplifies the sequence labelling problem by focusing only on the presence or absence of the target phenomenon at the token level.

In addition to the binary formulation, we also employ the BILOU tagging scheme. The BILOU label set includes the tag O and structured span labels of the form B-t, I-t, L-t, and U-t, where t denotes a target category. The tag B marks the beginning of a multi-token span, I marks a token inside the span, L marks the last token of the span, and U denotes a single-token span. The tag O indicates that the token does not belong to any labelled span.

Compared to binary tagging, the BILOU scheme explicitly models span boundaries, allowing the model to distinguish between the beginning, inside, and end of multi-token spans, as well as single-token spans. This provides richer structural information about entity boundaries.

5. Experimental Setup

First, we separate a test set that is used exclusively for the final evaluation in all tasks. This set remains untouched during all fine-tuning procedures and is used only for testing. Specifically, 105 articles are selected using stratified sampling by propaganda technique. The remaining data are used for model training and validation. For each task, the data are further split into training and validation sets, with 15% reserved for validation to monitor model performance. The split is performed at the article level to ensure that text chunks from the same article do not appear in both the training and validation sets, thereby preventing data leakage.

For all tasks, we use the same main fine-tuning hyperparameters. We did not perform an exhaustive hyperparameter optimization procedure, such as grid search or Bayesian optimization. Instead, the main fine-tuning hyperparameters were fixed across experiments using commonly adopted settings for transformer-based models. Specifically, all models were trained for 10 epochs using a batch size of 16 and the AdamW optimizer with a learning rate of 3×10^{-5} and a weight decay of 0.01. Full fine-tuning was applied in all experiments: all transformer encoder parameters and the task-specific classification layers were updated during training, and no layers were frozen. The reported results are based on a single fine-tuning run for each experimental configuration. These hyperparameters and fine-tuning settings were kept constant across experiments to ensure comparability between different modelling settings.

5.1. *Task 1: Span Identification*

In this task, we fine-tune the transformer model using different input sequence lengths (512, 1024, and 2048 tokens) and two tagging schemes: Binary and BILOU. In total, the model is fine-tuned six times, covering all combinations of these parameters: (512, Binary), (1024, Binary), (2048, Binary), (512, BILOU), (1024, BILOU), and (2048, BILOU).

Since some articles exceed the maximum input length, they are split into shorter textual fragments so that each fragment does not exceed this limit at the sentence level. This ensures that each fragment ends at a sentence boundary rather than in the middle of a sentence. The best-performing model is selected based on the overall span-level F1 score (see Section 4.2.2). The models are trained using the standard cross-entropy loss without class weighting. The goal of this step is to obtain the span identification model with the highest performance, which is selected based on the performance on test set.

Additionally, for comparative purposes, we fine-tune the XLM-RoBERTa and LitLat-BERT models on the same span identification task using the same parameters. These experiments are conducted solely to compare their performance with the *LT-MLKM-modernBERT* model and to assess whether the selected model provides superior results.

5.2. *Task 2: Technique Classification*

In this task, we fine-tune a Transformer model for sentence-level classification. The model is trained separately for each technique using two different training settings: (i) using all

sentences extracted from the articles, without incorporating span-level information, and (ii) using only sentences corresponding to gold-annotated spans, thereby explicitly leveraging span-level information. Overall, the model is fine-tuned for nine techniques under both settings, resulting in a total of 18 training runs.

Before fine-tuning, the data is preprocessed by splitting all articles into individual sentences. Based on the annotations, each sentence is assigned nine binary labels corresponding to the nine techniques. Each label indicates whether the respective technique appears anywhere within the sentence, where label 1 denotes the presence of the technique and label 0 indicates its absence. For the experiments using only span sentences, we remove sentences that contain only negative labels (i.e. sentences where all nine technique labels are 0).

The best performing model is selected based on the macro-averaged F1 score (see Section 4.3.1). Since most techniques are highly imbalanced, we apply a weighted cross-entropy loss during training. The only exception is the *Emotional Expression* technique, for which the class distribution is relatively balanced; therefore, the standard (unweighted) cross-entropy loss is used.

5.3. *Overall Performance Evaluation*

For the final evaluation, we assess the technique classification models in three different ways. First, we report the macro-F1 score of the sentence classification model trained using all sentences, without incorporating any span information. Second, we report sentence classification results when using gold span information. Third, we evaluate sentence classification performance using spans predicted by the best-performing span identification model obtained in Task 1 (Span Identification). This evaluation setup allows us to analyse the overall effectiveness of span-based information for technique classification and to estimate how much bias or performance variation is introduced when span identification predictions are used instead of gold annotations.

5.3.1. *Comparison with ChatGPT*

For the final evaluation, we also compare our technique classification results with the GPT-5.3 model, one of the latest GPT models. The GPT model is accessed through the agent interface and configured as a sentence-level labelling agent for each propaganda technique. We investigate two variants: zero-shot, where the prompt contains only the definition of the technique, and few-shot, where the prompt additionally includes ten examples: five sentences labelled with 1 (technique present) and five labelled with 0 (technique absent).

We aim to replicate the logic used in our approach, where each model is fine-tuned separately as a binary sentence-level classifier for each technique. Following the same setup with GPT-5.3, we provide a separate prompt for each technique and instruct the model to focus only on the specified technique during annotation. The prompts used in this approach are shown in Fig. 8.

