Bridging the Gap in Multilingual News Semantic Similarity: a Crowdsourced Benchmark for Ukrainian, Polish, Russian, and English

Evgeniya Sukhodolskaya[1]*, **Daryna Dementieva**[1,2], and **Alexander Fraser**[1,2,3]
[1]Technical University of Munich (TUM),
[2]Munich Center for Machine Learning (MCML), [3]Munich Data Science Institute
suxodolskaya97@gmail.com, daryna.dementieva@tum.de

Abstract

In the era of social networks and the rapid spread of misinformation, news analysis remains a critical task. Detecting fake news across multiple languages, particularly beyond English, poses significant challenges. Cross-lingual news comparison offers a promising approach to verify information by leveraging external sources in different languages (Chen and Shu, 2024). However, existing datasets for cross-lingual news analysis (Chen et al., 2022a) were manually curated by journalists and experts, limiting their scalability and adaptability to new languages. In this work, we address this gap by introducing a scalable, explainable **crowdsourcing pipeline for cross-lingual news similarity** assessment. Using this pipeline, we collected a novel dataset CROSSNEWS-UA of news pairs in **Ukrainian** as a central language with linguistically and contextually relevant languages—**Polish**, **Russian**, and **English**. Each news pair is annotated for semantic similarity with detailed justifications based on the **4W** criteria (Who, What, Where, When). We further tested a range of models, from traditional bag-of-words and Transformer-based architectures to large language models (LLMs). Our results highlight the challenges in multilingual and cross-lingual news analysis and offer insights into model performance.

1 Introduction

The detection of fake news, especially in a multilingual setup, is still a critical challenge. While numerous methods exist for fake news detection—ranging from internal linguistic features (Pérez-Rosas et al., 2018; Abualigah et al., 2024) to leveraging external knowledge (Ghanem et al., 2018; Chen et al., 2024a; Chen and Shu, 2024)—external verification from multilingual resources presents a promising direction (Dementieva et al., 2023).

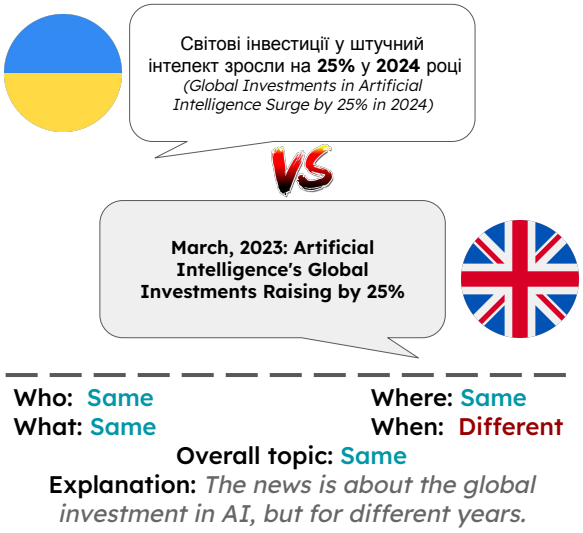


Figure 1: CROSSNEWS-UA: Example of the explainable cross-lingual news comparison obtained in this work for Ukrainian and related languages.

A key subtask in such a pipeline is cross-lingual news comparison, assessing the semantic similarity between news articles in different languages. However, beyond simple similarity scoring, there is a need for explainable data and approaches that provide insights into what main aspects of the content are similar or divergent (e.g., facts, entities, or temporal details, see Figure 1).

Cross-lingual document comparison is generally a complex task, and resources are scarce for many languages. A dataset for cross-lingual news comparison was introduced for SemEval-2022 Shared Task 8 (Chen et al., 2022a). The data collection was done by journalist experts but without clear criteria for a regular news reader and was difficult to reproduce. While the dataset indeed covered up to 10 languages, it is quite challenging to re-use the proposed expert-dependent pipeline to new languages easily. Given the current global events and the urgent need for reliable information, there is a strong motivation to support languages such as

*This work was conducted as part of Sukhodolskaya's Master's thesis at the Technical University of Munich.

Ukrainian, which are underrepresented in existing resources.

In this paper, we address these gaps with the following contributions:

- We introduce a novel **crowdsourcing pipeline** designed to collect high-quality cross-lingual news similarity data, ensuring both scalability and transparency in the annotation process. We designed the pipeline to not just give an overall similarity score, but **explain** news relations across several dimensions.
- We apply this pipeline to create a **multilingual dataset featuring Ukrainian** alongside **cross-lingual pairs** with contextually relevant languages—Polish, Russian and English.
- Using this dataset, we **test a range of models**, from traditional bag-of-words approaches to Transformer-based ones and LLMs, giving insights into model performance on the multilingual news similarity estimation task.

All data, code, and annotation instructions are available online.[1]

2 Related Work

Multilingual Documents Similarities Datasets The challenge of estimating text similarity—ranging from sentences to paragraphs and full articles—has been widely studied in the research community. For example, Ferrero et al. (2016) introduced a multilingual, multi-style, and multi-granularity dataset for cross-lingual textual similarity detection, covering French, English, and Spanish, with data sourced from both manual and machine translations. In SemEval 2017, the Semantic Textual Similarity Multilingual and Cross-lingual Focused Evaluation task was introduced (Cer et al., 2017), concentrating on sentence-level similarity across five languages. More recently, the SemEval-2022 shared task on multilingual news similarity (Chen et al., 2022a) expanded language coverage and evaluated documents based on seven distinct parameters, with datasets curated by domain experts. In recent work, Chen et al. (2024b) expanded the existing framework by incorporating automated news article scraping, further enriching the dataset. Despite these advancements, significant limitations remain: (i) none of the previous studies have introduced a scalable crowdsourcing pipeline adaptable to any language, and (ii) Ukrainian has consistently been excluded from the language coverage.

Multilingual Documents Similarities Estimation Methods Baseline methods for multilingual text similarity estimation often use sentence encoders like LaBSE (Reimers and Gurevych, 2020) and mBERT (Devlin et al., 2019), with similarity measured through metrics like cosine similarity. Another approach involves annotating topics from monolingual documents with cross-lingual labels, allowing similarity assessments through multilingual concept hierarchies derived from independently trained models (Badenes-Olmedo et al., 2019). Additionally, multilingual sequence-to-sequence training with a similarity loss function has been employed to improve cross-lingual document classification, incorporating similarity constraints during training to better identify semantically similar documents across languages (Yu et al., 2018). For SemEval-2022 Shared Task 8, the top solutions were based on fine-tuning, combining, and stacking various transformer models like DistilBERT, BERT, RoBERTa, and XLM-RoBERTa (Di Giovanni et al., 2022; Kuimov et al., 2022; Chen et al., 2022b). LLMs are usually evaluated on the MTEB benchmark (Muennighoff et al., 2023) for the STS task, however, this assessment does not cover document-to-document similarity (Zhao et al., 2023). Gatto et al. (2023) found that generative LLMs outperform encoder-based STS models in assessing complex semantic relationships, but their performance in low-resource languages or domain-specific tasks is still uncertain due to limited annotated data.

3 Multilingual News Pairs Corpus Annotation

We built our pipeline on the experience of the SemEval2022 Task 8 (Chen et al., 2022a), adapting it to a broader, more scalable scenario. Rather than relying on academic experts for labelling, we collected data through general crowdsourcing.

3.1 News Semantic Similarity Measurement Objective

While crowdsourcing offers scalability, it also presents challenges in ensuring labelling quality

[1]We release code for crowdsourcing project as well as the final dataset:
Code and details: https://github.com/TUM-NLP/crossnews-ua
CROSSNEWS-UA: https://huggingface.co/datasets/ukr-detect/crossnews-ua

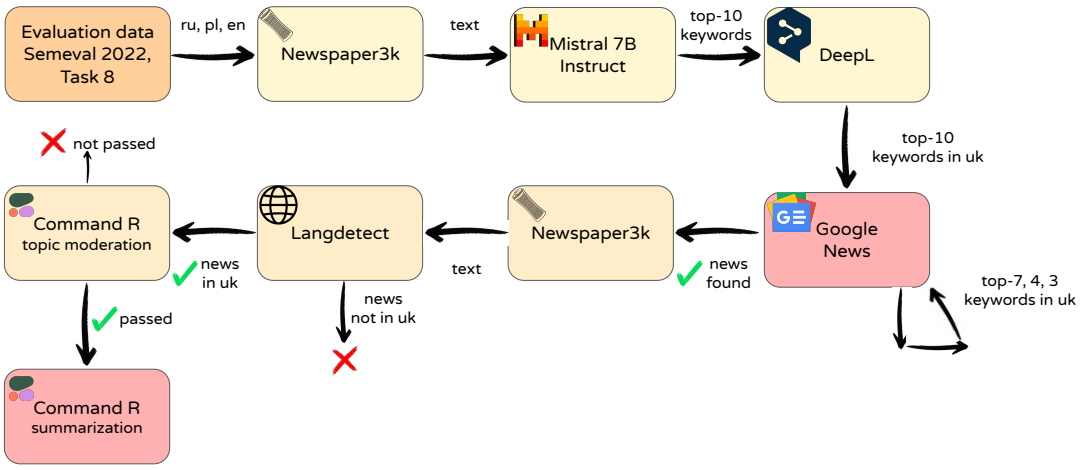


Figure 2: Initial scraping pipeline of news articles in Ukrainian.

and preventing fraud. Clear instructions, an intuitive interface, and a well-structured task design are essential for maintaining data reliability. To simplify the evaluation process, we decomposed the news similarity task using the **4W model**, derived from the traditional *5W1H* framework commonly used in journalism to summarise key story elements—What, Where, When, Who, Which, and How (Bleyer, 1913). Since comparing news articles aligns closely with evaluating these core elements, we focused on the **four most critical dimensions**: *Who*, *When*, *Where*, and *What*. Annotators assessed each of these dimensions using **three straightforward labels**—"Similar", "Somewhat related", and "Different"—with an optional "Incomparable" label ensuring both clarity and ease of use.

