A Lithuanian Image-Caption Dataset: Combining News Media and Translated Benchmarks

Tautvydas Kvietkauskas, Ūla Šveiterytė

Faculty of Fundamental Science, Vilnius Gediminas Technical University

Abstract

Vision-Language Models (VLMs) are widely used in image captioning tasks, but their development for less common languages is limited by a lack of high-quality multimodal datasets. This poster presents a hybrid approach for creating Lithuanian image-caption datasets, combining two different data sources: articles from Lithuanian news portals with associated photographs, and international image-caption datasets: NoCaps, VizWiz, MS COCO, TextCaps, Flickr30k, SBU, DOCCI and Winoground. Data from the news portal underwent multi-stage filtering; articles, images, and text were cleaned and standardized. Following this processing, a dataset of Lithuanian news portals was created, comprising 272,964 articles and 364,122 images. Descriptions of international datasets were automatically translated into Lithuanian and evaluated using a quality control methodology based on English and Lithuanian evaluation metrics. English descriptions were assessed according to indicators of fluency, grammar, spelling, relevance, and semantic anomalies, using GPT-2, LanguageTool and pyspellchecker tools, whilst Lithuanian translations were evaluated using metrics based on LaBSE, Sentence-BERT MPNet and Hunspell. Based on these indicators, a composite quality score of 0.6 was calculated, enabling the automatic identification and removal of low-quality image-caption pairs and semantically inaccurate translations. After cleaning and filtering, a final bilingual dataset was formed from all the international sets, consisting of 1,162,252 unique images and 2,009,222 captions. By applying the proposed methodology, high-quality Lithuanian multimodal datasets were prepared, which can be used for the training, adaptation, and evaluation of vision-language models in the context of the Lithuanian language.

Combining News Media Dataset

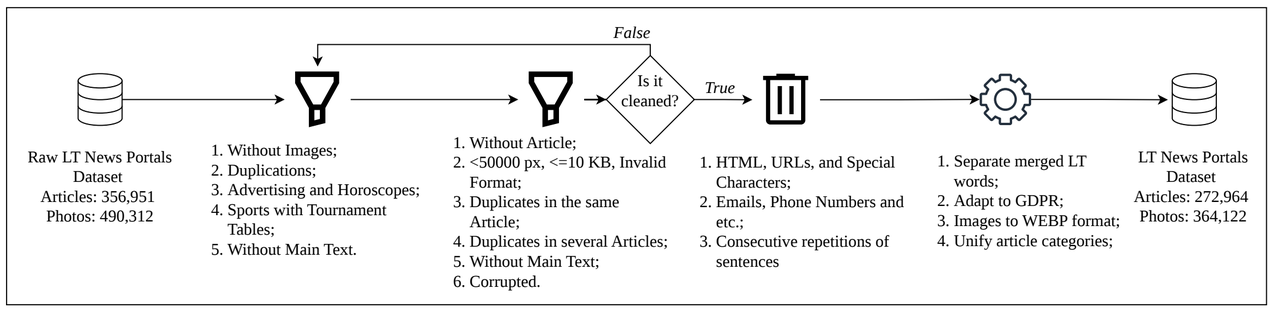


Figure 1. Data cleaning workflow for Lithuanian news media.

- The Raw LT News Portals Dataset initially contained **356,951 articles** and **490,312 images** collected from multiple Lithuanian news portals.
- Articles without images, duplicate entries, advertisements, horoscopes, sports articles dominated by tournament tables, and articles lacking meaningful content were removed.
- Corrupted files, low-quality images, unsupported formats, very low-resolution photographs, and duplicate images were filtered from the dataset.
- The textual data was cleaned by removing HTML tags, URLs, email addresses, telephone numbers, special characters, and repetitive sentences.
- The dataset was standardised by separating Lithuanian compound words, anonymising textual information according to GDPR principles, converting images to WEBP format, and harmonising article categories.
- The final LT News Portals Dataset contained **272,964 articles** and **364,122 images** and was used as the primary Lithuanian-language resource for image captioning experiments.