Zero-shot Prompt

You are an expert annotator of propaganda techniques.
Analyze each sentence and assign binary labels (0 or 1) for the technique listed below.
Label = 1 if the technique is present.
Label = 0 if the technique is not present.
Only analyze the sentence itself. Do NOT infer meaning from the rest of the article. Do NOT invent labels. Do NOT output explanations.
Return JSON only.
TECHNIQUE Doubt: Undermines credibility by casting suspicion or attacking character without presenting concrete evidence.
OUTPUT FORMAT
"articleid": <articleid>, "sentid": <sentid>, "sentence": "<sentence>", "label": "Doubt": 0
Rules: - Labels must be 0 or 1 - Do not rename labels - Output valid JSON only

Few-shot Prompt

You are an expert annotator of propaganda technique.
Analyze each sentence and assign binary labels (0 or 1) for the technique listed below.
Label = 1 if the technique is present.
Label = 0 if the technique is not present.
Only analyze the sentence itself. Do NOT infer meaning from the rest of the article. Do NOT invent labels. Do NOT output explanations.
Return JSON only.
TECHNIQUE Doubt: Undermines credibility by casting suspicion or attacking character without presenting concrete evidence.
FEW-SHOT EXAMPLES
sentence: *"Tai tik dar vienas jų bandymas mus apgauti"*; label: 1
OUTPUT FORMAT
"articleid": <articleid>, "sentid": <sentid>, "sentence": "<sentence>", "label": "Doubt": 0
Rules: - Labels must be 0 or 1 - Do not rename labels - Output valid JSON only

Fig. 8. Zero-shot and few-shot prompts used for propaganda technique annotation.

6. Results

6.1. *Span Identification*

Table 2 reports the results of span identification models fine-tuned with different input sequence lengths and tagging schemes. Overall, the results on the test set clearly show that the BILOU tagging scheme consistently outperforms the Binary scheme across all input lengths. This outcome is expected, since BILOU explicitly models span boundaries by distinguishing the beginning, inside, last, and unit tokens of spans, whereas the Binary scheme only indicates whether a token belongs to a span or not. The results also indicate that increasing the maximum input sequence length does not improve performance. For both tagging schemes, the best results are achieved with the smallest input size of 512 tokens.

From the perspective of precision and recall, a consistent pattern can be observed. Binary tagging achieves higher recall than precision across all experimental settings on the test set, indicating that it favours broader coverage of gold spans rather than strict boundary accuracy. In some cases, Binary tagging even achieves higher recall than BILOU. For example, with an input length of 512 tokens, Binary tagging reaches a recall of 77.08%, compared to 71.49% for BILOU. A similar pattern appears for the input length of 1024 tokens, where Binary achieves 75.86% recall compared to 67.23% for BILOU.

Table 2
Span identification performance of LT-MLKM-modernBERT under different input sizes and tagging schemes.

Max input size	Tagging scheme	Train			Validation			Test		
		Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
512	Binary	61.34%	80.44%	69.60%	55.77%	74.92%	63.94%	55.52%	77.08%	64.55%
512	**BILOU**	96.09%	89.05%	92.44%	66.84%	71.91%	69.28%	72.41%	71.49%	**71.95%**
1024	Binary	68.46%	88.28%	77.12%	54.53%	75.93%	63.47%	55.75%	75.86%	64.27%
1024	BILOU	95.72%	78.90%	86.50%	70.86%	66.58%	68.65%	73.31%	67.23%	70.14%
2048	Binary	67.15%	90.81%	77.21%	55.89%	76.14%	64.46%	54.85%	60.85%	57.70%
2048	BILOU	94.91%	79.25%	86.38%	71.19%	70.05%	70.62%	75.09%	64.55%	69.42%

Table 3
Comparison of different models for span identification using a maximum input size of 512 tokens and the BILOU tagging scheme.

Model	Train			Validation			Test		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
LT-MLKM-modernBERT	96.09%	89.05%	92.44%	66.84%	71.91%	69.28%	72.41%	71.49%	**71.95%**
XLM-RoBERTa	59.26%	92.28%	72.17%	55.27%	79.10%	65.07%	60.44%	81.71%	69.49%
LitLatBERT	61.51%	75.02%	67.59%	59.35%	60.74%	60.03%	62.35%	67.15%	64.66%

However, BILOU tagging consistently achieves substantially higher precision across all configurations, often outperforming Binary tagging by nearly 20 percentage points. Overall, BILOU tagging demonstrates a more balanced trade-off between precision and recall, whereas Binary tagging shows a clear imbalance between these two metrics. This suggests that BILOU tagging produces more stable span predictions by simultaneously capturing a larger proportion of gold span characters while also maintaining more accurate span boundaries. In contrast, Binary tagging primarily focuses on identifying tokens that belong to gold spans but does not explicitly model span boundaries. As a result, it often predicts spans that are overly broad or include additional characters that should not belong to the span. Considering the overall performance measured by the F1 score on the test set, the BILOU tagging scheme consistently yields better results. The best performance is achieved with BILOU tagging and a maximum input length of 512 tokens, reaching an F1 score of 71.95%. Therefore, this configuration is selected as the final span identification model.

