Natural Language Processing Projects

Winter semester 2022

Below, you can find concise descriptions of NLP projects. They show the general idea of the project. We would like to encourage you to elaborate on your research project proposals. If you don't like any of the following projects, don't worry and suggest your ideas. Maybe you can also look into the latest text contest, e.g. SemEval, PolEval, or go through my preprints then I will be able to help more: arxiv preprints by Anna

Good luck
Anna

All projects must have the same architecture stages:

1. **Research proposal (report - overleaf + PDF and presentation)**

The proposal is to focus on a description and plan of the solution and its justification - state-of-the-art (SOTA) review: literature analysis, datasets and ML methods review.

Template for the proposal is under the link:
https://www.overleaf.com/4925917446vwsyhvsjnfcc

2. **Proof of Concept - data analysis & ML (MR and presentation)**

The PoC part focuses on dataset gathering, exploratory data analysis (EDA) and initial modelling - building an initial system and commenting on its performance. EDA should include data structure discovery, and outlier/anomaly detection using visualisation and summarisation techniques. The Analysis part should consist of code (can be Jupiter notebook) and a report that gathers all of your current findings (datasets overview, and ML part). The task is not to create the best available model but to build a well-thought-out solution, taking into account the current SOTA.

3. **Final report and modelling: (presentation + poster + MR, after a week: final report regarding also remarks given in the presentation)**

The last part is to prepare the final solution (also with more profound analysis, e.g. interpretability techniques) and report that discusses how the proposed system performs, including the advantages and disadvantages of the solutions as well as possible improvements or architectures that could address the given weaknesses - this will be the scope of the second project.

Additional points (5 points)

Additional points are provided for teams that propose how to deploy the solutions for scalable usage in production. The deployment should be given along with the text discussion and deployment files, e.g. Dockerfile, JSON/YAML deployment definition etc., with the example usage and results.

The recommended solutions should meet the following objectives:

- Docker image(s) smaller than 4 GB
- The solution runs on CPU faster than 1 second per watch/request (performance test attached)
- CD4ML/MLOps best practices considered - Implementation using Kubernetes - Clean, understandable code with comments where necessary
- Assumptions and their justification
- Appropriate model validation
- Creative approach to modelling
- The balance between simplicity and performance

Project – few-shot learning

Model training and performance on small NLP datasets

Description:

Deep learning models often require a large amount of data to train or fine-tune. Unfortunately, we often struggle with the problem of undersized samples in the real world. This project aims to comprehensively analyse the techniques that can improve models' quality in a few-shot learning scenario. In this project, you should explore semi-supervised learning methods or different model architectures or learning scenarios (e.g. contrastive learning). This machine-learning approach combines a small amount of labelled data with a lathe amount of unlabeled data during training.

Datasets:

- PyTorch datasets such as AG News and IMDb (https://pytorch.org/text/stable/datasets.html)
- Kaggle datasets https://www.kaggle.com
- other…

Useful links and papers:

https://github.com/makcedward/nlpaug
https://arxiv.org/pdf/2104.08821.pdf
https://arxiv.org/abs/2108.11458
https://arxiv.org/abs/2112.08462
https://arxiv.org/pdf/2011.00362.pdf - survey

Project - spoilers

Spoiler detection and extraction

Description:

We collected a dataset for a spoiler classification (detection) task with annotated spoiler-responsible tokens. This project aims to build a model that can detect texts containing spoilers. Additionally, the team should analyse if the interpretability tools indicated the exact works/phrases as people annotated based on the dataset. Apart from that, it would be valuable to conduct research on other valid data and methods to explore the topic of interpretability.

