FRI Natural language processing course 2024
Project report

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

Natural language inference dataset

Lea Cvetko, Arja Hojnik, and Katarina Plementaš

Abstract

The abstract goes here. The abstract goes here.

Keywords
Keyword1, Keyword2, Keyword3 ...

Advisor: Aleš Žagar

Introduction

Natural Language Processing (NLP) continually evolves to bridge the gap between human language understanding and machine interpretation. Despite significant advancements, challenges remain, notably in processing less-resourced languages like Slovenian. This project focuses on constructing a Slovenian Natural Language Inference dataset, a critical resource for testing machine understanding of text in entailment, contradiction, and neutrality contexts. With an interest in making AI systems more accessible and effective across diverse linguistic landscapes, our project aims to push the boundaries of what's currently achievable with Slovenian language processing.

By utilizing Large Language Models (LLMs) to generate text pairs, we aim to produce a diverse and challenging dataset that reflects the complexity of natural language reasoning. Recognizing the foundational work done on the Slovenian Natural Language Inference (SI-NLI) dataset, we embark on an ambitious initiative to further enrich this dataset.

Related work

Current approaches to NLI rely heavily on datasets such as SNLI and MultiNLI, which are predominantly in English. While these datasets have driven significant advancements in NLP, the reliance on English limits the applicability of derived models to other languages. Even though recent efforts, like XNLI (The Cross-Lingual NLI Corpus), attempt to bridge this gap by providing multilingual extensions, the quality and quantity of data for Slovenian remain insufficient. On the LLM front, models like GPT-3 and BERT also show promising results in understanding and generating natural language.

Dataset

For the purpose of our project, we chose to work with the GigaFida (ccGigaFida 1.0) corpus, which is the largest corpus of written Slovene. It contains texts from newspapers, magazines, books, and web publications, which is why it presents a comprehensive resource that reflects contemporary usage of Slovene, making it valuable for generating text passages that capture current linguistic trends. This rich linguistic foundation is crucial for developing a Slovenian Natural Language Inference dataset aimed at enhancing machine understanding of textual entailment, neutrality, and contradiction.

Methodology

In this section, we focus on the methods used for developing such dataset.

1. **Preprocessing:** We chose to download the corpus encoded in TEI-like format with annotations in Slovenian, for easier processing in Python. It consists of 39,427 .xml files, each containing one excerpt from various different fields of public interest. Then we performed some basic preprocessing. The text in original files is written in the form of lemmas, so the first step included extracting relevant textual content, ignoring metadata, headers and footers. We also removed non-textual elements and any unnecessary whitespace and saved

the processed paragraphs into a new folder. Upon realizing that the whole process takes quite a lot of time and RAM, we decided to fix the code and divide the documents into three sections, with each of the members analyzing around 13,000 files.

Document similarity: To make our dataset more representative, we decided to identify the most common themes that appear in the corpus and then extract the paragraphs from there. Keep in mind that the code was implied three separate times, as we divided the corpora into three sections, so the search for the common themes was also carried out three times (once for each section). The process included removing punctuation, numbers, and common stopwords (commonly used words that may not add significant semantic value to the text). First, we tried to make our own list of stopwords, but later on realized it is better to use a pre-set collection of stopwords integrated in the Python environment because our own set was too large and made it more difficult to use the dataset for further analysis. We also put the words into their basic form (lemmatization). Then we defined a term-frequency-inverse document frequency (TF-IDF), which helped us determine the importance of specific words in the document. We decided to use K-means clustering, to divide the documents into 4 clusters with the help of previously calculated TF-IDF measure. For better visualization, see the analysis of one cluster below. The specific words that appeared in said cluster, show that Cluster 2 focused on the topic of sports.

Cluster 2: *nizozemci bežijo reprezentance amsterdam mesecev nizozemski hitrostni zoi v naganu osvojili*

2. **Paragraph selection:** We decided it is best to look through all of the clusters manually and choose 13 most representative paragraphs from each of the clusters, that we will later on, in phase 3, pair with a corresponding paragraph that they will be in clear relation with (entailment, contradiction, neutral). Each member of the team will produce 50 samples, each sample containing two paragraphs (each approximately 5 sentences long). Those corresponding paragraphs will be generated through the usage of creative prompts that will then be fed to the large language model.

3. **Designing prompts for LLMs:** For each selected paragraph, we will craft tailored prompts designed to guide LLMs in generating hypotheses that are either entailed by, neutral to, or contradict the paragraph's content. This creative step is vital for ensuring the generated hypotheses are relevant and diverse. We will then employ ChatGPT-4 to produce hypotheses. See example for creating a contradictory paragraph to one of the examples below:

"Napiši paragraf, ki opisuje, kako je avantgarda šestdesetih let močno cenila zgodovinske umetniške prakse in kako je ta zavezanost preteklosti vplivala na naslednjo generacijo mladih komponistov. Opisujte, kako so ti mladi umetniki našli navdih v tradicionalnih tehnikah in zvočnih materialih, ki so jih raziskovali njihovi predhodniki, kar je privedlo do renesanse klasičnih stilov v sedemdesetih letih. Poudarite, kako je ta kontinuiteta tradicije in spoštovanje preteklosti privedlo do večjega zanimanja in vrednotenja klasične glasbe med mlajšimi generacijami, kar je postopoma omililo potrebo po nenehnem iskanju novih tehnik ali radikalnih idej."