Image Captions Translated Dataset

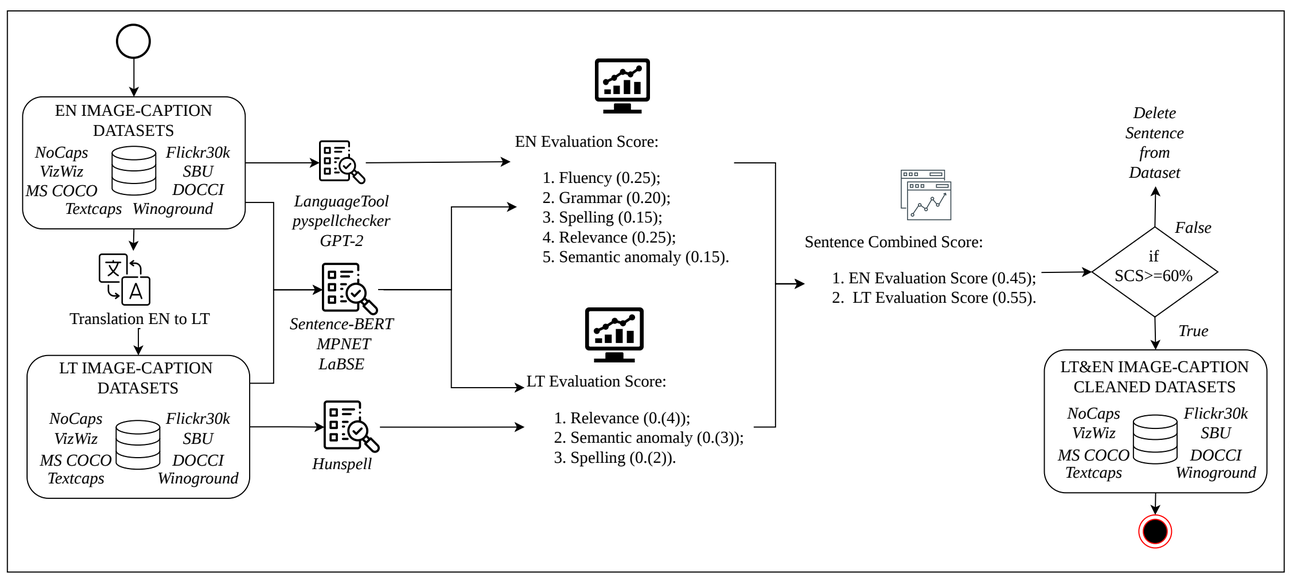
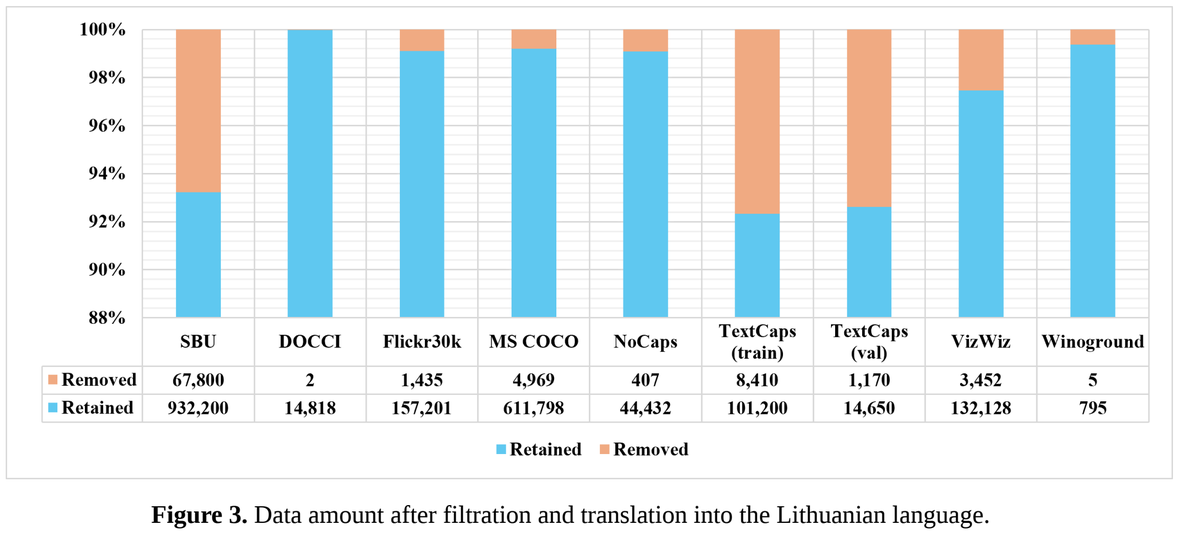


Figure 2. Data cleaning and benchmark workflow for Image-caption datasets.

- To increase the amount of Lithuanian image-caption data, several widely used international datasets, including NoCaps, VizWiz, MS COCO, TextCaps, Flickr30k, SBU, DOCCI, and Winoground, were combined and all English captions were automatically translated into Lithuanian, creating an initial bilingual image-caption dataset.
- Lithuanian caption quality was evaluated using LaBSE semantic similarity, Sentence-BERT MPNET semantic similarity, and Lithuanian spelling quality, with weights of 0.(4), 0.(3), and 0.(2), respectively.
- The quality of the original English captions was assessed using GPT-2 fluency evaluation, LanguageTool grammar checking, pyspellchecker spelling assessment, relevance analysis, and semantic anomaly detection, combined with weights of 0.25, 0.20, 0.15, 0.25, and 0.15.
- A combined quality score was calculated for each image-caption pair by assigning 0.55 weight to the Lithuanian evaluation and 0.45 weight to the English evaluation.
- Additional content safety analysis was performed to identify potentially inappropriate captions, classifying them as either confirmed inappropriate content or needs context cases requiring further review.
- An image-caption pair was included in the final dataset only if its combined quality score was at least 0.60 and it was not flagged as confirmed inappropriate content, resulting in a cleaned bilingual dataset suitable for vision-language model training and evaluation.

