Multilingual NLP – Project Presentation

Cross-lingual transfer via fully fine-tuning and BitFit

Code Presentation - Datasets

- Base training/validation datasets
  - WikiAnn (en)
  - Conll2003
- Test datasets:
  - WikiAnn/Conll2003
  - Masakhaner → training/validation/test merged to one big test set
- Preprocessing:
  - SAME → B_x, B_x, ..
  - IGNORE → B_x, -, ..
  - INSIDE → B_x, I_x

Code Presentation - Datasets

- Dataloaders
  - Training → shuffling, simple one
  - Validation → no shuffling, simple one
  - Test → no shuffling, combined one (forced to use sequential)

Code Presentation - Model

- Model:
  - Xml-roberta-base w/o pooling layers
  - Single dense layer (d → number of NER tags)
- BitFit modification
  - Set “requires_grad” respectively for each parameter
    - Bias = True
    - Else = False
  - Only pass relevant parameters to optimizer
    - → AdamW influences parameters too

Code Presentation - Run

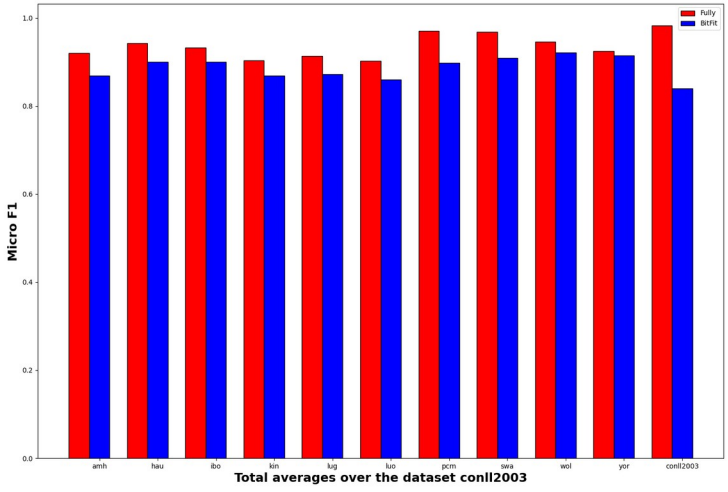
- Different configurations
  - Preprocessing
    - Same
    - Inside
    - Ignore
  - Batch size
    - 16
    - 32
    - 64
  - Base dataset
    - WikiAnn (en)
    - Conll2003

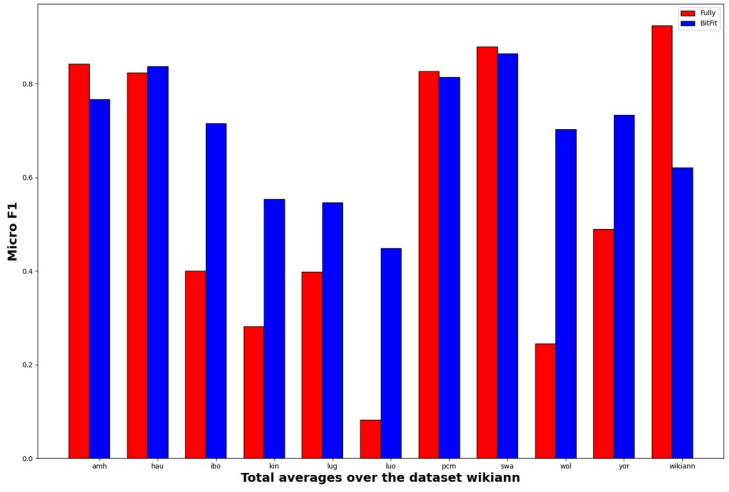
Code Presentation - Details

- GPU is assumed to be available
- Flag “save_dir”
  - Set save directory of runs and checkpoints
    (1.1GB BitFit / 3.3GB Fully)
- Uses .env
  - “CACHE_DIR” to set directory of huggingface cache
- Saving and reloading preprocessed datasets to/from disk
  - Due to certain issues in hf caching mechanisms
- 36(-3) runs in total
  - Plotted results via matplotlib + wandb api

