University *of Ljubljana*
Faculty *of Computer and Information Science*

# Automatic generation of Slovenian traffic news for RTV Slovenija

Miha Frangež, Matej Medved, and Jan Mrak

**Abstract**

This project addresses the manual and resource-intensive process of generating Slovenian traffic news reports at RTV Slovenija. Our goal was to develop an automated system capable of transforming raw traffic data from the *promet.si* portal into concise, accurate, and stylistically appropriate radio news segments. We explored various Large Language Models (LLMs), including proprietary APIs like Google's Gemini series (*Gemini Flash 1.5, 2.0, 2.5 Flash*) and open-source models such as *GaMS* and *Gemma_3*. Key methodologies included prompt engineering with progressively refined templates and fine-tuning the *Gemma_3 4B* model using the *Unsloth* library on a dataset of paired *promet.si* inputs and historical RTV Slovenija broadcast scripts. Evaluation involved a varied approach: LLM-based scoring (using *Gemini Flash 2.5*), standard quantitative metrics (ROUGE and BERTScore), and qualitative human analysis of generated examples. Results indicate that proprietary models with well-crafted prompts generally performed strongly. Fine-tuned *Gemma_3 4B* showed competitive semantic similarity (BERTScore) but faced challenges in LLM-based evaluation and adherence to nuanced broadcasting rules, potentially due to model size or over-tuning. The study highlights the importance of combining automatic metrics with human qualitative assessment for this domain-specific generation task and suggests that larger open-source models may offer further improvements.

**Keywords**

traffic news, llm, prompt engineering, slovenian language, news generation, fine tuning, Unsloth, Gemma, ROUGE, BERTScore, Gemini evaluation, qualitative analysis

*Advisors: Slavko Žitnik*

## Introduction

Traffic reporting is an essential service for public broadcasters, providing critical real-time information that affects citizens' daily commutes and travel plans. Currently at RTV Slovenija, traffic news is produced manually, with students reviewing and transcribing data from the *promet.si* portal into structured news segments that are broadcast every 30 minutes. This manual process introduces several challenges:

- It is labor-intensive, requiring constant human monitoring.
- It creates potential for human error in transcription and prioritization.
- It limits the speed at which critical traffic information can be disseminated.
- It introduces inconsistencies in reporting style and format.

Our project aims to address these limitations by developing an automated system for generating Slovenian traffic news reports. By leveraging Large Language Models (LLMs), prompt engineering techniques, and fine-tuning approaches, we intend to create a solution that processes structured traffic data from the *promet.si* portal and automatically generates concise, accurate, and standardized traffic reports that adhere to RTV Slovenija's broadcasting guidelines.

### Related work

Similar work on this topic includes papers like Fast Hybrid Approach for Thai News Summarization [1]. The paper is solving the problem of news summarization of news in Thai language using a fine-tuned mBART model.

Another paper Enhancing Large Language Model Performance through Prompt Engineering Techniques [2] focuses on prompt engineering techniques to improve the performance of large language models, which is one of the approaches we

will use to generate traffic news reports.

In a similar paper Leveraging the Power of LLMs: A Fine-Tuning Approach for High-Quality Aspect-Based Summarization [3] the authors investigate the use of fine-tuning LLMs for purposes of generating summaries. They hypothesize that fine-tuning open-source LLMs like Llama2, Mistral, Gemma, and Aya on a domain-specific dataset will lead to superior aspect-based summaries compared to current state-of-the-art methods.

## Initial ideas

As both the input as well as the intended output texts are in Slovene, it is essential that we use a language model that is proficient in that language. As no powerful foundation models trained specifically on Slovene are available, we will most likely employ a model with strong multi-lingual abilities. Some proprietary models such as ChatGPT have been shown to perform well in Slovene language tasks, however, the inability to fine-tune them, as well as the practical issues with using a proprietary model, only available as an API from one provider, make this a suboptimal choice.