Image Captions Data Amount

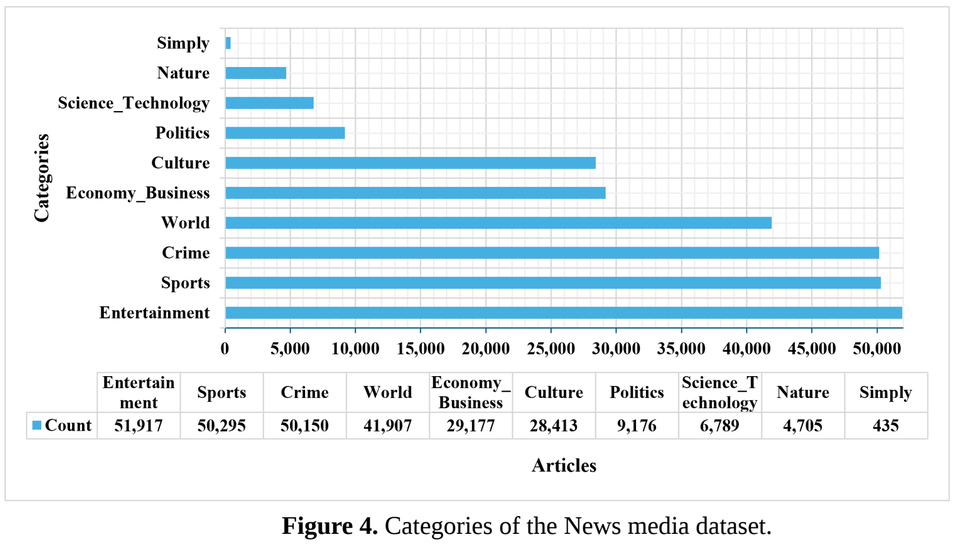


	SBU	DOCCI	Flickr30k	MS COCO	NoCaps	TextCaps (train)	TextCaps (val)	VizWiz	Winoground
Removed	67,800	2	1,435	4,969	407	8,410	1,170	3,452	5
Retained	932,200	14,818	157,201	611,798	44,432	101,200	14,650	132,128	795

Figure 3. Data amount after filtration and translation into the Lithuanian language.

The filtering pipeline removed 87,650 image-caption pairs (4.18%) from the original collection of 2.1 million samples, resulting in a final dataset of 2,009,222 high-quality image-caption pairs. The highest removal rates were observed in TextCaps (train) (7.67%), TextCaps (val) (7.40%), and SBU (6.78%), while datasets such as DOCCI, Winoground, MS COCO, and Flickr30k required minimal filtering. After cleaning and filtering, all international datasets were combined into a final bilingual dataset containing 1,162,252 unique images and 2,009,222 captions, providing a large-scale resource for vision-language model training and evaluation.

News Media Dataset Categories



	Entertainment	Sports	Crime	World	Economy_Business	Culture	Politics	Science_Technology	Nature	Simply
Count	51,917	50,295	50,150	41,907	29,177	28,413	9,176	6,789	4,705	435

Figure 4. Categories of the News media dataset.

The LT News Portals Dataset is stored in JSON format, where each record contains a news article together with its metadata and associated image information. Each entry includes the article title, summary, full text, publication date, URL, category, tags, and source portal information. In addition, image-related fields store image filenames, image URLs, and alternative text descriptions.

Problems and Limitations

The proposed QC pipeline substantially improved dataset quality; however, several limitations remain. The datasets differ in caption length, style, and domain, which can affect metric behavior and comparability. Semantic similarity measures (LaBSE and Sentence-BERT) estimate cross-lingual alignment but cannot guarantee factual translation correctness, while GPT-2, LanguageTool, Hunspell, and pyspellchecker provide only indirect indicators of quality and may incorrectly penalize valid names, technical terms, or uncommon expressions. In addition, the weighting scheme and filtering threshold were defined empirically and were not validated against human annotations.

Several limitations are also related to the automated collection and preprocessing of news portal data. Despite extensive cleaning, some articles with limited informational value (e.g., statistics-heavy or table-based content), duplicated sentences, merged words, or remaining text-formatting artifacts may still be present. Automatic GDPR-based anonymization can occasionally produce incorrect entity substitutions, and the evaluation process does not verify whether captions accurately describe the corresponding images. Therefore, while the proposed framework effectively removes many low-quality samples, it cannot guarantee complete semantic and visual correctness of all image-caption pairs.

Conclusions and Future works

This work presented the creation of large-scale Lithuanian image-caption resources by combining the LT News Portals Dataset with several international datasets, including MS COCO, NoCaps, VizWiz, TextCaps, Flickr30k, SBU, DOCCI, and Winoground. A comprehensive data cleaning, translation, standardisation, and quality-control pipeline was developed to improve dataset quality and remove noisy or problematic samples. The resulting datasets provide a valuable resource for Lithuanian image captioning and vision-language research.

Future work will focus on further improving the quality-control framework through stronger multilingual evaluation metrics, image-aware validation methods, and additional human evaluation. The LT News Portals Dataset and the bilingual image-caption dataset will continue to be expanded and refined, and will be used for vision-language model (VLM) training, fine-tuning, benchmarking, and evaluation tasks. Ultimately, these resources are intended to support the development, training, and assessment of Lithuanian vision-language models for image captioning and related multimodal applications.