Code Presentation - Results

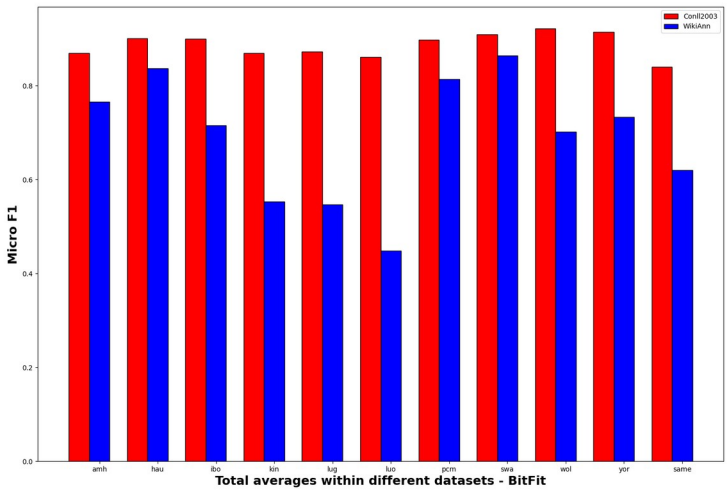
Datasets

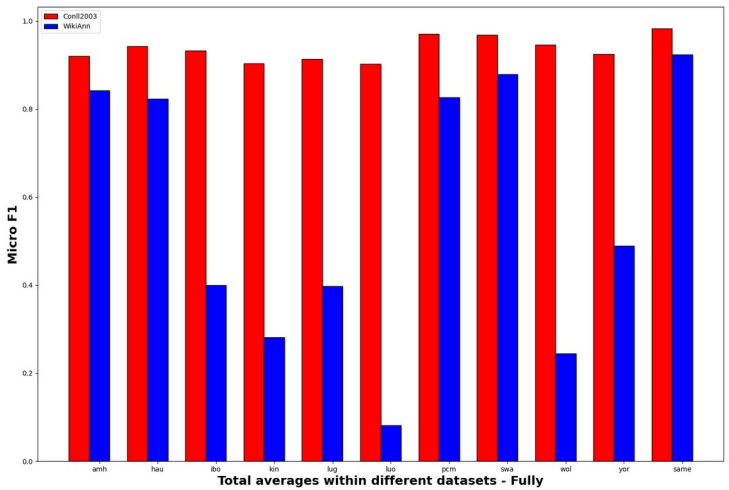




Code Presentation - Results

Datasets



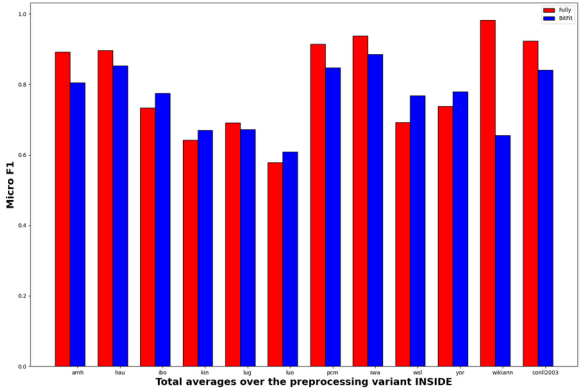


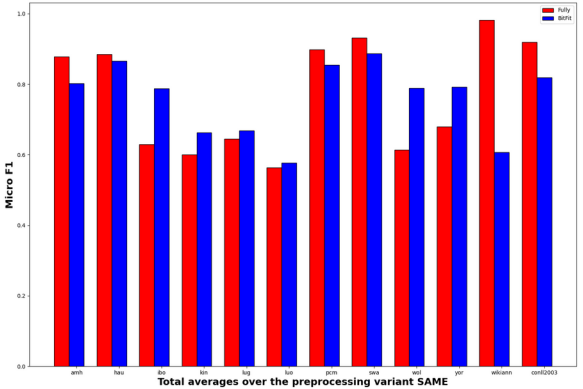
Code Presentation - Results

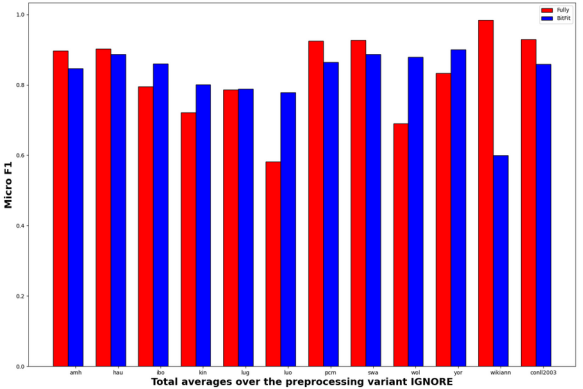
- Datasets
  - Fully adapts better to base dataset
  - Conll2003 > WikiAnn !!!
    - Huge drop in performance
  - Conll2003 → Fully, WikiAnn → BitFit

