FRI Natural language processing course 2025
Project report

University of Ljubljana
Faculty of Computer and
Information Science

Automatic generation of Slovenian traffic news for RTV Slovenija

Miha Lazić, Luka Gulič, and Matevž Crček

Abstract

/

Keywords

/

Advisors: Slavko Žitnik

Introduction

Traffic reporting is a crucial component of public broadcasting, providing real-time updates on road conditions, accidents, and congestion. At RTV Slovenija, traffic news is currently compiled manually by students who review traffic reports from the *promet.si* portal and transcribe them into structured news segments, which are then broadcast every 30 minutes. This manual approach is time-consuming, prone to human error, and inefficient given the volume of traffic data that needs to be processed.

With the advancement of large language models (LLMs), there is an opportunity to automate the generation of structured and readable traffic news. This project aims to leverage an existing LLM, apply prompt engineering techniques and fine-tune the model to automatically generate traffic reports that adhere to the broadcasting standards of RTV Slovenija. In the next chapter, we present various existing solutions that we consider using in this project.

Related Work and Existing Solutions

In their master's thesis, Bjørgo and Lundegaard explored the automation of traffic incident reporting. They developed a system that collects real-time traffic data, processes it to identify incidents, and automatically generates reports for dissemination to the public and relevant authorities. Their work provides a detailed description of approaches that could be highly beneficial to our project [1].

LLM Selection

For this project, we consider three potential LLMs: *LLaMA*, *FLAN-T5*, and *Bloom*. These models have been pre-trained on multilingual datasets, including Slovenian text, enabling them to process Slovenian-language prompts. However, none of these models have been specifically trained on Slovenian traffic data, making fine-tuning essential for achieving optimal performance.

Both *LLaMA* and *FLAN-T5* demonstrate strong reasoning capabilities, making them well-suited for structured traffic report generation. In contrast, *Bloom* is less effective in reasoning tasks, which may lead to lower performance when applying *Chain-of-Thought Prompting*.

For evaluation, we think about combination of automatic and semi-automatic evaluation criteria. Automatic evaluation quantitatively assesses LLM outputs without human intervention. For this, we employ algorithms such as *BLEU*, which measures similarity to reference texts, and *ROUGE*, which evaluates recall-based overlap. However, for a more comprehensive assessment, human involvement is essential. Therefore, we incorporate semi-automatic evaluation, leveraging AI-assisted scoring. Specifically, we use another LLM, to grade generated outputs based on predefined criteria. Each score is then manually reviewed and adjusted as needed to ensure accuracy.

Prompt Engineering

To generate traffic reports, we consider using a combination of *Role-Based Prompting* and *Few-Shot Prompting*. This approach is particularly suitable because our dataset consists of structured traffic data paired with human-written traffic reports. By assigning the LLM the role of a traffic reporter and providing it with well-structured example reports, we aim to guide the model towards generating high-quality outputs.

Before report generation, the model will be given explicit instructions on how to structure the reports, which include:

- The location of the incident, specifying the affected road and direction.
- The cause of the disruption and its consequences.
- The specific section of the road impacted.

Additionally, the dataset contains road names that differ from those used in official reports, requiring the model to handle these discrepancies appropriately.

To ensure proper prioritization of events, an event hierarchy must be established, as some incidents are more critical than others. If the model fails to respect this hierarchy, we will experiment with *Chain-of-Thought Prompting* before fine-tuning the model. This technique introduces step-by-step reasoning, helping the model determine which incidents should be reported first.

Data

The primary dataset is the traffic_reports_2022_2023_2024.xlsx file from the *promet.si* portal, which contains 170,823 traffic reports from the years 2022, 2023, and 2024, organized into 27 attributes, as illustrated in Figure 1. Based on the findings in [1], incorporating diverse data sources is essential. To enhance our dataset, we could integrate additional factors such as weather conditions and seasonal variations.