To improve the transparency of labelling decisions, annotators were also required to provide brief **explanations** justifying their similarity assessments, particularly regarding the "What" component. These explanations not only clarify annotation choices but also serve as valuable data for fine-tuning NLP models in tasks related to event summarisation and semantic similarity.

3.2 Selecting Multilingual News Articles for Annotation

The first step was to collect new data pairs that will include Ukrainian news. We developed a several-step pipeline that is schematically illustrated in Figure 2. For human evaluations of models' performance on Ukrainian text processing, we engaged three fluent Ukrainian speakers, each holding a Master's or PhD in Computer Science.

3.2.1 Initial News Selection

To ensure a balanced dataset across similarity levels and languages, we re-used the SemEval 2022 Task data as a base for the further search. We sampled equal numbers of news pairs from four original categories in this dataset: "Very Similar", "Somewhat Similar", "Somewhat Dissimilar", and "Very Dissimilar". Our focus was on only **related to Ukrainian languages**—Russian, Polish, and English news pairs—given their frequent coverage of events related to Ukraine. The primary objective was to identify a corresponding Ukrainian article for any language in each news pair.

3.2.2 Scraping News in Ukrainian

Generating Search Request in Ukrainian To find suitable Ukrainian news for pairs, we needed a concise summary of original articles' key events. Given that data samples' headlines often can be abstract and fail to capture the full content, we extracted the top 10 keywords from each article's content to guide our search. For keyword extraction, we employed the Mistral 7B Instruct model[2] which demonstrated strong performance in generating relevant keywords in the preliminary experiments (prompt in Appendix C.1). Then, the keywords were translated into Ukrainian using DeepL[3] and used as search queries.

Automated News Retrieval We sourced news articles using the `GoogleNews`[4] Python library. Our initial search queries consisted of the top 10 extracted keywords, joined by spaces. If no relevant

[2] https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2
[3] https://www.deepl.com
[4] https://pypi.org/project/GoogleNews

articles were retrieved, we incrementally simplified the query by reducing the keywords to the top 7, then 4, and finally the top 3, to broaden the search scope. The articles were scraped with the `Newspaper3k` library.[5] To ensure language accuracy, we used the langdetect[6] library to verify that the articles were indeed in Ukrainian.

3.3 Annotation Procedure

We implemented the annotation procedure using the Toloka platform,[7] chosen for its broad pool of speakers in our target languages and its robust quality control features. The following sections detail the entire annotation setup, including data preparation, instruction design, annotator selection, and the quality control measures employed.

To maintain high annotation quality in large-scale crowdsourcing projects, we considered two strategies: (i) providing **guiding questions** to direct annotators' attention to key sections of lengthy articles or generating concise, (ii) **information-dense summaries**, allowing annotators to quickly grasp the content and selectively verify details in the full text.

3.3.1 Preparing the Dataset for Annotation

Filtering Sensitive Topics To comply with the Toloka platform's guidelines on restricted content to protect annotators' well-being,[8] we pre-filtered news article pairs containing sensitive material. For automated filtering, we employed Cohere's Command-R[9] model, which demonstrated superior performance in detecting sensitive content compared to other open-source models in preliminary experiments.

Preparing Summarisation Format Summarisation was employed to assist annotators by structuring content in alignment with the similarity measurement criteria. The summaries emphasised key elements: *What* occurred, *When* and *Where* events took place, and *Who* the main actors were. For this task, we utilized again Command-R model as well. The English summarisation prompt example is provided in the Appendix C.2, with translated versions used for other languages. As Command-R's API supports Retrieval Augmented Generation (RAG),[10] we implemented two summarisation strategies: embedding the full news article directly within the prompt versus supplying it as retrieved content in RAG mode. Upon evaluating both summarisation approaches, we observed no consistent difference in overall quality, as both methods occasionally produced hallucinations. To mitigate this, we retained both summarisation versions and incorporated a feedback mechanism in the annotation interface, enabling annotators to flag cases where hallucinations rendered the task unfeasible. This approach ensured readiness for the subsequent crowdsourcing phase.

3.3.2 Instructions & Interface

To ensure annotators accurately assessed the information in the news pairs, we provided instructions detailing the task requirements, along with clear explanations of the dimensions and classification categories. We designed the interface to provide all needed information together with a comfortable quick look at the news. The instructions that annotators saw in the beginning:

> **Task description**
> In each task, you will compare two news articles and evaluate their similarity. The news articles in the pair may be in different languages, but this should not affect your comparison. For example, Wołodymyr Zełenski (Polish) and Volodymyr Zelenskyy (English) are the same actors in an event. For each news article, **we have provided**:
> • A **link** to the news website (be sure to follow it to see the original article, at least the headline and date of publication);
> • **Full text** of the article (if you want to check some details right in the task interface);
> • **Two summarisation** options made by specially trained algorithms (including basic information on four main parameters of events: WHO, WHEN, WHERE and WHAT happened).
> **Assessing news similarity**
> You will answer four questions to help us draw a conclusion about the similarity of the news.

The interface example is presented in Appendix D. The example of *Who* dimension question:

[5] https://newspaper.readthedocs.io
[6] https://pypi.org/project/langdetect
[7] https://toloka.ai
[8] https://toloka.ai/docs/guide/unwanted
[9] https://docs.cohere.com/docs/command-r
[10] https://docs.cohere.com/docs/retrieval-augmented-generation-rag

Do the main characters in the news match? (WHO)
The actor of an article can be a specific person (Volodymyr Zelenskyi), an organization/party or any other association of people.
• **Exactly** - the main and secondary actors in the articles are the same. At the same time, it is important to note that one actor can be called differently, for example, "Volodymyr Zelenskyi" and "President of Ukraine".
• **Partially** (there are overlaps) - Some of the main or secondary actors are the same.
• **No** - The protagonists in the articles do not coincide at all; they have nothing in common.
• **At least one article does not have any actors** - for example, one of the articles states something very vague, e.g. "the entire population of the Earth", or the article is a weather forecast.

The full detailed instruction text and the interface design are provided in Appendix E.

3.4 Annotators Selection

Language Proficiency To ensure high-quality annotations, we divided the labelling tasks into separate pools based on language combinations in the news article pairs: Ukrainian-Ukrainian, Ukrainian-Russian, Ukrainian-Polish, and Ukrainian-English. This allowed us to train and evaluate annotators according to their language proficiency. Toloka platform provided pre-filtering mechanisms to select annotators who had passed official language proficiency tests, serving as an initial screening step. In our scenario, we selected annotators who were proficient in both required languages. The language combinations did not affect pricing or quality control procedures, allowing for a consistent evaluation framework across pools.

Training and Exam Phases Annotators interested in participating first completed an unpaid training phase, where they reviewed detailed instructions and examples with explanations for correct labelling decisions (see Appendix E). Following this, annotators were required to pass a paid-by-us exam, identical in format to the actual labelling tasks, to demonstrate their understanding of the guidelines. Successful candidates gained access to the main assignments. Exams were manually reviewed by experts within Toloka's platform, using an interface that mirrored the labelling project setup, allowing reviewers to accept or reject submissions based on accuracy.

3.5 Quality Control

The quality control pipeline, illustrated in Figure 3, was designed with scalability in mind, leveraging automation wherever possible.

Annotators Overlap To ensure consistent and reliable annotations, each news article pair was assigned to **three annotators**, enabling the use of automatic *majority voting* for quality control in the classification tasks. This method worked effectively for structured responses to the Who, When, Where, and What questions, each constrained to a limited set of options. However, for the explanatory comments tied to the What question, which involve open-ended text, automatic verification through majority voting was not feasible due to the inherent variability in valid responses.

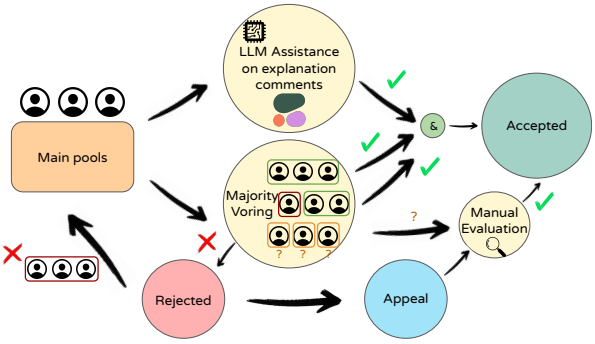
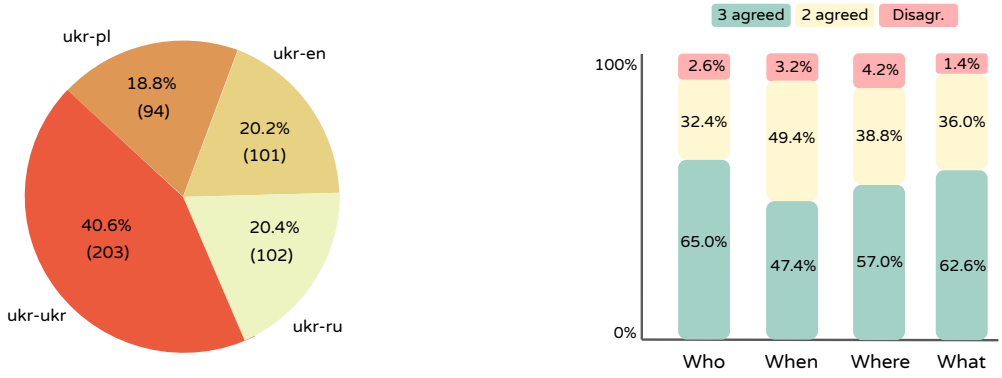


Figure 3: Quality control of labelled assignments overview.