Code Presentation - Results

Preprocessing

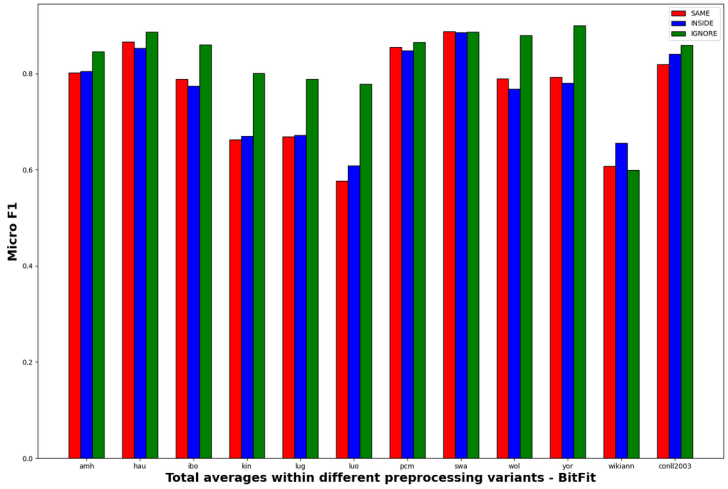


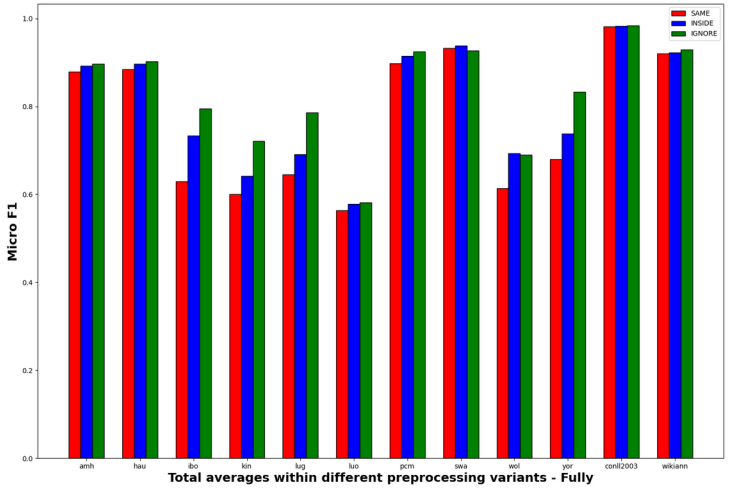




Code Presentation - Results

Preprocessing



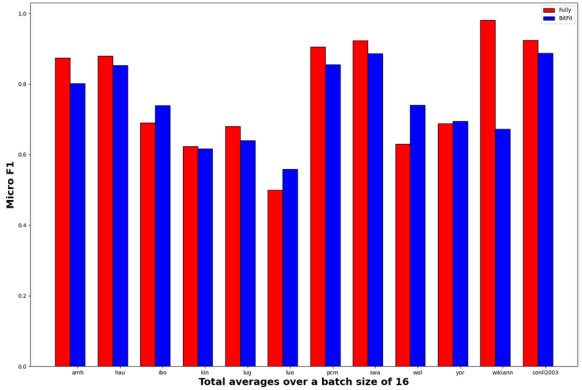


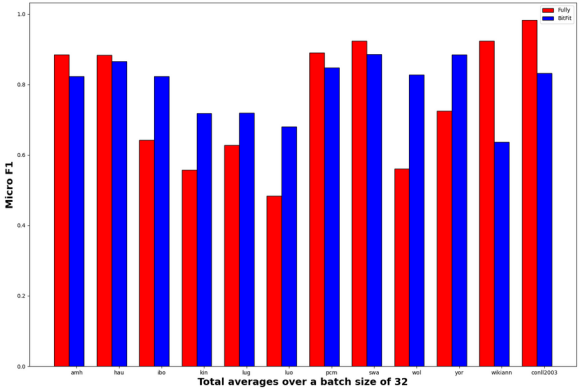
Code Presentation - Results

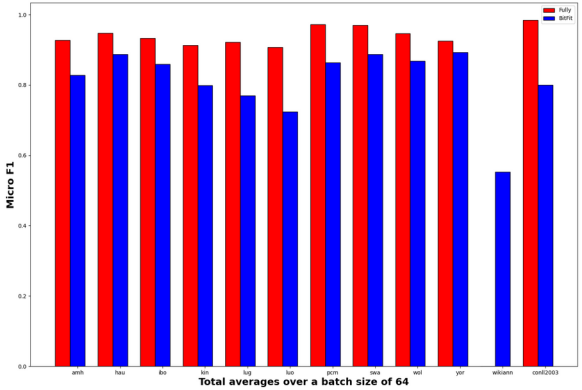
- Preprocessing:
  - “Smaller” changes in performance
  - Ignore >> Inside > Same
    - Might be due to bias to non entities