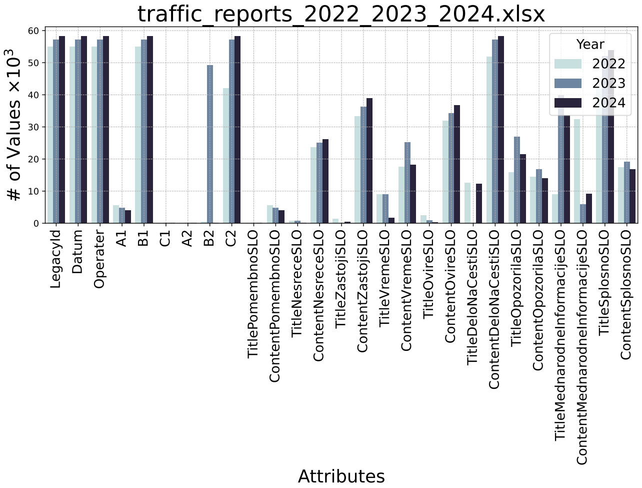

Figure 1. Project's primary dataset visualization.

Methods

Use the Methods section to describe what you did an how you did it – in what way did you prepare the data, what algorithms did you use, how did you test various solutions ... Provide all the required details for a reproduction of your work.

Below are LaTeXexamples of some common elements that you will probably need when writing your report (e.g. figures, equations, lists, code examples ...).

Equations

You can write equations inline, e.g. $\cos\pi = -1$, $E = m \cdot c^2$ and α, or you can include them as separate objects. The Bayes's rule is stated mathematically as:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}, \tag{1}$$

where A and B are some events. You can also reference it – the equation 1 describes the Bayes's rule.

Lists

We can insert numbered and bullet lists:

1. First item in the list.
2. Second item in the list.
3. Third item in the list.

- First item in the list.
- Second item in the list.
- Third item in the list.

We can use the description environment to define or describe key terms and phrases.

Word What is a word?.

Concept What is a concept?

Idea What is an idea?

Random text

This text is inserted only to make this template look more like a proper report. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Etiam blandit dictum facilisis. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Interdum et malesuada fames ac ante ipsum primis in faucibus. Etiam convallis tellus velit, quis ornare ipsum aliquam id. Maecenas tempus mauris sit amet libero elementum eleifend. Nulla nunc orci, consectetur non consequat ac, consequat non nisl. Aenean vitae dui nec ex fringilla malesuada. Proin elit libero, faucibus eget neque quis, condimentum laoreet urna. Etiam at nunc quis felis pulvinar dignissim. Phasellus turpis turpis, vestibulum eget imperdiet in, molestie eget neque. Curabitur quis ante sed nunc varius dictum non quis nisl. Donec nec lobortis velit. Ut cursus, libero efficitur dictum imperdiet, odio mi fermentum dui, id vulputate metus velit sit amet risus. Nulla vel volutpat elit. Mauris ex erat, pulvinar ac accumsan sit amet, ultrices sit amet turpis.

Phasellus in ligula nunc. Vivamus sem lorem, malesuada sed pretium quis, varius convallis lectus. Quisque in risus nec lectus lobortis gravida non a sem. Quisque et vestibulum sem, vel mollis dolor. Nullam ante ex, scelerisque ac efficitur vel, rhoncus quis lectus. Pellentesque scelerisque efficitur purus in faucibus. Maecenas vestibulum vulputate nisl sed vestibulum. Nullam varius turpis in hendrerit posuere.

Figures

You can insert figures that span over the whole page, or over just a single column. The first one, Figure 2, is an example of a figure that spans only across one of the two columns in the report.