Additionally, for comparative purposes, we fine-tuned the multilingual transformer models XLM-RoBERTa and LitLatBERT using the BILOU tagging scheme and a maximum input size of 512 tokens, which is the largest supported sequence length for these models. The results, presented in Table 3, confirm that *LT-MLKM-modernBERT* outperforms these transformers and achieves the highest performance in span identification. Based on these results, the other transformer models are not used in subsequent experiments, as *LT-MLKM-modernBERT* demonstrates superior performance.

Table 4
Sentence-level classification results for all techniques under two training settings: using all sentences and using only span sentences.

Technique	Setting	Training		Validation		Testing	
		Acc.	F1	Acc.	F1	Acc.	F1
Emotional Expression	**All sentences**	76.83%	71.38%	69.09%	65.18%	70.84%	**66.08%**
	Span sentences	73.13%	72.75%	63.13%	62.61%	61.24%	60.62%
Simplification	All sentences	98.28%	97.27%	78.51%	61.11%	78.23%	60.25%
	Span sentences	60.69%	57.61%	66.53%	62.94%	70.22%	**61.70%**
Doubt	All sentences	82.14%	70.71%	76.39%	61.99%	76.18%	59.90%
	Span sentences	64.51%	58.32%	76.92%	68.45%	74.22%	**64.97%**
Whataboutism/Red Herring/Strawman	All sentences	85.86%	50.73%	89.37%	53.83%	87.75%	49.36%
	Span sentences	75.38%	51.35%	89.99%	53.87%	90.21%	**51.74%**
Appeal to Authority	All sentences	75.35%	50.97%	79.65%	52.49%	77.08%	48.85%
	Span sentences	99.34%	97.62%	91.58%	69.37%	91.54%	**67.06%**
Waving the Flag	All sentences	97.91%	80.34%	96.29%	67.62%	96.31%	68.85%
	Span sentences	99.91%	99.40%	95.38%	70.61%	94.67%	**69.85%**
Uncertainty	All sentences	96.60%	66.97%	97.05%	53.90%	97.33%	50.46%
	Span sentences	94.51%	68.36%	89.29%	52.57%	91.02%	**57.47%**
Reductio Ad Hitlerum	All sentences	99.20%	85.23%	98.12%	61.93%	98.76%	67.43%
	Span sentences	99.83%	97.76%	97.85%	72.98%	98.32%	**74.13%**
Following Behind	All sentences	97.44%	52.20%	99.16%	51.92%	98.67%	51.89%
	Span sentences	95.73%	52.89%	96.89%	58.86%	96.93%	**71.32%**

6.2. *Techniques Classification*

Table 4 reports the results for the sentence-level propaganda technique classification task. It should be noted that in the span-sentence setting, the sentences are selected based on gold spans obtained directly from the annotations.

Overall, the results show that for most techniques using only propaganda span sentences improves classification performance. The only exception is the *emotional expression* technique, for which better results are achieved when training on all sentences.

One possible explanation relates to the distribution of this technique in the dataset. Emotional expression accounts for a large proportion of span sentences, covering approximately 56% of all annotated spans (Fig. 3). However, this becomes more complex when considering the overlap between techniques (Fig. 2), as emotional expression frequently co-occurs with other propaganda techniques and often appears in sentences containing multiple rhetorical patterns.

When training only on span sentences, non-propagandistic sentences are removed, reducing the number of negative examples that help distinguish emotional expression from other techniques. As a result, the model may learn less distinctive features for emotional expression and struggle to differentiate them from similar rhetorical patterns, such as simplification.

For all other techniques, a clear improvement can be observed when using span sentences instead of all sentences. This suggests that span identification helps the classifica-

tion model focus on propagandistic content and improves technique detection, particularly for less frequent techniques. A likely explanation is that removing non-propagandistic sentences effectively increases the proportion of sentences containing the target technique, which leads to a more balanced training distribution.

Interestingly, the highest performance is achieved for the techniques *Reductio Ad Hitlerum* (F1 = 74.13%) and *Following Behind* (F1 = 71.32%). This may be explained by the fact that these techniques often contain very distinctive linguistic cues. For example, the *Following Behind* (Bandwagon) technique is typically expressed through phrases that signal broad consensus, such as "everyone", "the majority", or "the whole nation". Similarly, *Reductio Ad Hitlerum* frequently appears in contexts referring to Nazism, fascism and etc. Such clearly identifiable patterns allow the model to learn more discriminative features.

A similar observation can be made for other relatively rare techniques such as *Waving the Flag* (69.85%) and *Appeal to Authority* (67.06%). These techniques also tend to appear in recognizable rhetorical contexts. For example, *Waving the Flag* often relies on patriotic language, while *Appeal to Authority* references influential figures or institutions, which makes these patterns easier for the model to detect.

For the most dominant techniques in the dataset, such as *Emotional Expression*, *Simplification*, and *Doubt*, the performance remains above 60% F1. However, the frequent overlap between these techniques may make it more difficult for the model to distinguish their boundaries. These techniques often appear together in the same sentences or in similar rhetorical contexts, which can complicate the learning of clearly separable features.

The lowest performance is observed for the *Uncertainty* and *Whataboutism/Red Herring/Strawman* techniques. This may be explained by the fact that these techniques are generally more difficult to identify, even for human annotators. In the HALT-PROP corpus (Rizgelienė *et al.*, 2025), these techniques were reported to have the lowest inter-annotator agreement scores.