Code Presentation - Results

Batches

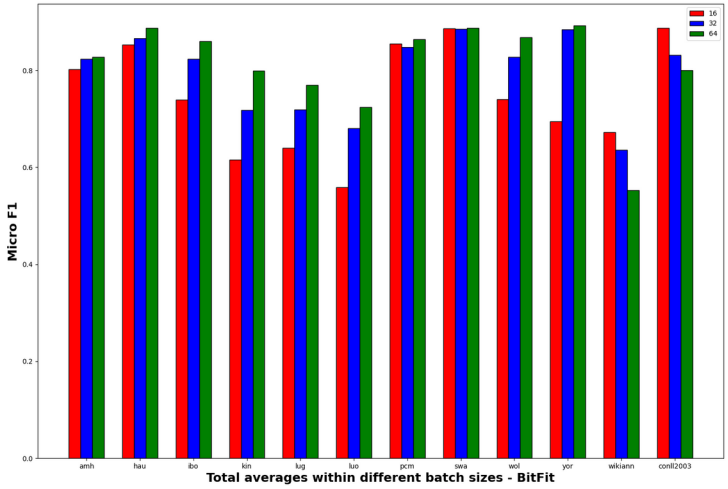


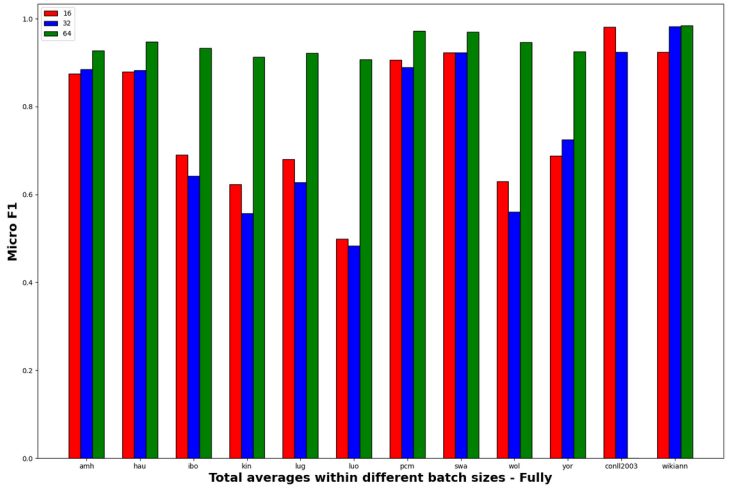




Code Presentation - Results

Batches



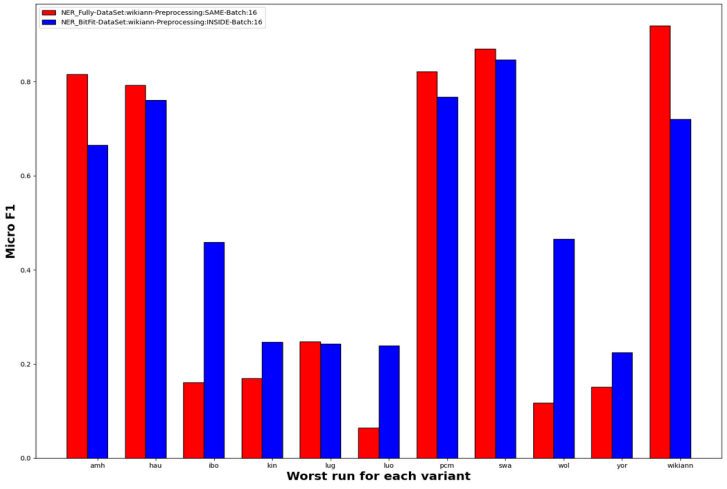