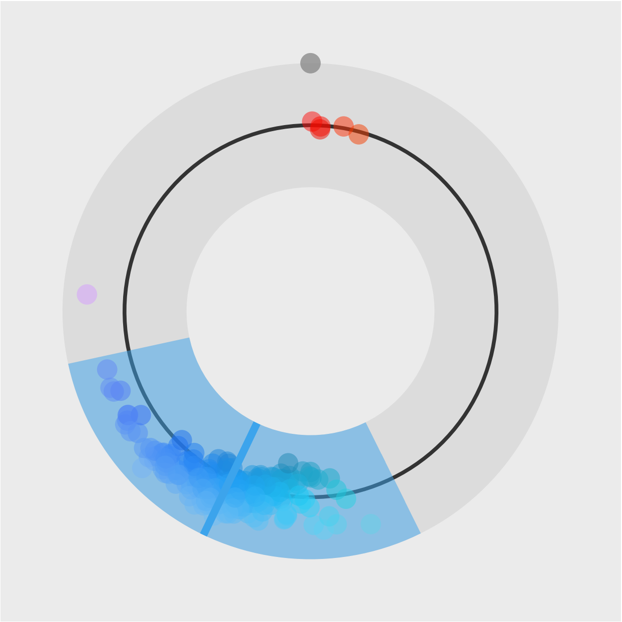

Figure 2. A random visualization. This is an example of a figure that spans only across one of the two columns.

On the other hand, Figure 3 is an example of a figure that spans across the whole page (across both columns) of the report.

Tables

Use the table environment to insert tables.

Table 1. Table of grades.

Name		
First name	Last Name	Grade
John	Doe	7.5
Jane	Doe	10
Mike	Smith	8

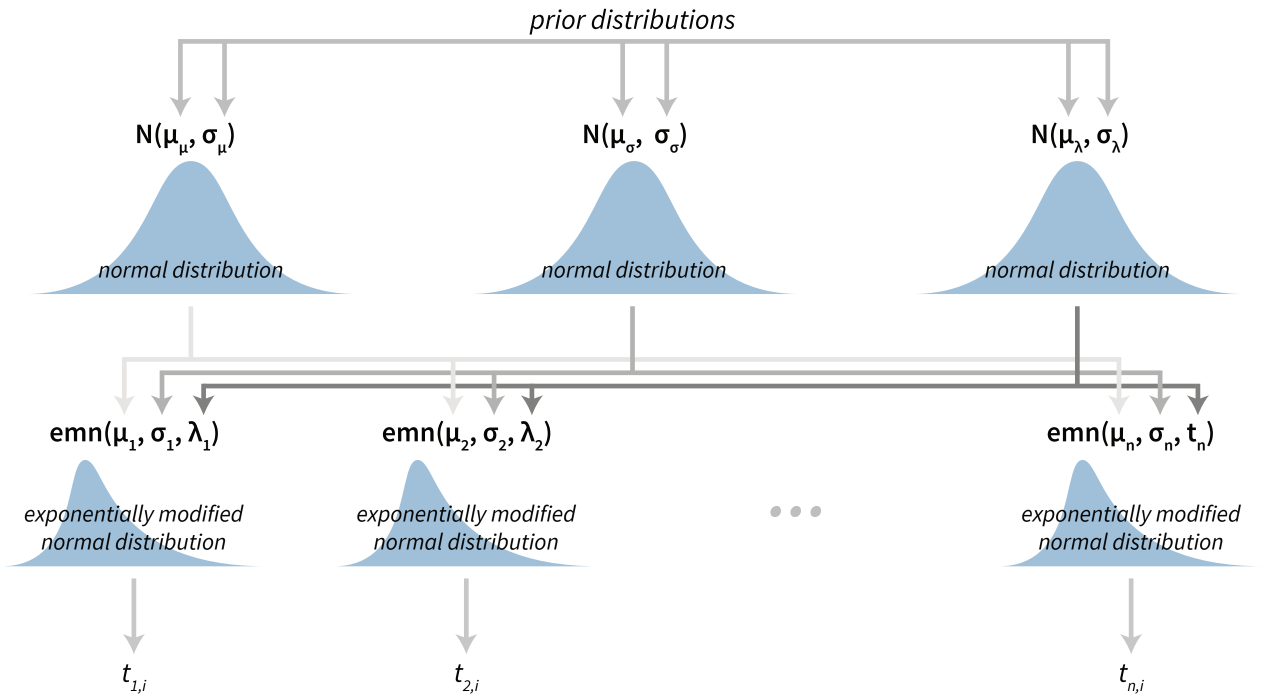
prior distributions
$N(\mu_\mu, \sigma_\mu)$
normal distribution
$N(\mu_\sigma, \sigma_\sigma)$
normal distribution
$N(\mu_\lambda, \sigma_\lambda)$
normal distribution
$emn(\mu_1, \sigma_1, \lambda_1)$
exponentially modified normal distribution
$emn(\mu_2, \sigma_2, \lambda_2)$
exponentially modified normal distribution
• • •
$emn(\mu_n, \sigma_n, t_n)$
exponentially modified normal distribution
$t_{1,i}$
$t_{2,i}$
$t_{n,i}$

Figure 3. Visualization of a Bayesian hierarchical model. This is an example of a figure that spans the whole width of the report.