Some open-source or open-weights models such as *Llama_3, Gemma_3* and *DeepSeek_R1* might be more suitable for the task, as their inferior multilingual ability could be improved by fine-tuning on the example RTV Slovenija texts, perhaps to the point of surpassing proprietary models. Such models are also more suitable to specifics of this task, as the public broadcaster has very strict rules about the use of Slovene language and it is unlikely this standard could be achieved without fine-tuning.

As both the input and the output text is very short and the task calls for little reasoning, a purely prompt-based approach might be sufficient to achieve semantically correct output. As this leaves significant space in the context window of most models, a significant about of example text (few-shot prompting) as well as written rules about the style can be provided to the model, which may be sufficient to achieve the required writing style. If this is not sufficient, a feedback loop could be constructed, wherein the model is prompted to analyze and revise its writing based on additional examples and/or written rules.

If prompt-only techniques prove insufficient in reproducing the desired style, this could be improved by fine-tuning the model on the example corpus using methods such as *LoRA*. Fine-tuning is also likely to improve the accuracy of task-specific parts of the text, such as place names.

## Dataset

Our primary source of input data for traffic news generation is the *promet.si* website, operated by the *Traffic Information Centre for Public Roads*. This data was provided as an Excel export, containing timestamped traffic alerts categorized by topic (e.g., *Accident, Delay, Weather, Obstructions, Road works, Warning, International information, General*). For training, fine-tuning, and evaluation, we obtained a three-year archive of the final broadcast scripts from RTV Slovenija, which were originally in RTF (Rich Text Format) files.

To prepare this data for use with our models, we developed custom parsing tools:

- An **Excel reader** was implemented to process the *promet.si* data, extracting relevant alert information and timestamps.
- An **RTF parser** was created to extract the textual content (the actual broadcast script) from the RTV Slovenija archive files, along with associated metadata like air times.

This processing allowed us to create a structured dataset, including paired input-output examples (i.e., *promet.si* alerts leading up to a broadcast, and the corresponding RTV script) crucial for supervised fine-tuning and robust evaluation of our generation models.

## Methods

The first technique we employed was *prompt engineering* with a series of progressively refined templates. These were primarily executed using Google's Gemini API; for rapid iteration, we selected lightweight models such as **Gemini Flash 1.5** and **Gemini Flash 2.0**, later migrating to the more capable **Gemini 2.5 Flash** for generation.

In parallel, we explored the use of open-source models. We performed limited local deployments of models like **GaMS 2B** and the base **Gemma_3** (before fine-tuning). A significant step in this direction was the fine-tuning of the **Gemma_3 4B** model. We utilized the *Unsloth* library for efficient fine-tuning on our prepared dataset of paired *promet.si* inputs and RTV Slovenija outputs. The resulting fine-tuned model was subsequently uploaded to Hugging Face (as *mmedved/gemma-3-finetune-v2* and *mmedved/gemma-3-finetune-v6*) and integrated into our testing pipeline for comparison. We also attempted to scale to the larger **GaMS 9B** variants on the HPC cluster; after overcoming initial setup challenges, their output quality was still not on par with that of the **Gemini 2.5 Flash** for this specific task.

### Data Selection for Evaluation

To ensure a fair comparison between model outputs and human-written reports, we established a process for aligning input data. For a given RTV Slovenija broadcast script (with a known air time, extracted via our RTF parser), we selected traffic alerts from our parsed *promet.si* Excel data that were logged within the 30-minute window preceding the report's air time. This curated set of recent alerts served as the input to our models, aiming to replicate the information an RTV radio host would have had available.

### Automatic Evaluation

Our evaluation strategy combined qualitative assessments with established quantitative metrics. To select optimal prompt templates and compare model performances (including our fine-tuned **Gemma_3 4B**), we initially designed an automatic grading loop driven by **Gemini Flash 2.5**. For every generated report, this evaluator compared the output against the human-written RTV Slovenija reference, assigning three sub-scores on a 0–100 scale for event accuracy (referred to as "Event similarity"), road naming accuracy, and report length similarity. The composite score was the arithmetic mean of these sub-scores, grouped by model and template.