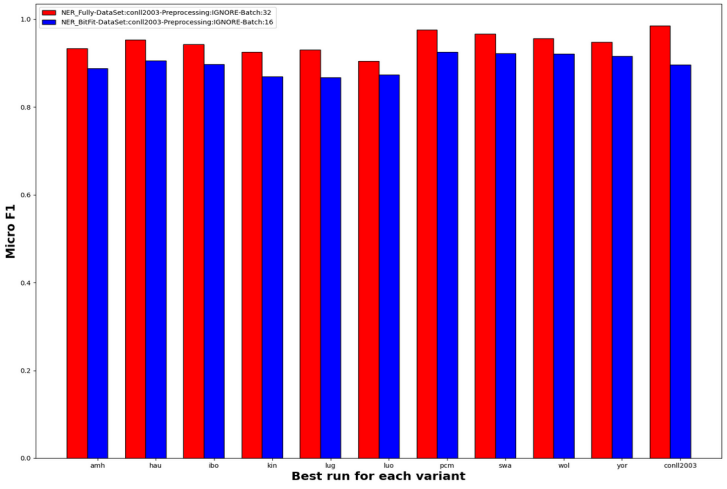
Code Presentation - Results

- Batches
  - 64 for WikiAnn not available (16GB VRAM were not enough)
    - Skewed plots
  - BitFit:
    - 64 > 32 > 16
  - Fully:
    - Indications of 16 > 32 > 64