6.3. *Final Evaluation*

Table 5 presents the overall results for propaganda technique detection. The table reports the performance of the LT-MLKM-modernBERT model fine-tuned under different experimental settings: using all sentences and using only sentences containing propaganda spans. *Gold spans* refer to results obtained using the span annotations from the corpus, while *predicted spans* refer to spans generated by the best-performing span identification model (see Table 2). In addition to the per-technique results, we report the average values of the evaluation metrics across all techniques.

For comparison, we also include results for a *multi-class multi-label* setup, where instead of fine-tuning LT-MLKM-modernBERT separately for each propaganda technique, the model was fine-tuned jointly for all techniques. In this setting, we again experimented with both all sentences and span-only sentences, and evaluated performance using both gold spans and predicted spans. We also report the results of the GPT-5.3 model obtained through prompting in both zero-shot and few-shot settings (Section 5.3.1).

Table 5
Comparison of LT-MLKM-modernBERT and GPT-5.3 performance across propaganda techniques under different sentence settings. Binary and multi-class classification results are reported.

Technique	Model	Setting	Binary		Multi-class	
			Acc.	F1	Acc.	F1
Emotional Expression	**LT-MLKM-modernBERT**	**All sentences**	70.84%	**66.08%**	51.05%	47.49%
		Span sentences (gold)	61.24%	60.62%	48.78%	48.67%
		Span sentences (predicted)	58.81%	57.82%	50.79%	48.69%
	GPT-5.3 Zero Shot	–	62.22%	52.64%	–	–
	GPT-5.3 Few Shot	–	60.08%	52.23%	–	–
Simplification	**LT-MLKM-modernBERT**	All sentences	78.23%	60.25%	55.15%	47.30%
		Span sentences (gold)	70.22%	**61.70%**	52.43%	50.21%
		Span sentences (predicted)	69.67%	**61.21%**	54.27%	47.31%
	GPT-5.3 Zero Shot	–	51.98%	43.46%	–	–
	GPT-5.3 Few Shot	–	84.89%	46.12%	–	–
Doubt	**LT-MLKM-modernBERT**	All sentences	76.18%	59.90%	50.28%	40.61%
		Span sentences (gold)	74.22%	**64.97%**	52.89%	47.02%
		Span sentences (predicted)	73.57%	**60.59%**	48.77%	40.21%
	GPT-5.3 Zero Shot	–	74.53%	52.75%	–	–
	GPT-5.3 Few Shot	–	86.56%	52.45%	–	–
Whataboutism/ Red Herring/ Strawman	**LT-MLKM-modernBERT**	All sentences	87.75%	49.36%	54.31%	38.18%
		Span sentences (gold)	90.21%	**51.74%**	51.11%	38.79%
		Span sentences (predicted)	93.60%	**51.64%**	51.63%	37.11%
	GPT-5.3 Zero Shot	–	71.18%	45.16%	–	–
	GPT-5.3 Few Shot	–	89.60%	51.10%	–	–
Appeal to Authority	**LT-MLKM-modernBERT**	All sentences	77.08%	48.85%	46.71%	35.01%
		Span sentences (gold)	91.54%	**67.06%**	56.73%	43.41%
		Span sentences (predicted)	92.19%	**54.97%**	52.59%	38.38%
	GPT-5.3 Zero Shot	–	79.40%	50.36%	–	–
	GPT-5.3 Few Shot	–	94.45%	52.67%	–	–
Waving the Flag	**LT-MLKM-modernBERT**	All sentences	96.31%	68.85%	47.18%	33.71%
		Span sentences (gold)	94.67%	**69.85%**	57.53%	42.66%
		Span sentences (predicted)	96.61%	**71.35%**	60.15%	41.59%
	GPT-5.3 Zero Shot	–	97.05%	52.22%	–	–
	GPT-5.3 Few Shot	–	97.52%	53.92%	–	–
Uncertainty	**LT-MLKM-modernBERT**	All sentences	97.33%	50.46%	53.81%	37.39%
		Span sentences (gold)	91.02%	**57.47%**	54.87%	39.26%
		Span sentences (predicted)	90.21%	51.64%	55.07%	37.81%
	GPT-5.3 Zero Shot	–	95.81%	51.73%	–	–
	GPT-5.3 Few Shot	–	95.07%	52.78%	–	–
Reductio Ad Hitlerum	LT-MLKM-modernBERT	All sentences	98.76%	67.43%	49.47%	34.26%
		Span sentences (gold)	98.32%	74.13%	52.37%	36.45%
		Span sentences (predicted)	98.64%	74.66%	56.93%	37.65%
	GPT-5.3 Zero Shot	–	98.86%	**77.42%**	–	–
	GPT-5.3 Few Shot	–	98.44%	66.71%	–	–
Following Behind	**LT-MLKM-modernBERT**	All sentences	98.67%	51.89%	66.35%	41.17%
		Span sentences (gold)	96.93%	**71.32%**	50.00%	35.92%
		Span sentences (predicted)	97.14%	**65.34%**	49.80%	34.76%
	GPT-5.3 Zero Shot	–	98.07%	59.77%	–	–
	GPT-5.3 Few Shot	–	94.78%	49.50%	–	–
Average across techniques	**LT-MLKM-modernBERT**	All sentences	86.79%	58.12%	52.70%	39.46%
		Span sentences (gold)	85.37%	**64.32%**	52.97%	42.49%
		Span sentences (predicted)	85.60%	**61.02%**	53.33%	40.39%
	GPT-5.3 Zero Shot	–	81.01%	53.95%	–	–
	GPT-5.3 Few Shot	–	89.04%	53.05%	–	–

