# Reproducibility Appendix
# Project Report for NLP Course, Winter 2023/24

**Sebastian Deregowski, Dawid Janus, Bartosz Jamroży, Klaudia Gruszkowska**
Warsaw University of Technology
sebastian.deregowski.stud@pw.edu.pl
klaudia.gruszkowska.stud@pw.edu.pl
bartosz.jamrozy.stud@pw.edu.pl
dawid.janus.stud@pw.edu.pl

**supervisor: Anna Wróblewska**
Warsaw University of Technology
anna.wroblewska1@pw.edu.pl

## Reproducibility checklist

The description below applies to XLNet, the most promising model of all tested throughout the project.

Overall results:

- MODEL DESCRIPTION –
  XLNet is an advanced language model introduced by Google AI in 2019. Its main features are Permutation Language Modeling (PLM), which takes into account all permutations of words in a sentence, and the integration of autoregression and autoencoder. The model is based on the Transformer architecture, trained on large datasets and can be adapted to different tasks through fine-tuning.
- LINK TO CODE –

  Source code: https://github.com/grant-TraDA/NLP-2023W/tree/main/6.%20NER%20for%20acknowledgments

  External libraries: https://github.com/grant-TraDA/NLP-2023W/blob/main/6.%20NER%20for%20acknowledgments/final/requirements.txt
- INFRASTRUCTURE –
  The calculations were performed on an i5-13600kf processor with a clock speed of up to 5.1GHZ . It has 14 cores and 20 threads. The computer has 32GB of RAM.
- RUNTIME PARAMETERS –
  Approximately one hour per epoch.
- PARAMETERS –
  XLNet Large has approximately 340 million parameters.
- VALIDATION PERFORMANCE – XLnet Large without silver-set reached 0.7893 F1-score on the validation set and 0.7747 on the test set. XLnet Large with silver-set reached 0.8078 on the validation set and 0.7924 on the test set.
- METRICS –

$$F1 = \frac{2 \cdot precision \cdot recall}{precision + recall}$$

$$Precision = \frac{TruePositives}{TruePositives + FalsePositives}$$

$$Recall = \frac{TruePositives}{TruePositives + FalseNegatives}$$

  https://github.com/grant-TraDA/NLP-2023W/tree/main/6.%20NER%20for%20acknowledgments/final/training_outputs

Multiple Experiments:

- NO TRAINING EVAL RUNS –
  10 Epochs
- HYPER BOUND –
  Learning rate: The variable learning rate value was tested manually from the range: 0.1, 0.01, 0.001, for learning rate like 0.1, 0.001 the model did not learn. So finally the 0.01 learning rate was used.
  Batch size: Due to the limitation of the machine on which the models were trained, the bach size had to be set to a small value. Finally, the bach size was set to 4, as larger values caused an overflow of RAM on the machine(32GB) used for training.

- HYPER BEST CONFIG –
  Not used
- HYPER SEARCH –
  Not used
- HYPER METHOD –
  Not used
- EXPECTED PERF –

  Model results with respect to entity (F1 score):

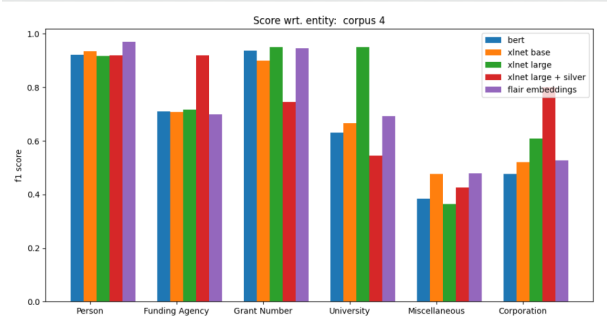


Figure 1: Comparison of F1 scores for each class and model

Datasets – utilized in the experiments and/or the created ones:

- DATA STATS –
  Sizes of each corpus are presented in the table 1.

| Corpus | Number of Samples |
|---|---|
| 1 | 1380 |
| 2 | 24,578 |
| 3 | 38,938 |
| 4 | 48,170 |

Table 1: Comparison of size of each corpus

- DATA SPLIT –
  Sizes of each corpus and their splits to train/test/dev sets are presented in the table 2.
- DATA PROCESSING –
  Data provided by Smirnova and Mayr has been already processed to the input suitable for model.
- DATA DOWNLOAD –
  `https://github.com/grant-TraDA/NLP-2023W/tree/main/6.%20NER%20for%20acknowledgments/final/data`

| Corpus | Train | Test | Val |
|---|---|---|---|
| 1 | 832 | 322 | 226 |
| 2 | 12,621 | 6,398 | 5,559 |
| 3 | 26,981 | 6,398 | 5,559 |
| 4 | 36,213 | 6,398 | 5,559 |

Table 2: Comparison of size of each corpus divided by train/test/dev splits

- NEW DATA DESCRIPTION –
  We prepared a silver standard set that allows automatic or semi-automatic ways to create labels, in our project labels of entities for text in acknowledgment sections. We manually collected acknowledgment text from 243 papers from the science domain (in total 20 773 samples). Next, we prepare predictions based on our best model at that time, which was the Flair Embedding model trained on corpus 4.
- DATA LANGUAGES –
  English