Code Presentation - Results

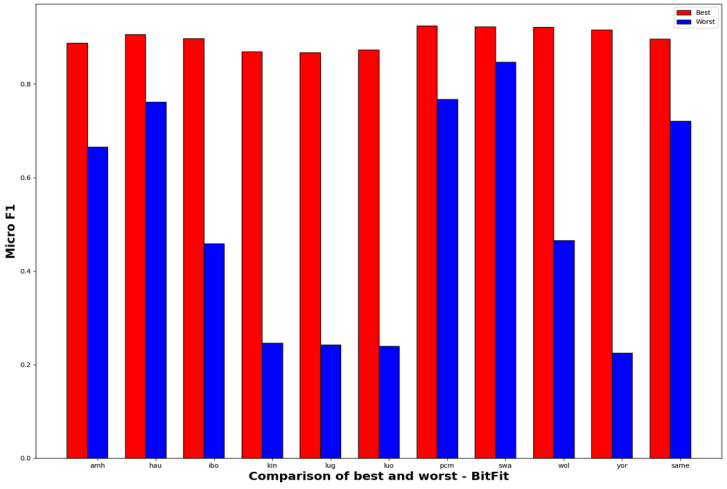
Best and Worst

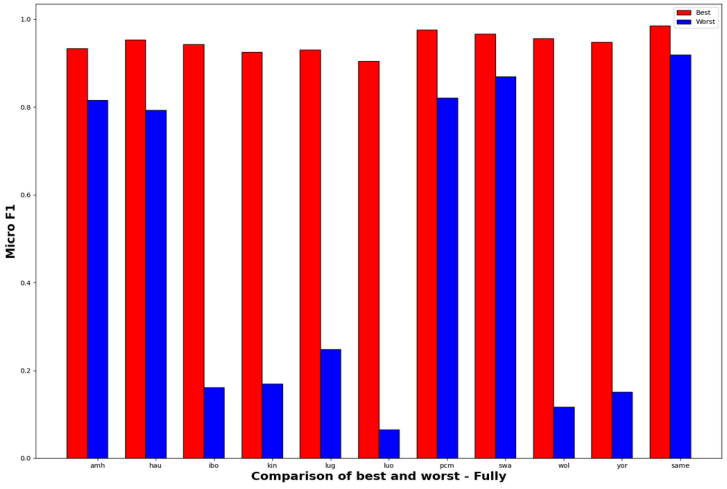




Code Presentation - Results

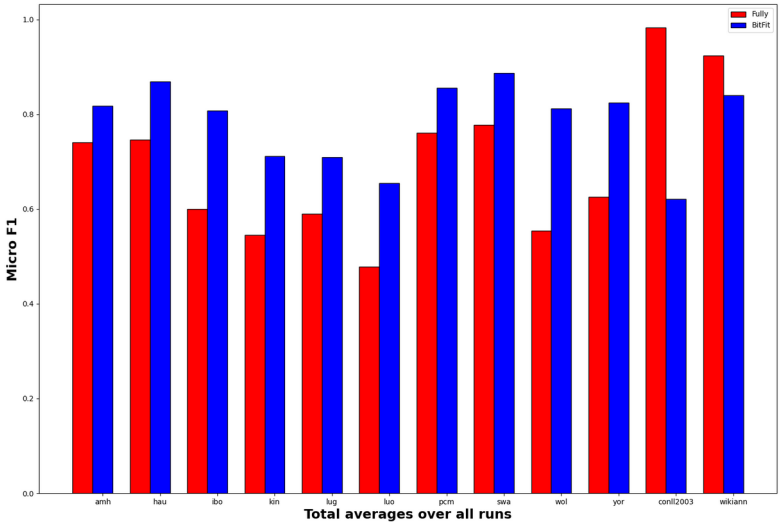
Best and Worst





Code Presentation - Results

Average



Code Presentation - Results

- Fully
  - With right configuration is the best
    - Conll2003, Ignore, 32
  - But also the worst
    - WikiAnn, Same, 16
  - Also, higher discrepancy between worst and best
    - Due to WikiAnn
- BitFit
  - Comparably good results
  - More resistant to base dataset

Thanks for listening !!!

Quality matters!

and

BitFit is decent at worst!