4. **Manual review and annotation:** Each NLI pair will undergo a manual review process to validate the logical relationship between the paragraph and the hypothesis. This step ensures the accuracy and reliability of the dataset, with multiple reviewers involved to mitigate subjectivity.

5. **Dataset Compilation:** After review and annotation, the pairs are compiled into a comprehensive dataset. We strive for a balanced representation of logical relationships, ensuring the dataset's utility across various NLP applications.

7. **Preliminary Evaluation:** Finally, we conduct an assessment of the dataset's effectiveness by training a baseline NLI model, ChatGPT-4. Using this particular model seems reasonable, as it can generate more complex and nuanced results, distinguishing our dataset assessment with a higher level of analytical depth not as readily achievable with the more widely available ChatGPT-3. This evaluation phase is crucial for identifying any areas for refinement, guaranteeing that the dataset serves as a robust tool for advancing NLP research in Slovenian.

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 1, is an example of a figure that spans only across one of the two columns in the report.

Figure 1. A random visualization. This is an example of a figure that spans only across one of the two columns.

On the other hand, Figure 2 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

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. 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

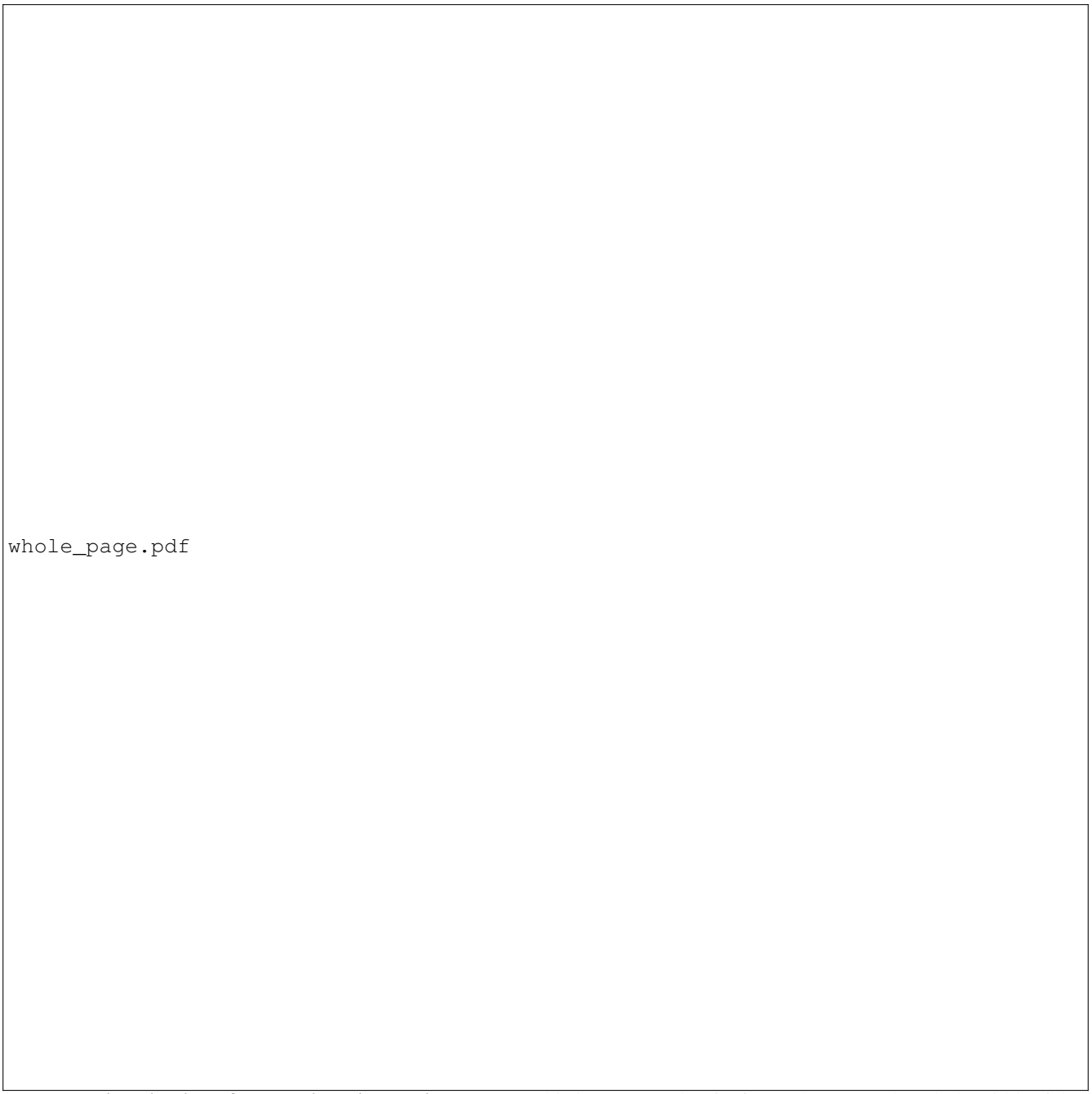

Figure 2. Visualization of a Bayesian hierarchical model. This is an example of a figure that spans the whole width of the report.