LLM as an Annotation Quality Control Assistant To ensure scalable and high-quality annotation, we combined majority voting with the semantic reasoning of Command-R, an LLM used to automatically evaluate annotators' textual explanations and adherence to guidelines. Manual reviews were conducted only in cases of complete annotator disagreement or when rejections were appealed. We fine-tuned prompts by analysing common annotation errors, especially in event similarity categories like "Different" and "Somewhat Related", while manually reviewing inconsistencies in "Similar" pairs for Who, Where, and When questions. The final prompt is listed in Appendix F. Command-R demonstrated strong potential as a moderation tool, with a True Positive rate of 96% and a False Negative rate of 75%. Additionally, the low rate of appealed rejections (under 10%) indicated the effectiveness of our quality control process. Tasks rejected were reassigned to highly engaged annotators with strong performance records. This hybrid approach of automated checks and selective manual review successfully balanced quality assurance with scalability.



(a) Distribution of Language Pairs

(b) Labels Annotation Agreement per Dimension

Figure 4: Collected final dataset statistics.

3.5.1 Annotators Well-Being

We aimed to design a fair, transparent, and user-friendly crowdsourcing project, featuring an intuitive interface, comprehensive instructions, and a clear feedback mechanism that allowed annotators to appeal rejected tasks.

Fair Compensation Payment rates were set to balance grant funding constraints with fair wages, aligning with Ukraine's minimum hourly wage at the time of labelling (**1.12 USD/hour**). Annotators received **0.20 USD** for each accepted exam assignment and **0.26 USD** for assignments in the main pools (assignment includes only one news pair). Even for tasks rejected by the Majority Vote Quality Control but containing valid comments aligned with instructions, annotators were compensated **0.15 USD**. The complete project budget is detailed in Table 1.

Positive Project Ratings Toloka provided annotators with tools[11] to rate project fairness in terms of payment, task design, and organiser responsiveness. Our projects received high ratings: **4.98/5.00** for the Main Project and **4.95/5.00** for the Training Project.

3.6 Annotators Statistics

Out of the 43 annotators who submitted work in the training phase, 29 participated in the main pools, with an average of 7 active users per day. The median age of the annotators was 49 years, ranging from a minimum of 28 to a maximum of 74. The highest number of assignments submitted by a single person was **138**, while the median number of submissions per person was **52**. On average, completing a labelling task took **9.06** minutes.

Category of Spending	Per Assignment ($)	Total ($)
Exam assignments	0.20	24.08
Main pools, accepted assignments	0.20	250.04
Bonus, rejected assignments with comments relevant to instructions	0.15	129.95
Additional bonus, accepted assignments	0.06	79.40
Total		**464.16**

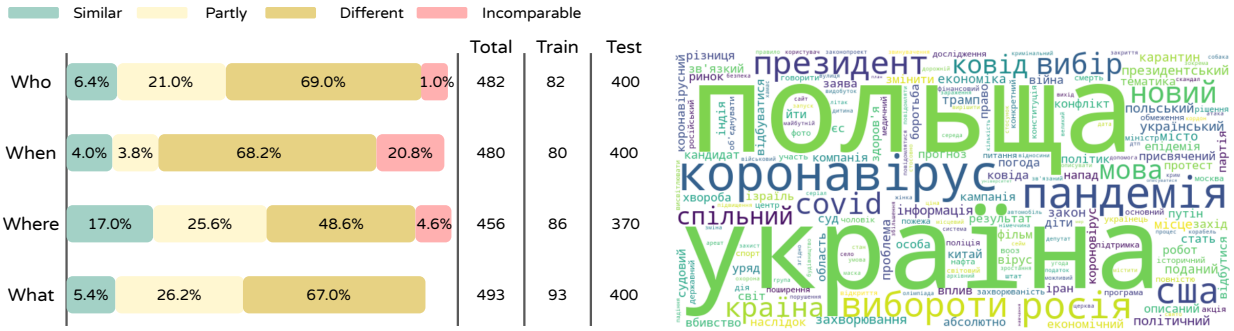
Table 1: Annotation Expenses.

4 CROSSNEWS-UA: Final Dataset

The resulting dataset contains **500** pairs of news articles, with the language distribution shown in Figure 4a with: `ukr-ukr` 203 pairs, `ukr-pl` 94 pairs, `ukr-en` 101 pairs, and `ukr-ru` 102 pairs. Additionally, in Figure 4b, we show the number of annotators who agreed on labels for each dimension. The high level of agreement indicates that, despite the complexity of the news semantic similarity task, our new setup effectively streamlined the process for reliable data collection through crowdsourcing.

To establish ground truth labels, we aggregated annotations from overlapping data by **major voting**, excluding instances of full disagreement within individual categories. The distribution of these aggregated labels across all dimensions is shown in Figure 5a. Notably, the dataset exhibits an imbalance, with the majority of samples labelled as *Different*. For the *When* dimension, a significant portion of samples is marked as *Incomparable*,

[11] https://toloka.ai/docs/guide/project_rating_stat



(a) Distribution of Labels per Dimension

(b) Popular Words in *What* Explanations

Figure 5: CROSSNEWS-UA: Labels details.

likely due to challenges in scraping timestamps from original sources and the absence of this information in many articles.

For the *What* dimension, the Ukrainian articles were initially scraped based on pre-existing topics in the dataset, resulting in thematic relationships across the samples without *Incomparable* labels. Figure 5b presents an overview of the most common topics mentioned in the *What* explanations, with prominent terms such as *Poland*, *Covid*, *Ukraine*, *elections*, *Russia*, *USA*, *president*, *economics*, *country*, and *borders*. These terms reflect discussions related to politics and economics in Ukraine and neighbouring regions.

Finally, we randomly split the dataset into train and test sets, ensuring that both sets contain all labels.

5 Baselines

We employed our curated dataset to benchmark multiple models on the Multilingual News Semantic Similarity task. We utilised two types of approaches—**embeddings-based** and **prompting-based**. The task was structured as a classification problem across four dimensions: WHO, WHERE, WHEN, and WHAT. However, the textual explanations provided for the WHAT dimension were excluded from this benchmarking, leaving room for future research focused on explanation generation.

Given the multi-class classification nature of the task, we assessed model performance using the standard classification metric **F1-Score**. To address the label imbalance, we report **macro-averaged F1-score**.

5.1 Embeddings

To the best of our knowledge, no existing models are specifically designed for multilingual semantic similarity measurement of news articles that include the Ukrainian language. Thus, we employ multilingual embedding models to generate contextualised embeddings of news article summaries, structured around the Who, Where, When, and What dimensions. The embedding of a text was calculated as the **mean of its token embeddings**. Semantic similarity between news article embeddings is evaluated using **cosine similarity**. We use a **decision tree classifier** to find the optimal class decision thresholds for each model and each question. The models evaluated include:

Bag-of-Words Due to its implementation design, the Bag-of-Words model cannot be directly applied for cross-lingual similarity measurement. Therefore, we translated the Russian, Polish, and English summarisation parts into Ukrainian using DeepL. Then, we transformed texts into vectors using `CountVectorizer`.

BERT We used the multilingual version of BERT[12] (Devlin et al., 2018) as it contains all target languages in the pre-trained data.

XLM-RoBERTa As an extension of BERT-alike models, we used XLM-Roberta-base[13] (Conneau et al., 2019) as it previously showed promising results in Ukrainian texts classification (Dementieva et al., 2025).

mT5 We also selected the encoder part of mT5-small model[14] (Xue et al., 2021) to get embeddings of our texts.

Multilingual E5-large Finally, we utilised the multilingual e5-large model[15] (Wang et al., 2024)

[12] https://huggingface.co/google-bert/bert-base-multilingual-cased
[13] https://huggingface.co/FacebookAI/xlm-roberta-base
[14] https://huggingface.co/google/mt5-small
[15] https://huggingface.co/intfloat/multilingual-e5-large

Category	Encoder-based Models					LLMs			
	BoW	mBERT	XLM-R	e5	mT5	Mistral7B	Mixtral7B	LLaMa3.1	Aya8B
Who	0.35	0.30	0.31	**0.52**	0.42	0.20	0.25	0.34	**0.42**
When	0.37	**0.49**	0.13	0.28	0.27	0.12	0.35	**0.39**	0.12
Where	0.41	0.31	0.35	**0.49**	0.28	0.26	0.34	**0.44**	0.26
What	0.50	0.17	0.15	**0.58**	0.37	0.35	**0.58**	0.51	0.35

Table 2: Macro-averaged F1-score of baselines tested on our proposed multilingual news semantic similarity dataset. The results in **bold** denote the best scores per dimension per model type.

that already showed promising results in various multilingual text similarity and extraction tasks.

5.2 LLMs

To test models based on another methodology, we also tried out various LLMs on our benchmark dataset, transforming our classification task into the `text-to-text` generation one. While Ukrainian and other target languages are not always explicitly present in the pre-training data reports, the emerging abilities of LLMs already showed promising results in handling new languages (Wei et al., 2022), including Ukrainian (Dementieva et al., 2025). With the prompt mentioned in Appendix G, we included the following models in evaluation:

Mistral We used several version of Mistral-family models (Jiang et al., 2023)—Mistral-7B-Instruct[16] and Mixtral-8x7B-Instruct.[17] The model cards do not mention explicitly Ukrainian and other languages; however, Mistral showed promising results in Ukrainian text classification tasks (Dementieva et al., 2025).

Llama3 Also, we tested the Llama-3.1-8B-Instruct model[18] (AI@Meta, 2024). The model card as well does not state Ukrainian explicitly; however, it encourages research in the usage of the model in various multilingual tasks.

Aya-23 8B Finally, we chose the Aya-23-8B[19] model (Aryabumi et al., 2024). All target languages were explicitly present in its pre-training data.

6 Results

The final multilingual results are presented in Table 2 with reports for each language pair in Appendix H.