For a more rigorous quantitative assessment, we implemented scripts to calculate **ROUGE** and **BERTScore** metrics. The process involved:

1. Programmatically collecting all generated report files, excluding auxiliary comparison or evaluation files. These were grouped by their model-template combination, extracted from filenames.
2. Matching each generated report with an original RTV Slovenija report from our dataset. This was achieved by parsing timestamps from generated report filenames (e.g., format *YYYY-MM-DD_HH-MM-SS*) and original report metadata (e.g., format *YYYY-MM-DD HH:MM:SS*), requiring an exact match on year, month, day, and hour.
3. For each successfully matched pair of a generated prediction and an original reference:
    - **ROUGE scores** (ROUGE-1, ROUGE-2, ROUGE-L, and ROUGE-Lsum, each providing Precision, Recall, and F1-score) were calculated using the *rouge-score* Python library, with stemming enabled for Slovenian.
    - **BERTScore** (Precision, Recall, and F1-score) was computed using the *bert-score* library, specifically configured for the Slovenian language (*lang="sl"*).
4. Average ROUGE and BERTScore values were then determined for each model-template combination across all its successfully matched pairs.

A persistent challenge throughout the evaluation was that the content of RTV scripts did not always perfectly align with the information available in the *promet.si* Excel feeds for the corresponding time window, occasionally causing the evaluators (both automated metrics and LLM-based grading) to penalize otherwise good generations.

## Results

Our evaluation process combined LLM-based scoring (using Gemini Flash 2.5), standard lexical (ROUGE) and semantic (BERTScore) similarity metrics, and qualitative human assessments. Detailed scores for each model-template combination can be found in the Appendix (Tables 1, 2, and 3).

### Gemini-based Evaluation

The Gemini Flash 2.5 model was used to evaluate generated reports against references on three criteria: Event Similarity, Road Naming Accuracy, and Length Similarity (each on a 0-100 scale). As shown in Table 1 in the Appendix, proprietary models like *gemini-2.5-flash-preview-05-20* with later prompt versions (v6) showed higher overall scores, with *gemini-2.5-flash-preview-05-20-v6* achieving an average of 55.17. The base *google-gemma-3-4b-it* models performed moderately; for instance, *google-gemma-3-4b-it-v4* scored 54.87. Our fine-tuned Gemma models (*mmedved-gemma-3-finetune-**) registered notably lower scores in this specific LLM-based evaluation, for example, *mmedved-gemma-3-finetune-v2-v2* scored 18.33.

### BERTScore Evaluation

BERTScore provides a measure of semantic similarity. The F1-scores (detailed in Appendix Table 2) were relatively clustered, with *gemini-2.5-flash-preview-05-20-v6-multi-shot* achieving the highest F1-score of 0.7194. Interestingly, our fine-tuned models (*mmedved-gemma-3-finetune-**) performed competitively in terms of BERTScore F1, with scores around 0.68, surpassing some of the non-fine-tuned Gemma and older Gemini configurations. This suggests good semantic coherence with the reference texts despite lower scores on some other criteria.

### ROUGE-Lsum Evaluation

ROUGE-Lsum measures the longest common subsequence, considering sentence-level structure (see Appendix Table 3 for full results). Again, *gemini-2.5-flash-preview-05-20-v6-multi-shot* had the highest ROUGE-Lsum F1-score (0.3703). The fine-tuned models showed moderate ROUGE-Lsum scores, with *mmedved-gemma-3-finetune-v6-v6* at 0.2617, generally lower than the top Gemini configurations but outperforming some zero-shot base models in this metric.

