Smilla Fox*, Martin Preiß*,
Florian Borchert, Aadil Rasheed, Matthieu-P. Schapranow
Hasso Plattner Institute, University of Potsdam, Germany
(*equal contribution)

Data Augmentation and Ensemble Learning for Multilingual Adverse Drug Event Detection

- Boosted overall accuracy of results with ensemble learning
- Improvement of low-frequency classes with data augmentation
- Second-best results in NTCIR-17 SM-ADE subtask across all languages

1. Task

- Datasets consisting of 7964 tweets each in Japanese, English, French or German
- Multi-label classification with 22 ADE classes

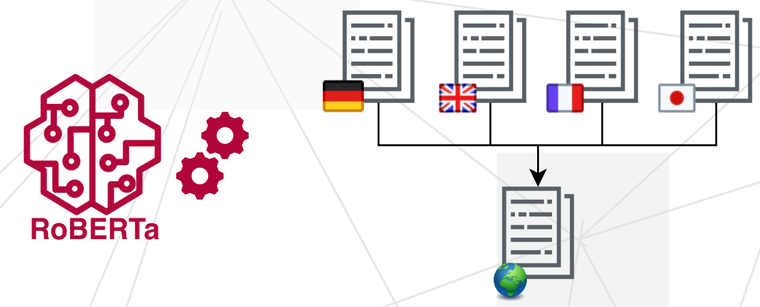
> *"I'm having **abdominal pain**, **nausea**, and **headaches** that I think are side effects of mesalazine..."*
>
> ADE-Labels: **Nausea, Abdominal Pain, Headache**

> *"I have a **stomach ache**. Reduced to 10mg prednisone and 40mg mesalazine today."*
>
> ADE-Labels: **None**

2. Methods

Baseline

- Finetune multi-lingual XLM-RoBERTa model encoder with multi-label classification head
- Hyperparameter sweep over learning rate, batch size, weight decay, decision threshold, number of training epochs

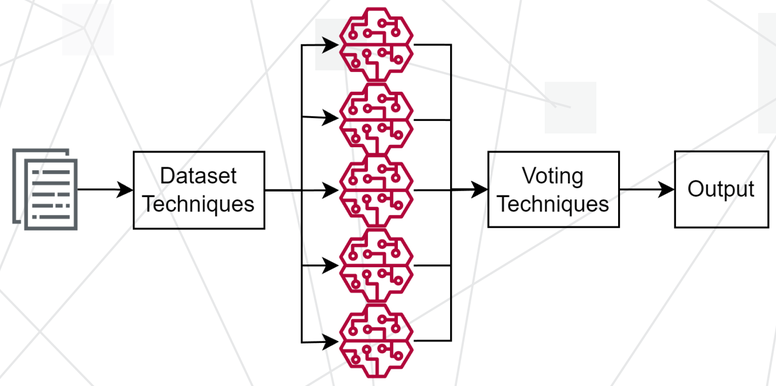


Finetune multi-lingual XLM-RoBERTa models on datasets in individual languages and multi-lingual dataset

Data Augmentation

- Generate artificial tweets using instruction-tuned GPT-3.5-turbo model through OpenAI API
- Prompt engineering
- For low-frequency classes → *Bone Marrow Dysfunction*, *Interstitial Lung Disease*, *Pain*

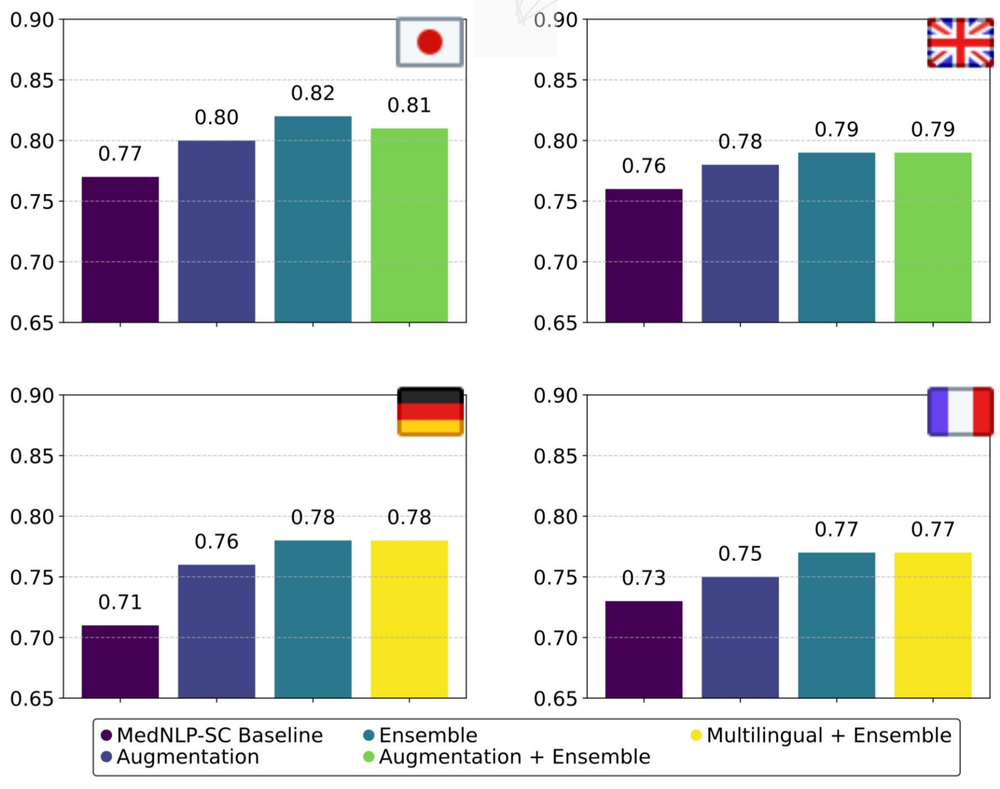
Ensemble Learning



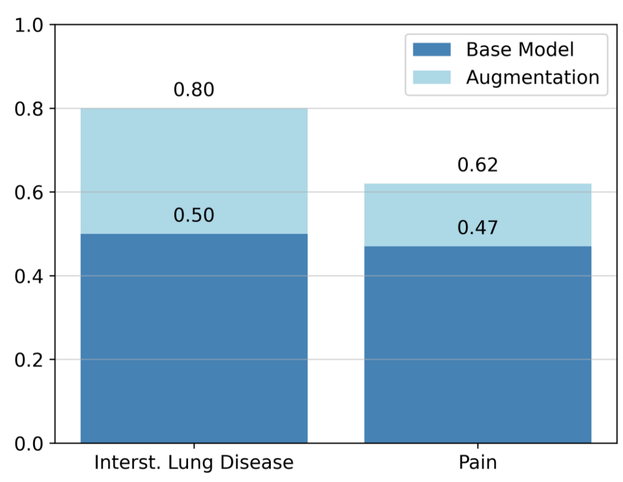
- Train multiple base learners
- Dataset Techniques: Kfold, Shuffle
- Voting Techniques: Majority Vote, Average Probability Vote, Weighted Vote

3. Results

- All our methods outperform MedNLP SC-Baseline (best run of task paper)
- Ensembles perform best
- Multi-lingual dataset + ensemble better for German and French
- Augmentation on mono-lingual dataset + ensemble better for Japanese and English
- Best overall results on Japanese data: original dataset



Micro-averaged F1 scores of 3 submitted runs vs. MedNLP SC-Baseline for individual languages



F1 scores of rare symptoms: Performance improvement through data augmentation