The results reveal several key insights. First, **When** emerged as the most challenging category for all models. This is understandable, as temporal information can be derived from both the article's content and its publication date. However, since annotators were instructed not to rely on publication dates as the primary source, this information was excluded from model inputs. For prompt-based models, incorporating publication dates into the instructions might have improved performance.

Second, among encoder models, **e5-large** achieved the highest performance. This is consistent with its status as a more advanced model widely adopted in modern NLP tasks like Information Retrieval, indicating its strong potential as a base model for future applications in news semantic similarity measurement. Interestingly, a simple **Bag-of-Words model** outperformed several multilingual, semantically aware encoders still prevalent in the field. This suggests that domain- and language-agnostic approaches like Bag-of-Words remain effective, particularly in low-resource or complex domains.

Lastly, among decoder-only models, **LLaMa 3.1 8B Instruct** delivered the best performance, which was unexpected given that **Aya-23 8B** was designed with multilingual tasks in mind. Despite this, Llama 3.1 8B Instruct currently appears to be the most promising candidate for future fine-tuning in the context of news semantic similarity tasks.

7 Conclusion

In this work, we addressed the gap in multilingual news similarity datasets by introducing a novel crowdsourcing pipeline for explainable data collection across underrepresented languages. We provided a comprehensive overview of each stage of the pipeline, including data scraping and prepara-

[16] https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3
[17] https://huggingface.co/mistralai/Mixtral-8x7B-Instruct-v0.1
[18] https://huggingface.co/meta-llama/Meta-Llama-3.1-8B-Instruct
[19] https://huggingface.co/CohereForAI/aya-23-8B

tion, annotation platform setup, instruction and interface development, annotator selection, and quality control processes. To enhance scalability and efficiency, we incorporated automation where possible, utilising LLMs as quality control assistants, thereby demonstrating the pipeline's adaptability to any digitalised language. Through this approach, we collected CROSSNEWS-UA: benchmark of 500 news article pairs in various language combinations, including Ukrainian, Polish, Russian, and English. Based on the collected data, we tested a range of embedding-based and LLM-based baselines, offering valuable insights into the challenges of the news semantic similarity task for these languages. We believe our dataset will contribute significantly to advancing explainable news comparison and trustworthy information detection.

Limitations

Presence of Underrepresented Languages While our dataset significantly addresses the underrepresentation of Ukrainian in the news semantic similarity task, there remains substantial room for enhancement. We provide detailed budget calculations to guide future data collection efforts, which could include expanding the dataset to cover additional languages, particularly other Slavic languages, to improve cross-linguistic representation.

Fixed Timestamp of Collection Furthermore, the current dataset is based on news topics from 2022; incorporating more recent events would enhance its relevance and applicability. Nonetheless, we believe that the news comparison pipelines developed using our data are adaptable and can be effectively transferred to new domains.

Limited Dataset Size Our dataset comprises 500 news article pairs with a relatively balanced representation of languages, though label distribution across dimensions remains uneven. Despite this, some of the tested out-of-the-box models demonstrated promising performance. We hope this dataset will serve as a valuable benchmark for future model evaluations involving Ukrainian. Additionally, exploring various data augmentation techniques presents a promising direction for further research.

Annotation Platform We hope that the provided instructions and annotation details make it possible to run the annotation pipeline on any annotation platform or interface without being limited to Toloka.

Baselines In terms of model benchmarking, we employed straightforward baseline approaches, leaving opportunities for more advanced experiments. Future work could involve end-to-end fine-tuning of Transformer-based models or assembling more complex architectures to improve performance.

LLMs Choice Additionally, while we focused on open-source LLMs for prompting, proprietary models were not included in our evaluation, representing another avenue for exploration.

Multilingual and Cross-lingual Experiments In our experiments, we perform experiments with only a multilingual setup. However, more nuanced insights could be gained by developing and evaluating language technologies tailored to specific language pairs.

Explainability Generation Experiments Finally, further research could delve into explanation generation techniques aligning with our dataset to enhance the interpretability and explainability of model outputs.

Ethics Statement

Our pipeline and dataset were developed with two primary objectives in mind: first, to enhance document representation and comparison for underrepresented languages, particularly Ukrainian; and second, to support research on robust, explainable methods for trustworthy news detection, especially concerning events in Ukraine and neighbouring countries. We believe this benchmark will contribute to advancing the field of multilingual news analysis and foster research in more robust and explainable fake news detection in any language.

Throughout the data collection process, we adhered to Toloka's ethical guidelines, placing a strong emphasis on annotator well-being as described in Section 3.5.1. We ensured fair compensation, aligned with or exceeding the minimum wage in Ukraine, and maintained open communication with annotators to address any concerns or questions promptly. Regarding data sharing, we follow the precedent set by SemEval 2022 Task 8 (Chen et al., 2022a) by releasing source links to the news articles rather than the full scraped texts, ensuring compliance with copyright regulations.

Acknowledgments

We would like to express our gratitude to Toloka.ai platform for their data annotation research grant. The work was also co-funded by the European Union (ERC, EPICAL, 101141712). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them.

References

Laith Abualigah, Yazan Yehia Al-Ajlouni, Mohammad Sh. Daoud, Maryam Altalhi, and Hazem Migdady. 2024. Fake news detection using recurrent neural network based on bidirectional LSTM and glove. *Soc. Netw. Anal. Min.*, 14(1):40.

AI@Meta. 2024. Llama 3 Model Card. https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md. Accessed: 2024-12-14.

Viraat Aryabumi, John Dang, Dwarak Talupuru, Saurabh Dash, David Cairuz, Hangyu Lin, Bharat Venkitesh, Madeline Smith, Jon Ander Campos, Yi Chern Tan, Kelly Marchisio, Max Bartolo, Sebastian Ruder, Acyr Locatelli, Julia Kreutzer, Nick Frosst, Aidan N. Gomez, Phil Blunsom, Marzieh Fadaee, Ahmet Üstün, and Sara Hooker. 2024. Aya 23: Open weight releases to further multilingual progress. *CoRR*, abs/2405.15032.

Carlos Badenes-Olmedo, José Luis Redondo García, and Óscar Corcho. 2019. Scalable cross-lingual document similarity through language-specific concept hierarchies. In *Proceedings of the 10th International Conference on Knowledge Capture, K-CAP 2019, Marina Del Rey, CA, USA, November 19-21, 2019*, pages 147–153. ACM.

Willard Grosvenor Bleyer. 1913. *Newspaper Writing and Editing*. Houghton Mifflin Company.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada. Association for Computational Linguistics.

Canyu Chen and Kai Shu. 2024. Combating misinformation in the age of llms: Opportunities and challenges. *AI Mag.*, 45(3):354–368.

Jifan Chen, Grace Kim, Aniruddh Sriram, Greg Durrett, and Eunsol Choi. 2024a. Complex claim verification with evidence retrieved in the wild. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3569–3587, Mexico City, Mexico. Association for Computational Linguistics.

Xi Chen, Mattia Samory, Scott Hale, David Jurgens, and Przemyslaw A. Grabowicz. 2024b. A multilingual similarity dataset for news article frame. In *Proceedings of the Eighteenth International AAAI Conference on Web and Social Media, ICWSM 2024, Buffalo, New York, USA, June 3-6, 2024*, pages 1913–1923. AAAI Press.

Xi Chen, Ali Zeynali, Chico Camargo, Fabian Flöck, Devin Gaffney, Przemyslaw Grabowicz, Scott Hale, David Jurgens, and Mattia Samory. 2022a. SemEval-2022 task 8: Multilingual news article similarity. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1094–1106, Seattle, United States. Association for Computational Linguistics.

Zhongan Chen, Weiwei Chen, YunLong Sun, Hongqing Xu, Shuzhe Zhou, Bohan Chen, Chengjie Sun, and Yuanchao Liu. 2022b. ITNLP2022 at SemEval-2022 task 8: Pre-trained model with data augmentation and voting for multilingual news similarity. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1184–1189, Seattle, United States. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Daryna Dementieva, Valeriia Khylenko, and Georg Groh. 2025. Cross-lingual text classification transfer: The case of Ukrainian. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 1451–1464, Abu Dhabi, UAE. Association for Computational Linguistics.

Daryna Dementieva, Mikhail Kuimov, and Alexander Panchenko. 2023. Multiverse: Multilingual evidence for fake news detection. *J. Imaging*, 9(4):77.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Marco Di Giovanni, Thomas Tasca, and Marco Brambilla. 2022. DataScience-polimi at SemEval-2022 task 8: Stacking language models to predict news article similarity. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1229–1234, Seattle, United States. Association for Computational Linguistics.

Jérémy Ferrero, Frédéric Agnès, Laurent Besacier, and Didier Schwab. 2016. A multilingual, multi-style and multi-granularity dataset for cross-language textual similarity detection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4162–4169, Portorož, Slovenia. European Language Resources Association (ELRA).

Joseph Gatto, Omar Sharif, Parker Seegmiller, Philip Bohlman, and Sarah Preum. 2023. Text encoders lack knowledge: Leveraging generative LLMs for domain-specific semantic textual similarity. In *Proceedings of the Third Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 277–288, Singapore. Association for Computational Linguistics.

Bilal Ghanem, Manuel Montes-y-Gómez, Francisco M. Rangel Pardo, and Paolo Rosso. 2018. UPV-INAOE - check that: An approach based on external sources to detect claims credibility. In *Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum, Avignon, France, September 10-14, 2018*, volume 2125 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *CoRR*, abs/2310.06825.

Mikhail Kuimov, Daryna Dementieva, and Alexander Panchenko. 2022. SkoltechNLP at SemEval-2022 task 8: Multilingual news article similarity via exploration of news texts to vector representations. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1136–1144, Seattle, United States. Association for Computational Linguistics.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2023. MTEB: massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2023, Dubrovnik, Croatia, May 2-6, 2023*, pages 2006–2029. Association for Computational Linguistics.

Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. 2018. Automatic detection of fake news. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3391–3401, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525, Online. Association for Computational Linguistics.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual E5 text embeddings: A technical report. *CoRR*, abs/2402.05672.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent abilities of large language models. *Trans. Mach. Learn. Res.*, 2022.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 483–498. Association for Computational Linguistics.

Katherine Yu, Haoran Li, and Barlas Oguz. 2018. Multilingual seq2seq training with similarity loss for cross-lingual document classification. In *Proceedings of The Third Workshop on Representation Learning for NLP, Rep4NLP@ACL 2018, Melbourne, Australia, July 20, 2018*, pages 175–179. Association for Computational Linguistics.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023. A survey of large language models. *CoRR*, abs/2303.18223.

A Licensing of Resources

Below is an overview of the licenses associated with each resource used in this work (Table 3).

Resource	License	Homepage
CROSSNEWS-UA	CC BY 4.0, news released as links to the original sources	https://huggingface.co/datasets/ukr-detect/crossnews-ua
SemEval 2022 Task 8 Dataset	CC BY 4.0, news released as links to the original sources	https://competitions.codalab.org/competitions/33835
Command-R	CC-BY-NC 4.0 License with Acceptable Use Addendum	https://cohere.com/c4ai-cc-by-nc-license
mBERT	Apache-2.0	https://huggingface.co/google-bert
XLM-RoBERTa	MIT	https://huggingface.co/FacebookAI
e5	MIT	https://huggingface.co/intfloat
mT5	Apache-2.0	https://huggingface.co/google
Mistral7B	Apache-2.0	https://huggingface.co/mistralai
Mixstral8x7B	Apache-2.0	https://huggingface.co/mistralai
LLaMa3.1	llama3.1	https://huggingface.co/meta-llama
Aya	CC-BY-NC 4.0 License with Acceptable Use Addendum	https://cohere.com/c4ai-cc-by-nc-license

Table 3: Overview of the licenses associated with each resource.

The licenses associated with the models and datasets utilised in this study are consistent with the intended use of conducting academic research on approaching various NLP applications for positive impact.

B Usage of AI Assistants

During this study, an AI assistant was utilised in the writing process. ChatGPT was employed for paraphrasing and improving clarity throughout the paper's formulation.

C Preparing the Dataset for Annotation: Prompts

Prompt, Mistral 7B Instruct, Keywords

C.1 Keywords Extraction for News Search in Ukrainian

```
<s>[INST] Extract top-10 keywords from a news article.
Order them from the most informative to the least,
in the following format:
1. <top-1 keyword>;
2. <top-2 keyword>;
... ;
10. <top-10 keyword>;
The news article:
"The weather will suddenly change in Kyiv on 13 March;
strong winds will rise in the capital of Ukraine.
This was reported by the press service of the State Emergency Service
of Ukraine in Kyiv with reference to the Ukrhydrometcenter.
In the city, the 1st level of danger is announced - yellow.
“In Kyiv in the next hour and until the end of the day on 13 March.
we will have wind gusts of 15-18 m/s strength,” - stated the message.
The weather will change dramatically in Kyiv. As reported by OBOZREVATEL,
in Kyiv, in March, real spring will come -
it will get warmer, and instead of snow, it will only rain." [/INST]
1. weather;
2. Ukraine;
3. danger;
4. Kyiv;
5. wind;
6. Ukrhydrometcenter;
7. March;
8. change;
9. spring;
10. strong; </s> [INST]
Extract 10 keywords from a news article.
The news article: "{NEWS ARTICLE}"
Order them from the most informative to the least,
in the following format:
1. <top-1 keyword>;
2. <top-2 keyword>;
... ;
10. <top-10 keyword>; [/INST]"
```

C.2 Preparing Summarization Format

Prompt, Command R Cohere, Summarization, English

```
## Instructions

    Summarize a news article focusing on four main points:
    1. WHO is the main/secondary actor in the news article?
    2. WHEN did the situation described in the news article occur?
    3. WHERE did the situation described in the news article take place?
    4. WHAT happened in the news article?
    (200-250 words summarization of the news article)

    ## Example output
    WHO: <answer to the 1st point of the Instructions>
    WHEN: <answer to the 2nd point of the Instructions>
    WHERE: <answer to the 3rd point of the Instructions>
    WHAT: <answer to the 4th point of the Instructions
    (200-250 words summary of the news article)>.

    ## News Article

    {NEWS ARTICLE}
```

D Interface design

D.1 Input data component

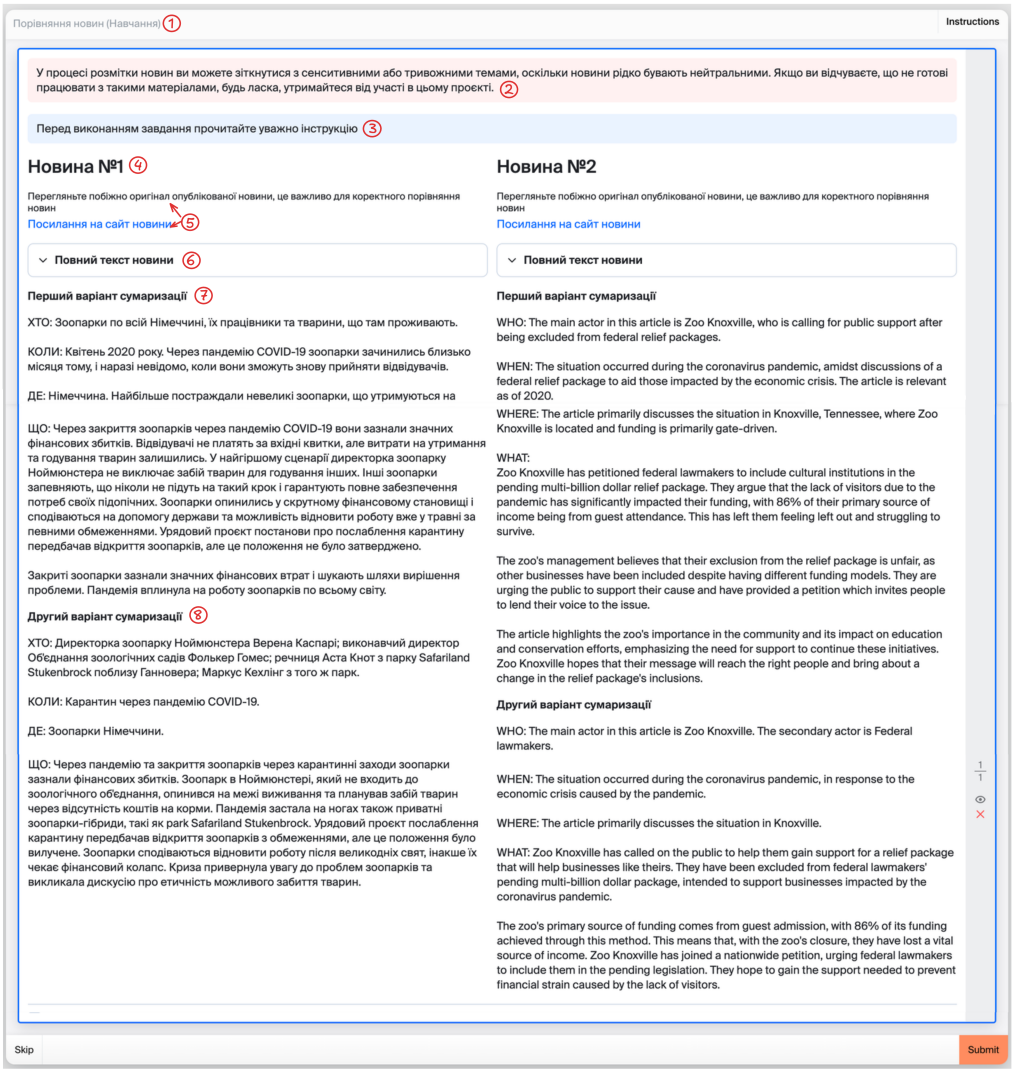


Figure 6: Interface, comparison of news articles, Shared Part

An English translation of the shared part displayed in Figure 6 of train/main projects interfaces:

1. **Project Name**: "News comparison (Training)"
2. **Disclaimer**: "In the process of labelling, you may come across sensitive or disturbing topics, as news articles are rarely neutral. If you feel that you are not ready to work with such materials, please refrain from participating in this project."
3. **Reminder**: "Read the instruction carefully before completing the task"
4. **News article header & Reminder** News Article №1. Skim through the original news article briefly; this is important for correctly comparing news articles.
5. **Link to the news article**
6. **Full text of the news article**