Code examples

You can also insert short code examples. You can specify them manually, or insert a whole file with code. Please avoid inserting long code snippets, advisors will have access to your repositories and can take a look at your code there. If necessary, you can use this technique to insert code (or pseudo code) of short algorithms that are crucial for the understanding of the manuscript.

Listing 1. Insert code directly from a file.

```
import os
import time
import random

fruits = ["apple", "banana", "cherry"]
for x in fruits:
  print(x)
```

Listing 2. Write the code you want to insert.

```
import(dplyr)
import(ggplot)

ggplot(diamonds,
           aes(x=carat, y=price, color=cut)) +
  geom_point() +
  geom_smooth()
```

Results

Use the results section to present the final results of your work. Present the results in a objective and scientific fashion. Use visualisations to convey your results in a clear and efficient manner. When comparing results between various techniques use appropriate statistical methodology.

More random text

This text is inserted only to make this template look more like a proper report. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Etiam blandit dictum facilisis. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Interdum et malesuada fames ac ante ipsum primis in faucibus. Etiam convallis tellus velit, quis ornare ipsum aliquam id. Maecenas tempus mauris sit amet libero elementum eleifend. Nulla nunc orci, consectetur non consequat ac, consequat non nisl. Aenean vitae dui nec ex fringilla malesuada. Proin elit libero, faucibus eget neque quis, condimentum laoreet urna. Etiam at nunc quis felis pulvinar dignissim. Phasellus turpis turpis, vestibulum eget imperdiet in, molestie eget neque. Curabitur quis ante sed nunc varius dictum non quis nisl. Donec nec lobortis velit. Ut cursus, libero efficitur dictum imperdiet, odio mi fermentum dui, id vulputate metus velit sit amet risus. Nulla vel volutpat elit. Mauris ex erat, pulvinar ac accumsan sit amet, ultrices sit amet turpis.

Phasellus in ligula nunc. Vivamus sem lorem, malesuada sed pretium quis, varius convallis lectus. Quisque in risus nec lectus lobortis gravida non a sem. Quisque et vestibulum sem, vel mollis dolor. Nullam ante ex, scelerisque ac efficitur vel, rhoncus quis lectus. Pellentesque scelerisque efficitur purus in faucibus. Maecenas vestibulum vulputate nisl sed vestibulum. Nullam varius turpis in hendrerit posuere.

Nulla rhoncus tortor eget ipsum commodo lacinia sit amet eu urna. Cras maximus leo mauris, ac congue eros sollicitudin ac. Integer vel erat varius, scelerisque orci eu, tristique purus. Proin id leo quis ante pharetra suscipit et non magna. Morbi in volutpat erat. Vivamus sit amet libero eu lacus pulvinar pharetra sed at felis. Vivamus non nibh a orci viverra rhoncus sit amet ullamcorper sem. Ut nec tempor dui. Aliquam convallis vitae nisi ac volutpat. Nam accumsan, erat eget faucibus commodo, ligula dui cursus nisi, at laoreet odio augue id eros. Curabitur quis tellus eget nunc ornare auctor.

Discussion

Use the Discussion section to objectively evaluate your work, do not just put praise on everything you did, be critical and exposes flaws and weaknesses of your solution. You can also explain what you would do differently if you would be able to start again and what upgrades could be done on the project in the future.

Acknowledgments

Here you can thank other persons (advisors, colleagues ...) that contributed to the successful completion of your project.

References

[1] Anna Jansdatter Bjørgo and Mads Lundegaard. Automating the process of traffic incident reporting. Master's thesis, NTNU, 2024.