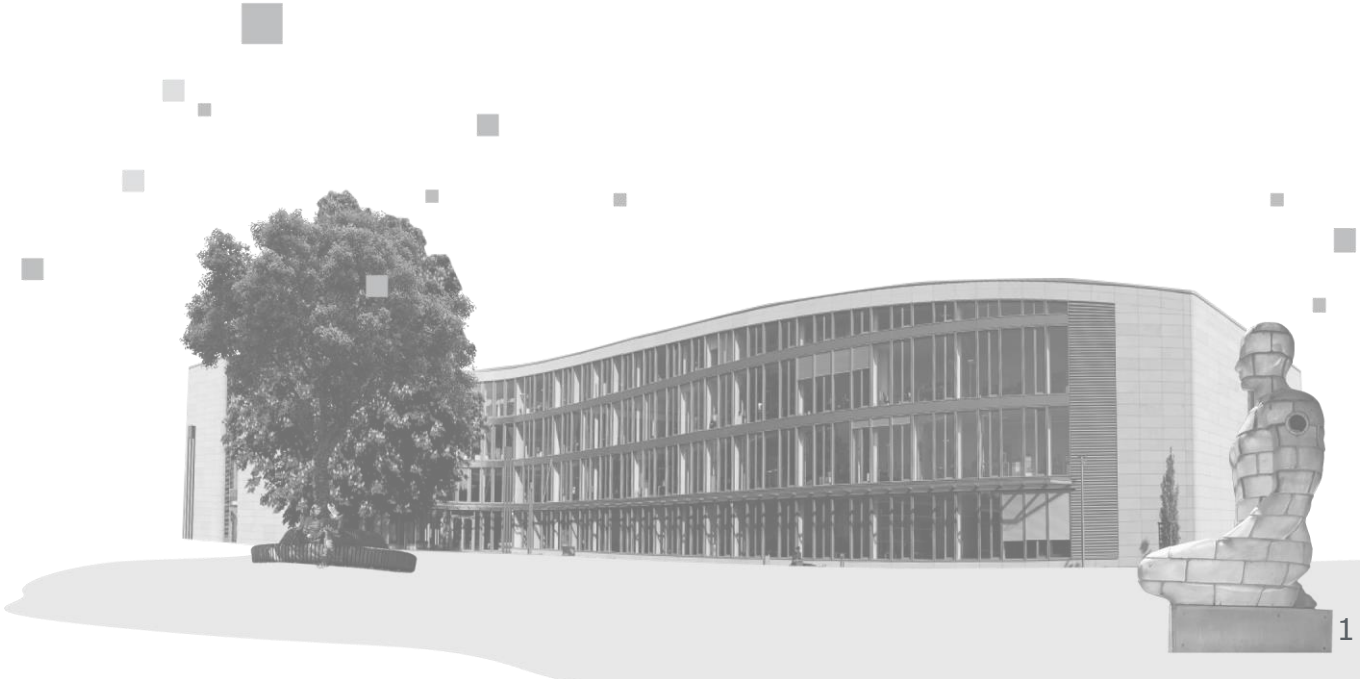
Hasso Plattner Institut
Digital Engineering · Universität Potsdam

NTCIR-17 MedNLP-SC

Data Augmentation and Ensemble Learning for Multilingual Adverse Drug Event Detection

Smilla Fox, Martin Preiß, Florian Borchert, Aadil Rasheed, Matthieu-P. Schapranow
Hasso Plattner Institute for Digital Engineering, University of Potsdam, Germany



Design IT.
Create Knowledge.

www.hpi.de

Task

Adverse Drug Event (ADE) Detection

- Input: Text
- Output: Labels for 22 symptoms
- Task: Multi-label classification

Dataset

- 7964 Generated social media texts
- Available in Japanese, English, German and French