7. **The first version of summarization**

8. **The second version of summarization**

D.2 Training

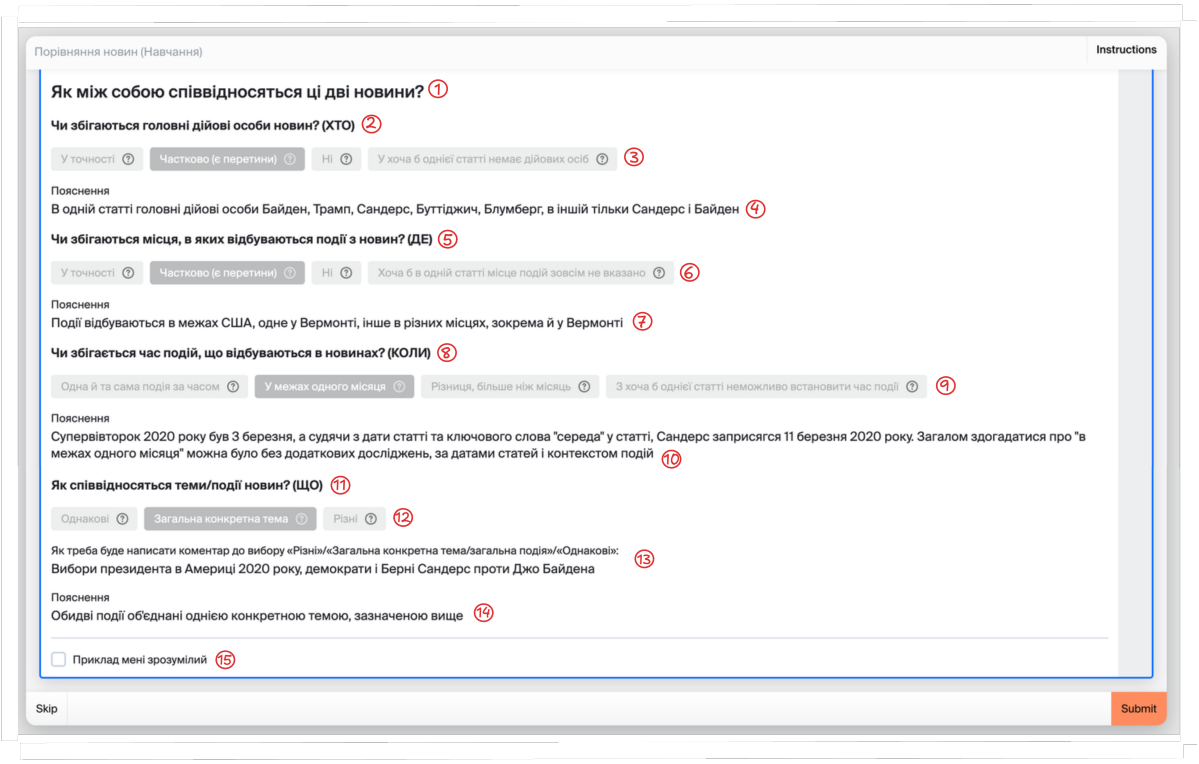


Figure 7: Interface, comparison of news articles, Training

The translation of the component of the training interface, shown in Figure 7, is provided below:

1. **How do these two news articles relate to each other?**

2. **Do the main actors in the news articles match? (WHO)**

3. **Options (from left to right)**: Exactly; Partly (there are overlaps); No; At least one article has no explicitly defined actors. *Question marks are action buttons showing hints, explaining the meaning of each option*

4. **Explanation**: In one article, the main actors are Biden, Trump, Sanders, Buttigieg, and Bloomberg. In the other, only Sanders and Biden are mentioned.

5. **Do the places where the events in the news are taking place match? (WHERE)**

6. **Options (from left to right)**: Exactly; Partly (there are overlaps); No; At least one article does not mention the place of the events.

7. **Explanation**: The events take place within the United States, one in Vermont, the other in various locations, including Vermont

8. **Do time ranges of events in the news coincide with each other? (WHEN)**

9. **Options (from left to right)**: The same event in time; Within one month; More than a month difference; It is impossible to determine the time of the event from at least one article.

10. **Explanation**: Super Tuesday 2020 was on March 3rd, and judging by the date of the article and the keyword “Wednesday” in the article, Sanders vowed to continue his campaign on March 11, 2020. It was generally possible to guess “within one month” without additional research based on the articles’ dates and the events’ context.

11. **How do news topics/events relate to each other? (WHAT)**

12. **Options (from left to right)**: Similar; Common specific topic; Different;

13. **Labeling example**: How to write a comment to the "Similar"/"Common specific topic/event"/"Different" choice: "American presidential election 2020, Democrats and Bernie Sanders vs. Joe Biden"

14. **Explanation** Both events are united by one specific topic mentioned above

15. **The example is clear to me**

D.3 Main pools

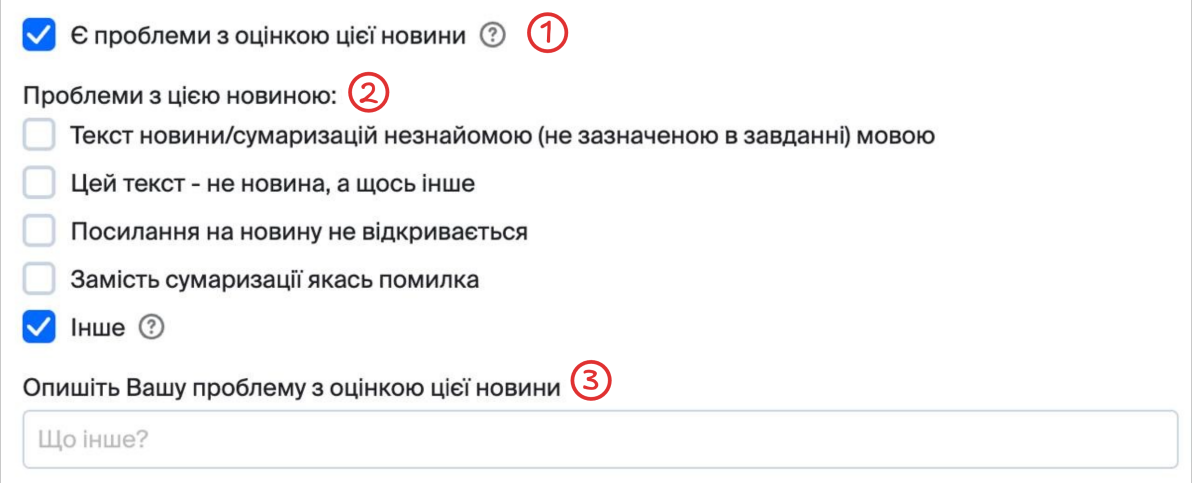


Figure 8: Interface, comparison of news articles, Errors Detection

A component for reporting problems with the input data is presented in a fully expanded form in Figure 8. Here’s the translation of it:

1. **There’s a problem with evaluating this news article** *Question mark is an action button explaining what’s meant by the "problem"*

2. **List of problems**: Problems with this news article:

   - Text of news article/summaries is in an unfamiliar language (not specified in the assignment)
   - This text is not a news article but something else
   - The link to the news article does not open
   - Instead of summarization, there is some kind of error
   - Other *Question mark is an action button explaining when to select "Other"*

3. If "Other" was chosen: Describe your problem with the evaluation of this news article: "What’s other?"

The main labelling component is presented in Figure 9.

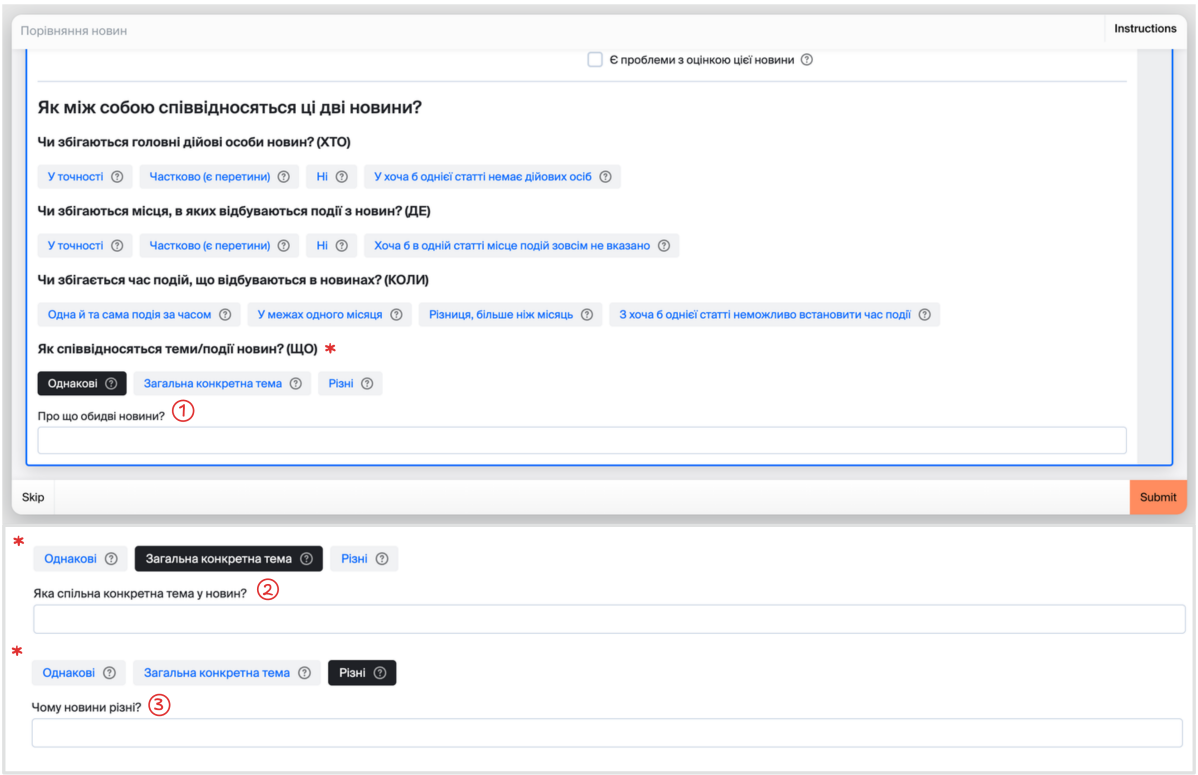


Figure 9: Interface, comparison of news articles, Exam & Main pools

E Instructions

In this section, we provide the Instructions we used for the training and the main project. We provide them translated to English, as the original language of instructions is Ukrainian.

E.1 Training

`Instruction for Training, translated to English`

This is the pre-project training. In it, we show you examples of completed tasks and corresponding explanations. Your goal is to familiarize yourself with the instructions and then study them deeper with the help of examples. After completing the training, you will gain access to the main project, where the assignments are paid. You will be provided a paid exam, which, when passed, will give you access to all the main tasks.
Thank you for your participation! Your input helps us to improve the quality of the news.