Overall, the results indicate that the selected *binary per-technique training approach*, in which a separate classifier is fine-tuned for each propaganda technique, achieves substantially better detection performance than the multi-class multi-label model across all techniques. This outcome can be explained by the high degree of overlap between propaganda techniques in the corpus. In a joint multi-label setting, the model must learn several overlapping technique patterns simultaneously, which may make it more difficult to capture features that are specific to individual techniques. In contrast, training separate binary classifiers allows each model to focus on detecting one technique at a time.

When comparing the fine-tuned models trained using span sentences and those trained using all sentences, span-based training performs better for all techniques except *emotional expression*. This observation is consistent with earlier findings discussed in Section 6.2. Furthermore, when predicted spans were used instead of gold spans, the results remained comparable and still outperformed the all-sentences setting. These findings suggest that span identification indeed improves propaganda technique detection.

However, high performance degradation is observed for some techniques when predicted spans are used instead of gold spans. For example, the macro-F1 score for *appeal to authority* decreases from 67.06% to 54.97%, for *uncertainty* from 57.47% to 51.64%, and for *following behind* from 71.32% to 65.34%. This can be explained by the strong class imbalance present in these techniques, each representing less than 4% of all span sentences. If the predicted span model fails to identify even a small number of sentences that contain these techniques, their proportion in the dataset decreases further, significantly affecting overall performance.

Comparing the fine-tuned transformer models with GPT results, the fine-tuned model outperforms GPT-5.3 for all techniques except *reductio ad hitlerum*. This exception can be explained by the fact that this technique has a highly distinctive contextual pattern and is generally easier to identify, as it involves references to Nazism or fascism. Consequently, providing GPT with a clear definition of the technique appears sufficient for accurate detection. In the zero-shot setting, GPT achieved a macro-F1 score of 77.42%, outperforming the fine-tuned transformer model. Interestingly, this high performance was achieved only in the zero-shot setting, while the few-shot configuration produced the lowest results for this technique. This may suggest that the additional examples unintentionally biased the model.

Overall, the best-performing techniques were *waving the flag*, *reductio ad hitlerum*, and *following behind*. For *waving the flag*, the span-based model achieved 69.85% (gold spans) and 71.35% (predicted spans). For *reductio ad hitlerum*, GPT-5.3 zero-shot achieved 77.42%, while the span-based models achieved 74.13% (gold) and 74.66% (predicted). For *following behind*, the span-based model achieved 71.32% (gold) and 65.34% (predicted). As previously noted, these techniques have clear contextual patterns and are among the easiest to recognize. They also achieved the highest annotation agreement rates in the HALT-PROP corpus (Rizgelienė *et al.*, 2025), indicating that they are relatively easy for human annotators to identify as well.

The averaged results across all techniques show the same general trend: the binary per-technique approach outperforms the multi-class multi-label model. In addition, the

transformer model fine-tuned on span-only sentences outperforms the model fine-tuned on all sentences, both when using gold spans and predicted spans. The span-based transformer models also outperform GPT-5.3 in both zero-shot and few-shot settings.

7. Conclusion

In this study, we present the first approach for propaganda technique detection in Lithuanian. In addition, we investigate several research questions that are not only specific to Lithuanian but are also applicable to other languages, particularly low-resource settings. These questions are designed to inform the development of propaganda techniques detection models in similar settings, focusing on: (i) whether span boundary information and longer input sequences improve span identification, (ii) whether incorporating span-level information improves technique classification, and (iii) whether modelling techniques as separate binary classifiers outperform a multi-class approach in scenarios with high overlap between techniques.

Using the monolingual Lithuanian transformer *LT-MLKM-modernBERT*, we experiment with different input lengths, tagging schemes, and classification setups (binary vs. multi-class, with and without span-level information). Overall, the results show that:

- Incorporating span boundary information improves span identification performance. The BILOU tagging scheme consistently outperforms binary tagging (71.95% vs 64.55%), producing more accurate and consistent spans and capturing a larger proportion of gold span characters.
- Incorporating span-level information improves techniques classification for most techniques, particularly those with class imbalance. Training only on sentences within propaganda spans yields better performance, and this effect persists even when using predicted spans instead of gold annotations.
- In scenarios with high overlap between techniques, modelling each technique as a separate binary classifier outperforms the multi-class approach, highlighting the benefit of learning technique-specific features.
- Fine-tuned transformer models outperform GPT-5.3 in both zero-shot and few-shot settings for most techniques. The only exception is reductio ad Hitlerum, where GPT-5.3 achieves the best zero-shot performance. This result is likely driven by the distinctive context and clear definition of this technique, which also makes it one of the easiest for humans to recognize. Additionally, the results indicate that fine-tuned models also perform better on techniques with clearly separable contexts, despite class imbalance. In contrast, techniques that are more frequent in the corpus but lack a clearly distinctive context tend to achieve lower performance, even when the training data is relatively balanced.