### Interpretation of Automatic Scores

It is important to note that while these automatic metrics provide valuable quantitative insights, they serve as a rough estimate of overall quality. Human evaluation revealed instances where generations with lower automatic scores were qualitatively judged as more 'up to par' with RTV Slovenija's broadcasting standards. This discrepancy arises because metrics like ROUGE primarily measure n-gram overlap (lexical similarity), BERTScore assesses semantic similarity to the reference text, and the Gemini-based evaluation focuses on specific content aspects (event matching, naming, length) as judged by another LLM.

However, none of these fully capture strict adherence to implicit or explicit broadcasting rules (e.g., specific phrasing, conciseness beyond length, avoidance of certain constructions), stylistic nuances, or the critical aspect of factual correctness and prioritization of information when multiple alerts are present, especially if the reference texts themselves vary in

these aspects. For instance, the fine-tuned Gemma models, despite lower Gemini-based scores, sometimes produced outputs that were subjectively closer to the desired radio style in terms of fluency or phrasing, aspects not fully reflected in these particular automatic evaluations. Therefore, these scores are best interpreted in conjunction with qualitative human assessment and as guides for further model refinement.

The de-duplication pre-processing positively impacted report conciseness, observed qualitatively. Data alignment issues between *promet.si* feeds and RTV reference scripts also sometimes affected the scores. (Further quantitative results, including detailed tables comparing all models and prompt templates across all ROUGE and BERTScore metrics, will be presented as analysis is finalized, with key summaries presented in the Appendix.)

## Discussion

The evaluation results highlight several key points. Proprietary models, particularly newer versions of Gemini Flash with refined prompts, generally achieved higher scores on both LLM-based evaluation and lexical/semantic similarity metrics like ROUGE and BERTScore. This underscores their strong general capabilities.

The performance of the open-source Gemma 4B models, both base and fine-tuned variants, was mixed. While their BERTScore F1 values were often competitive, indicating good semantic understanding, they tended to score lower on the Gemini-based evaluation criteria and sometimes on ROUGE. Several factors might contribute to this. The 4B variant of Gemma, while capable, might be relatively small for consistently generating nuanced and stylistically precise Slovenian traffic reports that adhere to all implicit RTV guidelines. There's also a possibility that the fine-tuned Gemma models became somewhat over-tuned to the training data. This could manifest as occasionally reporting events that were not the absolute latest news in a dynamic test scenario or struggling to generalize perfectly when faced with a very large number of simultaneous input alerts, a situation where human journalists excel at prioritization. For instance, when many events are reported, the models sometimes struggled to select the most pertinent ones or maintain the concise radio style.

Future improvements could explore the use of larger Gemma variants (e.g., 9B or future larger open models proficient in Slovenian) which might offer better capacity for nuance and adherence to complex stylistic rules. Further refinement of the fine-tuning dataset and techniques, perhaps incorporating more varied scenarios or negative examples, could also help mitigate any over-tuning effects.

### Qualitative Analysis of Generated Examples

To illustrate the practical differences and challenges, below are some examples of generated reports.

**Fine-tuned Gemma (*mmedved/gemma-3-finetune-**)**
**Good Example:** This output is generally good in terms of language and structure, adhering to the expected radio format.

Zaradi del na cesti Duplica - Kamnik med Volčjim Potokom in bencinskim servisom velja do 22.7. zaprt promet. Zaradi nesreče sta zaprti regionalni cesti Lesce - Bled in Dragonja - Koper.

Na gorenjski avtocesti pred predorom Karavanke občasno zapirajo predor.

Do 22. ure velja prometna omejitev tovornih vozil, katerih največja dovoljena masa presega 7,5 t.

**Bad Example (due to English output):** The primary issue here is the model reverting to English, which is unacceptable for Slovenian traffic news.

Due to the Ljubljana marathon, several streets in Ljubljana will be closed today.

Due to the Ljubljana marathon, the Bežigrad exit will be closed in both directions, and the Dravlje exit towards Zadobrova.

Until 22:00, trucks with a maximum weight greater than 7.5 tons are restricted.