Disclaimer *We assume that if you have stated certain languages (Polish, Ukrainian, Russian or English) in your profile, you are able to complete the tasks in them.*

Task description

In each task, you will compare two news articles and evaluate their similarity.
The news articles in the pair may be in different languages, but this should not affect your comparison. For example, Wołodymyr Zełenski (Polish) and Volodymyr Zelenskyy (English) are the same actors in an event.
For each news article, we have provided:

- A link to the news website (be sure to follow it to see the original article, at least the headline and date of publication);
- Full text of the article (if you want to check some details right in the task interface);
- Two summarization options made by specially trained algorithms (including basic information on four main parameters of events: WHO, WHEN, WHERE and WHAT happened)

Assessing news similarity

You will answer four questions to help us draw a conclusion about the similarity of the news. Examples of answers to these questions with explanations will be shown throughout this training.

Do the main characters in the news match? (WHO)
The actor of an article can be a specific person (Volodymyr Zelenskyi), an organization/party or any other association of people.

- **Exactly** - the main and secondary actors in the articles are the same. At the same time, it is important to note that one actor can be called differently, for example, "Volodymyr Zelenskyi" and "President of Ukraine".
- **Partially** (there are overlaps) - Some of the main or secondary actors are the same.
- **No** - The protagonists in the articles do not coincide at all; they have nothing in common.
- **At least one article does not have any actors** - for example, one of the articles states something very vague, e.g. "the entire population of the Earth", or the article is a weather forecast.

Do the places where the events in the news occur match? (WHERE)

- **Exactly** - they are talking about the same place. If one article mentions a specific place/city/-country, it should be mentioned in the second article. If one article mentions only a country and the other a city, this is not an exact match. In this case, the same place may be referred to differently, such as "St. Petersburg" and "St. Petersburg".
- **Partially** (there are overlaps) - The place mentioned in one article is located within the boundaries of the place mentioned in another article. For example, if one article refers to events taking place in Ukraine and the other in Kyiv, this option is selected. Also, if different places (e.g., countries) are mentioned in the articles and some of the names of the places are the same.
- **No** - The places are entirely different (different places in the city, in the country, different countries).
- **At least one article does not mention the location at all** - something very vague, for example, something that happens worldwide, all over the Internet, or if the location is not specified. However, suppose the place of action can be understood, for example. In that case, the article says "in our country," and the author is obviously a Ukrainian, then, this option should not be selected. Do the times of the events in the news coincide?

Do the times of the events in the news coincide? (WHEN)

It is important to understand that we are interested in the time of the event in an article, NOT the publishing date. However, the article publication date can be very useful in determining the time of the event. Therefore, we ask you to follow the link to the original news article to use this information.

- **Same event in time** - it is obvious from the article that it is about the same event that happened simultaneously.
- **Within one month** - two events occurred within the same month.
- **More than a month apart** - the difference between the times of the events is more than a month.
- **From at least one article, it is impossible to establish the time of the event**. Even if the article's publication date is present, which is almost always the case. If, after reading the text of the article, it is impossible to determine when the event itself took place, or the time period in the article is very vague, for example, it states "at all times".

How do news topics/events relate to each other? (WHAT)

- **The same**
- **Common specific topic**
- **Different**

Same

News articles describe the same event in the same place, with the same people, at approximately the same time.
However, the events in the news may be described as follows:

- in a different tone/style;
- using different words/slang;

- with a different focus (one focuses more on one part of the event, the other on another);
- from different points of view, political position;
- with errors (to your knowledge) in the details, which makes the news seem different

If you are not 100% sure that the news is about the same event with the same actors, place and approximate time, then you should choose the option "Common specific topic".

Common specific topic
News stories are related to one narrow topic or one joint event but describe different incidents within it.

Examples:

- Curfews imposed due to COVID-19 (in different countries or changes in its status within the same country);
- Robbery of a grocery store done by teenagers (in different places or different events within the same major case);
- Joe Biden's 2020 election campaign and events within it.

News topics should be narrow and specific, so, for example, two news stories on broad and popular news area topics such as COVID-19 (in general), death (in general), and elections (in general ;) do not have to have any commonality in themselves.

Different
Events described in the news articles have little to do with each other. They either do not have a common theme, or the common theme is broad (e.g., COVID-19), and it is not a reason to consider the news relatable.

Justify your choice
Examples of choice justification are shown in the training.
If news articles:

- **are the same:** briefly describe, summarizing in 1-2 sentences, what event both articles mention.
- **united by a specific topic:** briefly describe this topic in 1-2 sentences.
- **are different:** briefly summarize in 1-2 sentences why, from your point of view, these topics are different (one is about ..., another is about ...)

The instructions for the main project are mostly identical to the ones for the training, with the only following difference:

E.2 Exam & Main pools

`Instruction for Exam & Main pools, translated to English`

Same as instruction in Training

Payment
We check and accept all assignments within the span of **10 days**. Incorrect assignments will be

rejected. We take the subjectivity of grades into account. You can file an appeal if you believe your assignment has been rejected unfairly.

Task description

Same as instruction in Training

Assessing news similarity

Problems with news articles' evaluation
If you encounter such problems as:

- Instead of summarization, there is some kind of error;
- It seems to you that it's not a news article in front of you but something else;
- An unfamiliar language (not the one you completed the Training on) prevails in the text of the news article/summaries so that it is impossible to understand what is being said;
- The link to the news article does not open (you waited a minute+);
- Any other problem.

The interface has a button "*There's a problem with evaluating this news article*" below each news article.
When you click on it, you can choose one of the existing options or select the option "*Other*" and describe a problem that is not listed.
In this case, you do NOT need to label the news articles; the task will be fully paid if we reproduce the error you specified when checking the task.

Same as instruction in Training

F LLM as an Annotation Quality Control Assistant: Prompt

Prompt, Acceptance-Rejection with Command R Cohere, "Different"

```
## Instruction
    You are provided with a comment explaining
    why the news articles in a pair are different.
    Your goal is to judge if, out of this comment, one can understand
    what these two news articles are about (Yes), or if
    the comment just says that the news articles are different (No).

    ## Examples
    For instance, out of the following comments:
    - "A lawsuit from journalists and the trial of a police officer"
    Yes, both news are briefly summarized.
    - "Martial arts in the U.S.A.;
    a review of five old martial arts fighters"
    Yes, both news are briefly summarized.
    - "The first article reports that Dagestan has launched
    the production of oxygen valves for medical equipment
    due to the increased demand caused by the pandemic.
    The second article is about a “counterterrorist operation”
    in Makhachkala and Kaspiysk."
    Yes, both news are briefly summarized.
    - "First covid-19 in Ukraine, second - parade in Russia."
    Yes, both news are briefly summarized.
    - "Assaulting a pensioner and the weather"
    Yes, both news are briefly summarized.
    - "Events are different"
    No, it just says that news articles are different.
    - "Completely different events"
    No, it just says that news articles are different.
    - "The events are not related in any way"
    No

    ## Input
    - "{COMMENT}"
```

G LLMs for News Semantic Similarity Assessment: Prompt Example

Template One-Shot Prompt for LLMs Benchmarking

```
Your goal is to compare news articles.
For each news article, two summarization options are provided
(including basic information on four parameters: WHO, WHEN, WHERE and
WHAT happened). News in a pair can be in different languages, but this
should not affect your comparison. For example, Wołodymyr Zełenski
(Polish) and Volodymyr Zelenskyy (English) are the same actors in an
event. You have to answer 4 sub-questions about the similarity of news
in JSON format:

**Do the main characters in the news match? (WHO?)**
The actor of an article can be a specific person (Volodymyr Zelenskyy),
an organization/party or any other association of people.
1. **yes**. The main and secondary protagonists in the articles are the
same. At the same time, it is important to note that one actor can be
called differently, for example,
"Volodymyr Zelenskyi" and "the President of Ukraine".
2. **partly**. (There are overlaps) Some of the main or secondary actors
are the same.
3. **no**. The actors in the articles do not overlap at all;
they have nothing in common.
4. **incomparable**. At least one article has no actors. For example,
one of the articles states something very vague, like
“the entire population of the Earth” or the article is a weather forecast.

**Do the places where the events in the news are taking place match
the places in the news? (WHERE?)**
The same place may be referred to in different ways, such as
"St. Petersburg" and "Peter".
1. **yes**. Articles mention exactly the same place. That is, if one
article mentions a specific place/city/country, it should be mentioned
in the second article. If one article mentions only a country and the
other - a city, this is not an exact match.
2. **partly**. (There are overlaps) The place mentioned in one article is
located within the boundaries of the place mentioned in another article.
For example, if one article refers to events taking place in Ukraine
and the other in Kyiv, this option is selected. It is also a valid choice
if different places (e.g., countries) are mentioned in the articles and
some of the names of the places are the same.
3. **no**. The places are completely different (different places
in the city, in the country, different countries).
4. **incomparable**. At least one article does not mention the location
at all. Something very vague, for example, something that happens all
over the world, all over the Internet, or if the location is not really
specified. At the same time, if the place of action can be understood,
for example, the article says "in our country," and the author is
obviously a Ukrainian, then this option should not be selected.
```

Does the timing of the events in the news coincide? (WHEN?)
It is important to understand that we are interested in the time/date of the news event, NOT the time/date of the article's publication.
1. **yes**. Same event time-wise. It is obvious from the article that it is about the same event that happened at the same time.
2. **partly**. Two events in news articles occurred within the same month.
3. **no**. The difference between the times of the events is more than a month.
4. **incomparable**. From at least one news article, it is impossible to establish the time of the event; it is impossible to determine when the event itself took place or the time period in the article is very vague, for example, “at all times”.

