Natural Language Inference Dataset

**Group name:** NLP Hitchhikers

**Group members:** Maja Blažič, Veronika Durn, Ela Novak

**GitHub repository:** <https://github.com/UL-FRI-NLP-2023-2024/ul-fri-nlp-course-project-nlp-hitchhikers>

**Introduction**: Natural language inference (NLI) is a fundamental task in the field of natural language processing (NLP), enabling the ability to understand and infer relationships between texts. The goal of the project is to produce paragraphs of texts that test the understanding of involvement, neutrality, and contradictions in longer texts. Each team member will select 50 text samples from a corpus. To generate two-paragraph text (each containing approximately five sentences) we will use LLMs. After the generation of texts we will perform a manual validation to ensure the accuracy of the logical relationships and correct the mistakes within each sample. The validated samples will be combined into one large dataframe.By training and using a small model we will determine if the created dataset is challenging enough.

**Keywords:** NLP, NLI, LLMs

**Existing solutions and related work**: In the development of a Natural Language Inference (NLI) dataset for Slovene or English, we plan to utilise the insights and methodologies described in the provided links below. These documents offer foundational knowledge on enhancing the robustness of machine learning models, the importance of interpretable and reliable explanations, and innovative approaches to data generation and manipulation. Employing these concepts will ensure the creation of a quality unbiased NLI dataset, contributing to the advancement of Slovene or English language processing by incorporating state-of-the-art techniques in dataset creation and model evaluation. This approach underscores the commitment to fostering trust, transparency and ethical considerations in AI practices.

* <https://physionet.org/content/mednli/1.0.0/>
* <https://arxiv.org/abs/2106.09449>
* <https://www.researchgate.net/publication/347837670_The_ssj500k_Training_Corpus_for_Slovene_Language_Processing>
* <https://www.frontiersin.org/articles/10.3389/frai.2023.932519/full>
* <https://www.researchgate.net/publication/376202165_XL-WA_a_Gold_Evaluation_Benchmark_for_Word_Alignment_in_14_Language_Pairs>
* <https://www.researchgate.net/publication/373888629_Spremljevalni_korpus_Trendi_in_avtomatska_kategorizacija>
* <https://www.researchgate.net/publication/347837670_The_ssj500k_Training_Corpus_for_Slovene_Language_Processing>
* <https://www.researchgate.net/publication/366159461_Preventing_deception_with_explanation_methods_using_focused_sampling>

**Initial ideas**:

1. Find and choose a source from which we extract the samples
2. Each team member chooses 50 prompt samples
3. LLM text generation
4. Manually annotate and validate the generated texts (entailment, neutrality or contradiction) and correct mistakes
5. Combine the samples into one dataframe
6. Train a small language model
7. Report the findings

**References and resources:**

* Yu, F., Zhang, H., & Wang, B. (2023). Nature language reasoning, a survey. arXiv preprint arXiv:2303.14725.
* Bowman, S. R., Angeli, G., Potts, C., & Manning, C. D. (2015). A large annotated corpus for learning natural language inference. arXiv preprint arXiv:1508.05326.
* <https://paperswithcode.com/task/natural-language-inference>
* <http://nlpprogress.com/english/natural_language_inference.html>
* <https://www.clarin.si/repository/xmlui/handle/11356/1707>
* <https://www.kaggle.com/datasets/stanfordu/stanford-natural-language-inference-corpus>
* <https://aclanthology.org/D18-1007.pdf>
* <https://d2l.ai/chapter_natural-language-processing-applications/natural-language-inference-and-dataset.html>
* <https://paperswithcode.com/datasets?task=natural-language-inference>
* <https://www.sciencedirect.com/science/article/abs/pii/S0885230816302054>
* <https://aclanthology.org/2020.calcs-1.2.pdf>