Road closures on the Ljubljana - Zagorje road: - Near Beričevo, it will be closed until 24:00. - At Šklendrovci, full closures will occur from 21:00 to 25.10. Every day from 8:00 to 17:00.

Road Škofja Loka - Gorenja vas will be closed in the Sten tunnel until 13:00.

The road over Vršič pass will be closed for roadworks until approximately November 8th, from Monday to Friday, between 7:30 and 17:00.

## References

[1] Kietikul Jearanaitanakij, Suratan Boonpong, Kirttiphoom Teainnagrm, Thanakrit Thonglor, Tiwat Kullawan, and Chankit Yongpiyakul. Fast hybrid approach for thai news summarization. *Engineering and Technology Horizons*, 41(3):410307, Sep. 2024.

[2] Arlo Octavia and Meade Cleti. Enhancing large language model performance through prompt engineering techniques, 10 2024.

[3] Ankan Mullick, Sombit Bose, Rounak Saha, Ayan Kumar Bhowmick, Aditya Vempaty, Pawan Goyal, Niloy Ganguly, Prasenjit Dey, and Ravi Kokku. Leveraging the power of llms: A fine-tuning approach for high-quality aspect-based summarization, 2024.

# 1. Appendix

## A.1 Gemini-based Evaluation Scores

**Table 1.** Gemini-based Evaluation Scores (0-100 scale)

| Model-Template | Event Sim. | Road Name Acc. | Length Sim. | Overall Avg. | Count |
|---|---|---|---|---|---|
| *cjvt-GaMS-2B-Instruct-v6* | 0.00 | 0.00 | 95.00 | 31.67 | 1 |
| *cjvt-GaMS-2B-Instruct-v6-multi-shot* | 1.67 | 3.33 | 90.00 | 31.67 | 3 |
| *gemini-2.0-flash-v1* | 42.50 | 50.00 | 58.33 | 50.28 | 6 |
| *gemini-2.0-flash-v2* | 35.00 | 55.00 | 45.83 | 45.28 | 6 |
| *gemini-2.0-flash-v3* | 37.50 | 60.00 | 57.50 | 51.67 | 6 |
| *gemini-2.0-flash-v4* | 37.50 | 53.33 | 67.50 | 52.78 | 6 |
| *gemini-2.0-flash-v5* | 32.50 | 41.67 | 62.50 | 45.56 | 6 |
| *gemini-2.0-flash-v6* | 27.50 | 45.00 | 47.50 | 40.00 | 6 |
| *gemini-2.0-flash-v6-multi-shot* | 30.00 | 40.83 | 65.00 | 45.28 | 6 |
| *gemini-2.0-flash-v6-zero-shot* | 29.17 | 35.83 | 62.50 | 42.50 | 6 |
| *gemini-2.5-flash-preview-05-20-v1* | 30.83 | 47.50 | 40.83 | 39.72 | 6 |
| *gemini-2.5-flash-preview-05-20-v2* | 37.50 | 51.67 | 46.67 | 45.28 | 6 |
| *gemini-2.5-flash-preview-05-20-v3* | 34.17 | 54.17 | 63.33 | 50.56 | 6 |
| *gemini-2.5-flash-preview-05-20-v4* | 32.50 | 51.67 | 54.17 | 46.11 | 6 |
| *gemini-2.5-flash-preview-05-20-v5* | 40.00 | 47.50 | 53.33 | 46.94 | 6 |
| *gemini-2.5-flash-preview-05-20-v6* | 33.33 | 64.17 | 68.00 | 55.17 | 6 |
| *gemini-2.5-flash-preview-05-20-v6-multi-shot* | 41.67 | 53.33 | 64.17 | 53.06 | 6 |
| *gemini-2.5-flash-preview-05-20-v6-zero-shot* | 36.67 | 47.50 | 53.33 | 45.83 | 6 |
| *google-gemma-3-1b-it-v6-multi-shot* | 21.00 | 24.00 | 18.00 | 21.00 | 5 |
| *google-gemma-3-4b-it-v1* | 33.00 | 54.00 | 60.00 | 49.00 | 5 |
| *google-gemma-3-4b-it-v2* | 36.00 | 42.60 | 45.00 | 41.20 | 5 |
| *google-gemma-3-4b-it-v3* | 38.33 | 42.50 | 60.83 | 47.22 | 6 |
| *google-gemma-3-4b-it-v4* | 36.00 | 67.00 | 61.60 | 54.87 | 5 |
| *google-gemma-3-4b-it-v5* | 31.67 | 55.83 | 72.50 | 53.33 | 6 |
| *google-gemma-3-4b-it-v6* | 37.50 | 47.50 | 62.50 | 49.17 | 6 |
| *google-gemma-3-4b-it-v6-multi-shot* | 40.83 | 55.00 | 56.33 | 50.72 | 6 |
| *google-gemma-3-4b-it-v6-zero-shot* | 35.00 | 45.83 | 48.83 | 43.22 | 6 |
| *mmedved-gemma-3-finetune-v2-v2* | 7.00 | 14.00 | 34.00 | 18.33 | 5 |
| *mmedved-gemma-3-finetune-v2-v6* | 39.17 | 31.67 | 51.67 | 40.83 | 6 |
| *mmedved-gemma-3-finetune-v6-v6* | 9.00 | 13.40 | 54.00 | 25.47 | 5 |