How do the news topics/events relate to each other? (WHAT?)
1. **similar**. News articles are about the same event.
2. **somewhat_related**. Common specific topic/common event.
3. **different** News are about different events.
1. **similar**
News articles describe the same event that happened in the same place, with the same people at approximately the same time.
However, the event might be described:
- with different focus (one focuses more on one part of the event, the other - on another)
- from different points of view, political positions
- with errors in the details, which makes the news seem “different”
If you are not 100% sure that the news is about the same event with the same actors, in the same place and at approximately the same time (some details make them different), then you should choose the option “Common specific topic/common event”.
2. **somewhat_related**
News articles are related to one close, narrow topic/event but describe different incidents within it.
Examples of narrow common topics:
- Curfews imposed due to COVID-19 (in different countries or changes in its status within the same country);
- Robbery of a grocery store by teenagers (in different places or different events within the same case);
- Joe Biden's 2020 election campaign and various events described from it.
News common topics should be narrow and specific: for example, two news stories on broad and popular topics in the news, aka COVID, death, and elections do not necessarily have to have any commonality in themselves.
3. **different**
Events in the news have little to do with each other. They either do not have a common theme, or the common theme is so broad (e.g., COVID) that it is not a reason to consider the news similar.

Provide **a short reasoning** for **WHAT?** answer in the following format:
- News are **similar**: Briefly describe, summarizing in 1-2 sentences, what the event described in both news is about.

- Both news are **somewhat_related**: Briefly, in 1-2 sentences, identify the common specific topic of news.
- News articles are about **different** events: Briefly summarize in 1-2 sentences: what are the first and the second news articles about?

Input format example:
Summarization of the first news article, version 1

ХТО: 27 учнів середньої школи з Марбурга та їхній водій.
КОЛИ: В неділю, 2023.
ДЕ: На автобані 45 у німецькій федеральній землі Саар.
ЩО: Двоповерховий автобус, який віз 27 учнів середньої школи та їхніх вчителів в Англію, потрапив у серйозну аварію. Три дівчини та хлопець отримали серйозні травми і були госпіталізовані, а також водій автобуса отримав легкі травми. Автобус з'їхав з дороги та перекинувся на бік. Поліція розслідує причину ДТП. Через аварію автострада у напрямку Дортмунда була повністю блокована. Шкільна екскурсія була скасована, і дітей повернули додому. Це не перша така аварія за участі шкільного автобуса в Німеччині - у квітні та грудні минулого року також були серйозні ДТП за участі шкільних автобусів, у яких постраждали діти.

Summarization of the first news article, version 2

ХТО: 27 учнів середньої школи у Марбурзі та їх 54-річний водій автобуса.
КОЛИ: У неділю о 6:27, 2023.
ДЕ: Автобан 45 у Саарі, неподалік Вендена в районі Ольпе.
ЩО: Двоповерховий автобус, який віз 27 учнів та п'ятеро вчителів зі старшої школи з німецького Марбурга в Англію, потрапив у аварію. За словами поліції Дортмунда, автобус з'їхав у праву сторону та перекинувся на бік. У результаті аварії постраждали 27 учнів та їх водій, чотири людини отримали серйозні травми, а решта - легкі. Усі постраждалі були доставлені до лікарні, їхньому життю нічого не загрожує. Шкільна екскурсія була скасована, а дітей повернули до Марбурга. Через аварію автобан у напрямку Дортмунда був повністю закритий протягом дня. Ця подія є не єдиною аварією шкільного автобуса в Німеччині. Раніше у квітні в місті Гольцмінден аварія за участю шкільного автобуса призвела до госпіталізації 16 учнів, а в грудні минулого року - до загибелі дитини через сильну ожеледицю.

Summarization of the second news article, version 1

ХТО: Стрілянину в Огайо влаштував 27-річний чоловік, через його дії одна людина загинула, а 26 отримали поранення. А перед цим у Словаччині було скоєно замах на прем'єра Роберта Фіцо, а в Туреччині - на відділок поліції, де загинули двоє людей.
КОЛИ: Масштабна стрілянина в Огайо відбулася в ніч на неділю, 2 червня, 2024. Замах на Роберта Фіцо відбувся 15 травня, а стрілянина в Туреччині - у відділенні поліції - сталася невдовзі після цього.
ДЕ: Події розгорталися в американському штаті Огайо, specifically на вулиці міста Акрон. У Словаччині замах на прем'єра відбувся перед місцевим Будинком культури в місті Хандлов.
А в Туреччині - у місті Адияман, у відділку поліції.

ЩО: В результаті стрілянини в Огайо одна особа загинула, а 26 отримали поранення; поліція прибала на місце події і виявила зброю. Це вже другий за останній час інцидент зі стріляниною в цьому штаті. Замах на прем'єра Словаччини залишив його у важкому стані, а в результаті стрілянини в Туреччині загинули двоє правоохоронців. Обидві події показують зростання насильства з використанням зброї у цих регіонах.

Summarization of the second news article, version 2

ХТО: Нападником виявився 27-річний чоловік, який відкрив стрілянину і вбив одну людину, а ще поранив 26 людей. Прем'єр Роберт Фіцо.
КОЛИ: Стрілянина відбулася опівночі у неділю, 2 червня, 2024. А замах на політика Роберта Фіцо стався 15 травня.
ДЕ: Інцидент із стріляниною відбувся у штаті Огайо, на вулиці міста Акрон. А напад на Роберта Фіцо відбувся у місті Хандлов.
ЩО: У штаті Огайо відбулася стрілянина, за участі 27-річного стрільця, внаслідок якої 1 особа загинула та ще 26 отримали поранення різного ступеня тяжкості. Правоохоронці вилучили зброю на місці події. Окрім того, у своїй відповіді я розповіла про замах на прем'єра Словаччини Роберта Фіцо, який відбувся 15 травня та про стрілянину у відділку поліції в турецькому місті Адияман, де правоохоронець відкрив вогонь по колегах, в результаті чого загинуло двоє людей та восьмеро було поранено.

Output format example:
```
{
"WHO": "no",
"WHO_reasoning": "German schoolchildren vs shooter from Ohio",
"WHERE": "no",
"WHERE_reasoning": "Saar vs Ohio",
"WHEN": "no",
"WHEN_reasoning": "2nd of June, 2024, and any event in 2023 are more than
a month apart",
"WHAT": "different",
"WHAT_reasoning": "The bus with schoolchildren that flipped over in
Saarland, Germany, has nothing to do with the shooting in Ohio,
which resulted in 27 people dead",
}
```

Input:
Summarization of the first news article, version 1
{SUMMARIZATION 1.1}
Summarization of the first news article, version 2
{SUMMARIZATION 1.2}
Summarization of the second news article, version 1
{SUMMARIZATION 2.1}
Summarization of the second news article, version 2
{SUMMARIZATION 2.2}

Return answer in JSON format:

H Evaluation Results of Multilingual News Semantic Similarity: Experiments per Language Pair

Category	Encoder-based Models					LLMs			
	BoW	mBERT	XLM-R	e5	mT5	Mistral7B	Mixtral7B	LLaMa3.1	Aya8B
Who	0.45	0.33	0.30	**0.63**	0.39	0.25	0.25	**0.44**	0.40
When	0.39	**0.51**	0.23	0.30	0.25	0.18	0.20	**0.30**	0.15
Where	0.60	0.30	0.35	**0.46**	0.27	0.35	0.34	**0.47**	0.33
What	0.65	0.18	0.17	**0.61**	0.35	0.44	**0.60**	0.47	0.35

Table 4: Macro-averaged F1-score of baselines tested on our proposed multilingual news semantic similarity dataset: `ukr-ukr`. The results in **bold** denotes the best scores per dimension per model type.

Category	Encoder-based Models					LLMs			
	BoW	mBERT	XLM-R	e5	mT5	Mistral7B	Mixtral7B	LLaMa3.1	Aya8B
Who	0.29	0.31	0.35	**0.57**	0.46	0.22	0.29	0.39	**0.45**
When	0.30	**0.40**	0.18	0.27	0.28	0.20	**0.40**	**0.40**	0.18
Where	0.39	0.31	0.39	**0.47**	0.31	0.33	0.37	**0.51**	0.30
What	0.47	0.21	0.13	**0.59**	0.33	0.41	**0.43**	0.41	0.37

Table 5: Macro-averaged F1-score of baselines tested on our proposed multilingual news semantic similarity dataset: `ukr-en`. The results in **bold** denotes the best scores per dimension per model type.

Category	Encoder-based Models					LLMs			
	BoW	mBERT	XLM-R	e5	mT5	Mistral7B	Mixtral7B	LLaMa3.1	Aya8B
Who	0.27	0.20	0.18	**0.49**	0.41	0.10	0.25	0.24	**0.49**
When	0.15	0.25	**0.33**	**0.33**	0.15	0.17	**0.41**	0.34	0.25
Where	0.45	0.27	0.33	**0.51**	0.28	0.14	0.23	**0.45**	0.18
What	0.34	0.17	0.10	0.35	**0.45**	0.11	0.50	**0.53**	0.50

Table 6: Macro-averaged F1-score of baselines tested on our proposed multilingual news semantic similarity dataset: `ukr-pl`. The results in **bold** denotes the best scores per dimension per model type.

Category	Encoder-based Models					LLMs			
	BoW	mBERT	XLM-R	e5	mT5	Mistral7B	Mixtral7B	LLaMa3.1	Aya8B
Who	0.35	0.17	0.17	**0.49**	0.39	0.27	0.27	0.25	**0.40**
When	0.20	0.10	0.10	**0.27**	0.26	0.15	0.15	**0.35**	0.11
Where	0.42	0.35	0.31	**0.50**	0.21	0.20	0.29	**0.44**	0.23
What	0.60	0.13	0.20	**0.55**	0.30	0.20	0.35	**0.58**	0.30

Table 7: Macro-averaged F1-score of baselines tested on our proposed multilingual news semantic similarity dataset: `ukr-ru`. The results in **bold** denotes the best scores per dimension per model type.