Datasets and useful links and papers:

- https://arxiv.org/abs/2112.12913
- https://www.eraserbenchmark.com,
- https://arxiv.org/pdf/1911.03429.pdf

Project – SLR - semantic keywords for systematic literature reviews

Topic Modeling & Semantic Tagging

Description:

The goal of the task is to build a system that can find keywords in given longer texts and tag them with a given ontology or/and dictionary.
One approach might be to compute topic representations for the entire dataset (in the form of sentence embedding clusters) and then calculate the number of sentences belonging to a given group for data (BERTopic solution). This approach could include using pre-trained and fine-tuned language models for the selected corpus. However, the minimum solution (PoC) can also be based on the Latent Dirichlet Allocation (LDA) algorithm and NCBO annotator.

Datasets:

Dataset for systematic literature reviews or any other documents (abstracts)
https://academic.oup.com/nar/article/49/D1/D1534/5964074?login=false

Useful links and papers:

https://maartengr.github.io/BERTopic/index.html
https://github.com/MaartenGr/BERTopic

https://arxiv.org/pdf/1902.09476.pdf - MEDmentions – annotation with UMLS

Raymon van Dinter, Bedir Tekinerdogan, Cagatay Catal, Automation of systematic literature reviews: A systematic literature review,Information and Software Technology 2021

https://www.vosviewer.com/ - VoS viewer

https://towardsdatascience.com/medcat-introduction-analyzing-electronic-health-records-e1c420afa13a

https://www.frontiersin.org/articles/10.3389/fphar.2022.812338/full - ML based solution

https://datalanguage.com/blog/pico-cochrane-and-schema-org

https://data.cochrane.org/ontologies/ - ontology to systematize the research question

Project - recipes

Recipes data extraction

versions for Polish, English, maybe multilingual

Description:

The task aims to enrich the RecipeNLG dataset with additional nutritional facts and dietary tags. You can compare classification methods for predicting dietary tags, gather more nutritional values, or extract crucial information about values and units of ingredients and gather more nutritional and dietary data.

This approach could include pre-trained and fine-tuned language models for the selected corpus of recipes. To enrich the dataset, you can use the NCBO annotator to tag the recipes with the FoodOn ontology concept.

Datasets:

https://recipenlg.cs.put.poznan.pl/
https://paperswithcode.com/dataset/tasteset

Useful links and papers:

multilabel recipe classification -
https://sol.sbc.org.br/index.php/eniac/article/view/12133/11998
https://bioportal.bioontology.org/annotator
https://foodon.org/

Project - recipe reviews

Recipe reviews

versions for Polish, English, maybe multilingual

Description:

The task aims to enrich, e.g. the RecipeNLG dataset with reviews and extract attributes that are assessed in the Internet reviews. You can find out how to extract particular recipe aspects of each review and overall assessment of the recipe, and suggestions to improve the recipe often given in reviews.

Datasets:

https://recipenlg.cs.put.poznan.pl/
https://paperswithcode.com/dataset/tasteset
also datasets on sentiment analysis: http://nlpprogress.com/english/sentiment_analysis.html

Project - e-commerce products

Taxonomy, similarity between products

Description:

The task aims to create automatic methods for measuring similarity between products on multilevel dimensions, i.e. building/enriching a taxonomy (categories), extracting crucial information from description and titles, experimenting with generating product attributes (tags) based on product images. This task aims to aid e-commerce search for products and help in loading data on a new product into e-commerce platforms, and finding available price range for similar products.

Datasets:

https://arxiv.org/abs/2205.15712
WDC dataset + German dataset

Project – fake news and misinformation

Description:

The task aims to create automatic methods for finding fake news and assessing the misinformation or information overload.

Useful links:

M. Choraś et al. Advanced Machine Learning techniques for fake news (online disinformation) detection: A systematic mapping study. Applied Soft Computing 2020. DOI: 10.1016/j.asoc.2020.107050. https://www.sciencedirect.com/science/article/pii/S1568494620309881

Islam, M.R., Liu, S., Wang, X. *et al.* Deep learning for misinformation detection on online social networks: a survey and new perspectives. *Soc. Netw. Anal. Min.* **10**, 82 (2020). DOI: 10.1007/s13278-020-00696-x. https://link.springer.com/article/10.1007/s13278-020-00696-x