7.1. *Proposed Approach*

Since the results show that span identification improves propaganda technique classification, we propose a two-stage approach for propaganda technique detection in Lithuanian.

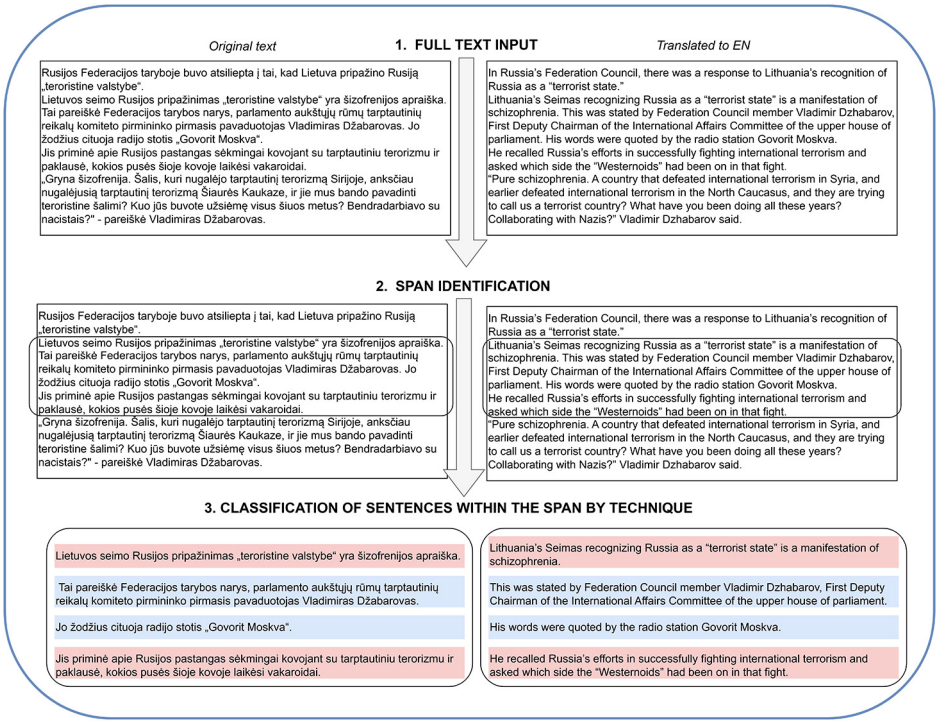

Fig. 9. Illustration of the proposed two-stage approach: full-text input, propaganda span identification, and technique classification of sentences containing the identified spans. English translations are provided alongside the original Lithuanian examples.

Figure 9 illustrates the proposed method. First, the full text is given as input to the span identification model, which detects propaganda spans. Then, only the identified spans are further analysed, and the corresponding sentences are classified according to propaganda technique. In this example, the pink colour indicates the *emotional expression* technique, while the blue colour indicates the *appeal to authority* technique.

8. Limitations and Future Work

Although this study presents the first approach for propaganda technique detection in Lithuanian, several limitations remain. These limitations also point to important directions for future work.

- ***Model comparison under fixed hyperparameters.*** This work mainly focused on LT-MLKM-modernBERT, one of the latest monolingual Lithuanian transformer models. However, for the span identification task, we also compared its performance with two multilingual transformer models: XLM-RoBERTa and LitLatBERT. Although all three models are transformer encoders, they differ in architecture. XLM-RoBERTa and LitLatBERT are RoBERTa-based models, whereas LT-MLKM-modernBERT is based on

the ModernBERT architecture. To ensure comparability, the same hyperparameters were used across all models and tasks. We also did not investigate layer-freezing strategies; instead, all model layers were fine-tuned in each experiment. Future work should include model-specific hyperparameter optimization and different fine-tuning strategies, such as freezing selected layers, to provide a more comprehensive comparison of transformer models for Lithuanian propaganda detection, since different hyperparameter choices may affect classification behaviour (Perišić *et al.*, 2025).

- ***Single-run evaluation.*** Another limitation of this study is that each experimental configuration was evaluated using a single fine-tuning run. Since transformer fine-tuning can be affected by stochastic factors, such as random initialization of classification heads, performance may vary across runs with different random seeds. Future work should evaluate each configuration over multiple runs and report averaged results together with standard deviations or confidence intervals.
- ***Technique overlap.*** The technique overlap analysis conducted in this study shows that techniques frequently co-occur, suggesting that some of them may share similar linguistic patterns. In the current modelling approach, techniques are treated as independent labels, which may overlook important relationships between them. When several techniques occur within the same sentence or span, modelling them as fully independent categories may limit the model's ability to capture these dependencies. A more fine-grained linguistic analysis could examine whether different co-occurrence patterns are associated with different ways in which techniques are expressed in text. For example, *emotional expression* may be expressed differently when it co-occurs with *simplification* than when it co-occurs with *doubt*. Future research could therefore explore modelling approaches that explicitly capture dependencies between techniques. This may improve the detection of techniques that frequently overlap.
- ***LLM-based approaches.*** Another limitation of this study is that the LLM-based comparison is limited to one model. Future research could extend this analysis by evaluating multiple LLMs, including multilingual and open-weight models. In addition, fine-tuning such models on the annotated dataset could be explored, as they may better capture broader contextual information relevant to propaganda technique detection.