## A.2 BERTScore Evaluation Results

## A.3 ROUGE-Lsum Evaluation Results

**Table 2.** BERTScore Evaluation Results

| Model-Template | Precision | Recall | F1-Score |
|---|---|---|---|
| *cjvt-GaMS-2B-Instruct-v6-multi-shot* | 0.6616 | 0.6723 | 0.6669 |
| *gemini-2.0-flash-v6-multi-shot* | 0.6590 | 0.7005 | 0.6787 |
| *gemini-2.0-flash-v6-zero-shot* | 0.6698 | 0.7143 | 0.6910 |
| *gemini-2.5-flash-preview-05-20-v6-multi-shot* | 0.6910 | 0.7514 | 0.7194 |
| *gemini-2.5-flash-preview-05-20-v6-zero-shot* | 0.6739 | 0.7382 | 0.7041 |
| *google-gemma-3-4b-it-v6-multi-shot* | 0.6714 | 0.7317 | 0.6999 |
| *google-gemma-3-4b-it-v6-zero-shot* | 0.6463 | 0.6981 | 0.6709 |
| *mmedved-gemma-3-finetune-v2-v2* | 0.6642 | 0.7004 | 0.6808 |
| *mmedved-gemma-3-finetune-v6-v6* | 0.6631 | 0.7048 | 0.6831 |

**Table 3.** ROUGE-Lsum Evaluation Results

| Model-Template | Precision | Recall | F1-Score |
|---|---|---|---|
| *cjvt-GaMS-2B-Instruct-v6-multi-shot* | 0.1605 | 0.1773 | 0.1674 |
| *gemini-2.0-flash-v6-multi-shot* | 0.3155 | 0.2971 | 0.2918 |
| *gemini-2.0-flash-v6-zero-shot* | 0.3209 | 0.3198 | 0.3113 |
| *gemini-2.5-flash-preview-05-20-v6-multi-shot* | 0.3516 | 0.4229 | 0.3703 |
| *gemini-2.5-flash-preview-05-20-v6-zero-shot* | 0.2910 | 0.3904 | 0.3185 |
| *google-gemma-3-4b-it-v6-multi-shot* | 0.2878 | 0.3803 | 0.3167 |
| *google-gemma-3-4b-it-v6-zero-shot* | 0.1940 | 0.2534 | 0.2117 |
| *mmedved-gemma-3-finetune-v2-v2* | 0.2097 | 0.2840 | 0.2265 |
| *mmedved-gemma-3-finetune-v6-v6* | 0.2717 | 0.2958 | 0.2617 |