A. English Translation of the Example in Fig. 4

"The United Kingdom, Germany, Italy, Canada, the United States, France, and Japan recognized the success of the special military operation of the Russian armed forces," he emphasized.
According to the former Prime Minister of the State Duma, **the statement adopted by the leaders of the G7 countries about the intention "not to allow Russia to win in Ukraine" can only be assessed in this way.** Vyacheslav Volodin added that **in this way, the G7 countries oppose the demilitarization and denazification of Ukraine, the status of a non-aligned and non-nuclear Ukraine, as well as the right to self-determination of the citizens of the Donetsk and Luhansk People's Republics.**
"Now it is clear why there was no response to our President's proposed initiatives on security issues. Russia thwarted their plans to expand NATO further east," he concluded.
"Everyone understands that Ukraine must be liberated from Nazism and Banderites. Everyone understands that the decision of the President of the Russian Federation to begin the operation for the demilitarization and denazification of Ukraine is historic. However, not everyone understands that if we stop and the Nazis remain in power, the situation will become even worse."
I would like to remind you that today we are not ending the Great Patriotic War. **At that time, we failed to stop the Banderites;** during the ten post-war years, attempts were made to eliminate them; 57 thousand Soviet soldiers, military personnel, state security officers, party and Komsomol Soviet activists were killed," said Gennady Zyuganov, the head of the Communist Party faction, at the plenary session.

B. English Translation of the Example in Fig. 6

While Ukrainians heroically defend their homeland and us, while the entire Lithuanian nation is fighting in the information war, leftist traitors strike Lithuania in the back with the Partnership Law that destroys the core of the nation – the family.
As if that were not enough, **they equate defenders of the natural family with the Kremlin regime and personally with Vladimir Putin.**
Apparently, he opposes genderism, so everyone who opposes genderism sits at the same table as the Moscow aggressor.
Reality is completely the opposite.
There is no partnership law in Ukraine.
Nor is there such a law in the countries of our strategic partners – Poland, Latvia, and Georgia.
Ukrainians are fighting not for genderism or globalism, but for their national state, their families, and their freedom.

Funding

This research was supported by the Lithuanian Government Priority Research Program "Building Societal Resilience and Crisis Management in the Context of Con temporary 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."

References

Alam, F., Mubarak, H., Zaghouani, W., Da San Martino, G., Nakov, P. (2022). Overview of the WANLP 2022 shared task on propaganda detection in arabic. In: Bouamor, H., Al-Khalifa, H., Darwish, K., Rambow, O., Bougares, F., Abdelali, A., Tomeh, N., Khalifa, S., Zaghouani, W. (Eds.), *Proceedings of the Seventh Arabic Natural Language Processing Workshop (WANLP)*. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, pp. 108–118. https://doi.org/10.18653/v1/2022.wanlp-1.11.

Barrón-Cedeño, A., Jaradat, I., Da San Martino, G., Nakov, P. (2019). Proppy: organizing the news based on their propagandistic content. *Information Processing & Management*, 56(5), 1849–1864. https://doi.org/10.1016/j.ipm.2019.03.005. https://www.sciencedirect.com/science/article/pii/S0306457318306058.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. In: Jurafsky, D., Chai, J., Schluter, N., Tetreault, J. (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, pp. 8440–8451. https://doi.org/10.18653/v1/2020.acl-main.747.

Da San Martino, G., Yu, S., Barrón-Cedeño, A., Petrov, R., Nakov, P. (2019). Fine-grained analysis of propaganda in news articles. In: Inui, K., Jiang, J., Ng, V., Wan, X. (Eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, pp. 5636–5646. https://doi.org/10.18653/v1/D19-1565.

Da San Martino, G., Barrón-Cedeño, A., Wachsmuth, H., Petrov, R., Nakov, P. (2020). SemEval-2020 Task 11: detection of propaganda techniques in news articles. In: Herbelot, A., Zhu, X., Palmer, A., Schneider, N., May, J., Shutova, E. (Eds.), *Proceedings of the Fourteenth Workshop on Semantic Evaluation*. International Committee for Computational Linguistics, Barcelona, pp. 1377–1414. https://doi.org/10.18653/v1/2020.semeval-1.186.

Dimitrov, D., Bin Ali, B., Shaar, S., Alam, F., Silvestri, F., Firooz, H., Nakov, P., Da San Martino, G. (2021). SemEval-2021 Task 6: detection of persuasion techniques in texts and images. In: Palmer, A., Schneider, N., Schluter, N., Emerson, G., Herbelot, A., Zhu, X. (Eds.), *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*. Association for Computational Linguistics, pp. 70–98. https://doi.org/10.18653/v1/2021.semeval-1.7.

Dimitrov, D., Alam, F., Hasanain, M., Hasnat, A., Silvestri, F., Nakov, P., Da San Martino, G. (2024). SemEval-2024 Task 4: multilingual detection of persuasion techniques in memes. In: Ojha, A.K., Doğruöz, A.S., Tayyar Madabushi, H., Da San Martino, G., Rosenthal, S., Rosá, A. (Eds.), *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*. Association for Computational Linguistics, Mexico City, Mexico, pp. 2009–2026. https://doi.org/10.18653/v1/2024.semeval-1.275.

Hasanain, M., Hasan, M.A., Ahmad, F., Suwaileh, R., Biswas, M.R., Zaghouani, W., Alam, F. (2024). ArAIEval shared task: propagandistic techniques detection in unimodal and multimodal arabic content. In: Habash, N., Bouamor, H., Eskander, R., Tomeh, N., Abu Farha, I., Abdelali, A., Touileb, S., Hamed, I., Onaizan, Y., Alhafni, B., Antoun, W., Khalifa, S., Haddad, H., Zitouni, I., AlKhamissi, B., Almatham, R., Mrini, K. (Eds.), *Proceedings of the Second Arabic Natural Language Processing Conference*. Association for Computational Linguistics, Bangkok, Thailand, pp. 456–466. https://doi.org/10.18653/v1/2024.arabicnlp-1.44.

Horák, A., Sabol, R., Herman, O., Baisa, V. (2024). Recognition of propaganda techniques in newspaper texts: fusion of content and style analysis. *Expert Systems with Applications*, 251, 124085. https://doi.org/10.1016/j.eswa.2024.124085. https://www.sciencedirect.com/science/article/pii/S0957417424009515.

Jose, J., Geeng, C., Morales, K.O., McCoy, D., Greenstadt, R. (2025). What's in a label? Propaganda labels and user sharing behavior on social media platforms. *Proceedings of the International AAAI Conference on Web and Social Media*, 19(1), 918–934. https://doi.org/10.1609/icwsm.v19i1.35853.

Moral, P., Marco, G., Gonzalo, J., Carrillo-de-Albornoz, J., Gonzalo-Verdugo, I. (2023). Overview of DIPROMATS 2023: automatic detection and characterization of propaganda techniques in messages from diplomats and authorities of world powers. *Procesamiento del Lenguaje Natural*, 71, 397–407. http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6569.

Moral, P., Fraile, J.M., Marco, G., Peñas, A., Gonzalo, J. (2024). Overview of DIPROMATS 2024: detection, characterization and tracking of propaganda in messages from diplomats and authorities of world powers. *Procesamiento del Lenguaje Natural*, 73, 347–358.

Perišić, A., Vanbelle, S., Petričević, R.B. (2025). Quantifying binary classifier algorithms similarity with a consensus agreement approach. *Informatica*, 36(3), 657–676. https://doi.org/10.15388/25-INFOR601.

Piskorski, J., Stefanovitch, N., Da San Martino, G., Nakov, P. (2023). SemEval-2023 Task 3: detecting the category, the framing, and the persuasion techniques in online news in a multi-lingual setup. In: Ojha, A.K., Doğruöz, A.S., Da San Martino, G., Tayyar Madabushi, H., Kumar, R., Sartori, E. (Eds.), *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*. Association for Computational Linguistics, Toronto, Canada, pp. 2343–2361. https://doi.org/10.18653/v1/2023.semeval-1.317.

Rashkin, H., Choi, E., Jang, J.Y., Volkova, S., Choi, Y. (2017). Truth of varying shades: analyzing language in fake news and political fact-checking. In: Palmer, M., Hwa, R., Riedel, S. (Eds.), *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Copenhagen, Denmark, pp. 2931–2937. https://doi.org/10.18653/v1/D17-1317.

Ratinov, L., Roth, D. (2009). Design Challenges and Misconceptions in Named Entity Recognition. In: *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL-2009)*. Association for Computational Linguistics, Boulder, Colorado, pp. 147–155. https://aclanthology.org/W09-1119/.

Rizgelienė, I., Zubaitienė, V., Maliukevičius, N., Marcinkevičius, V. (2025). HALT-PROP: Human-Annotated Lithuanian Textual Corpus for Propaganda Narratives and Techniques. *Scientific Data*, 13(1), 47. https://doi.org/10.1038/s41597-025-06367-w.

State Digital Solutions Agency (SDSA) (2025). LT-MLKM-modernBERT: Lithuanian ModernBERT Language Model. https://huggingface.co/VSSA-SDSA/LT-MLKM-modernBERT. Developed by Vytautas Magnus University (VMU), UAB Neurotechnology, UAB Tilde informacinės technologijos, MB Krilas.

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

I. Rizgelienė is a PhD student at the Institute of Data Science and Digital Technologies, Vilnius University. Her primary research interests include propaganda detection and analysis, with an emphasis on low-resource languages.

P. Zaranka received his master's degree in computer modelling from Vilnius University in 2025 and is currently a lecturer in NLP at Vilnius University. His primary research interests include large language models, natural language processing, and agent-based modelling.

G. Korvel received the PhD degree from the Institute of Data Science and Digital Technologies, Vilnius University, in 2013. Currently, she is a Research Fellow and a Professor with the Institute of Data Science and Digital Technologies, Vilnius University. Her research interests include speech and music signal processing, natural language processing, the development of mathematical models, and the applications of computational intelligence.

V. Marcinkevičius is a senior researcher, head of the Intelligent Technologies Research Group, and head of Artificial Intelligence Laboratory at Vilnius University, Institute of Data Science and Digital Technologies. His research interests include machine learning, information security and natural language processing. His academic background includes MSc in mathematics from Vilnius Educational University an PhD in informatics from Vytautas Magnus University.