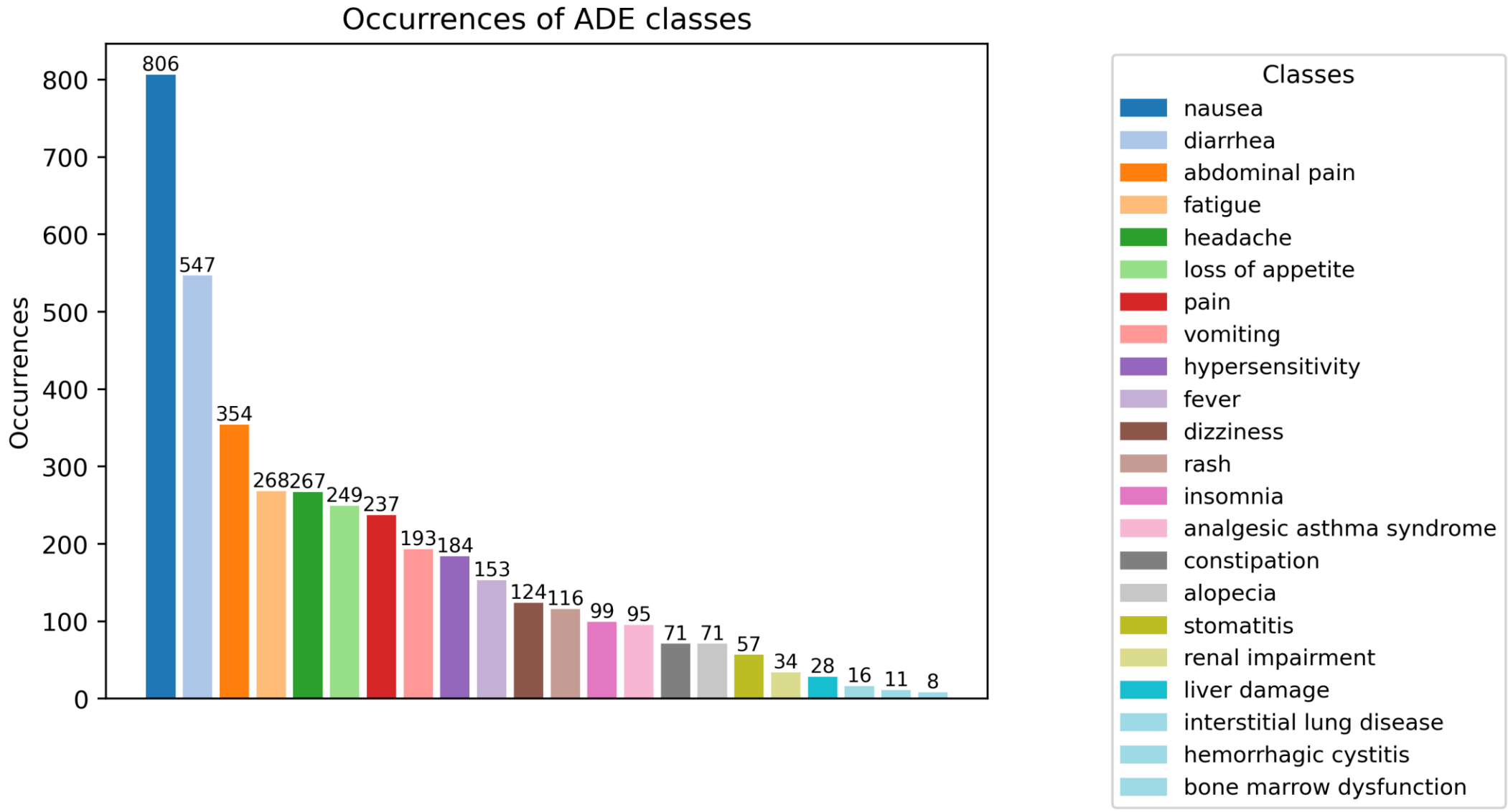
*"I'm having **abdominal pain**, **nausea**, and **headaches** that I think are side effects of mesalazine..."*

ADE Labels: **Nausea, Abdominal Pain, Headache**

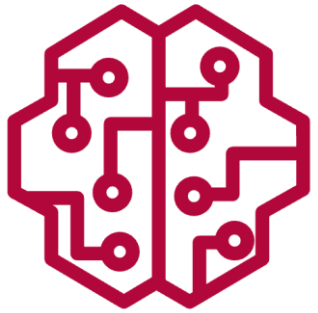
Not all symptoms are ADEs.

*"I have a **Stomach Ache**. Reduced to 10mg prednisone and 40mg mesalazine today."*

ADE Labels: **None**

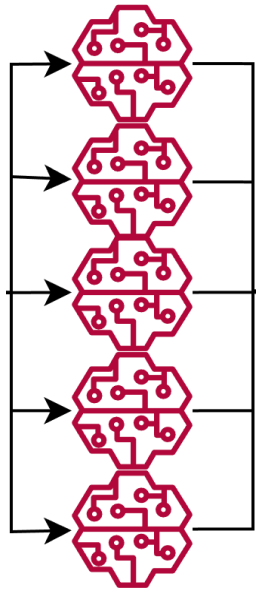

Occurrences of ADE classes
Occurrences
0
100
200
300
400
500
600
700
800
806
547
354
268
267
249
237
193
184
153
124
116
99
95
71
71
57
34
28
16
11
8
Classes
nausea
diarrhea
abdominal pain
fatigue
headache
loss of appetite
pain
vomiting
hypersensitivity
fever
dizziness
rash
insomnia
analgesic asthma syndrome
constipation
alopecia
stomatitis
renal impairment
liver damage
interstitial lung disease
hemorrhagic cystitis
bone marrow dysfunction

Methods



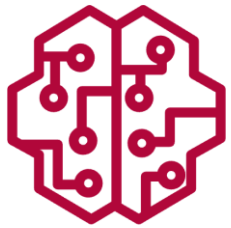
1. Base Model

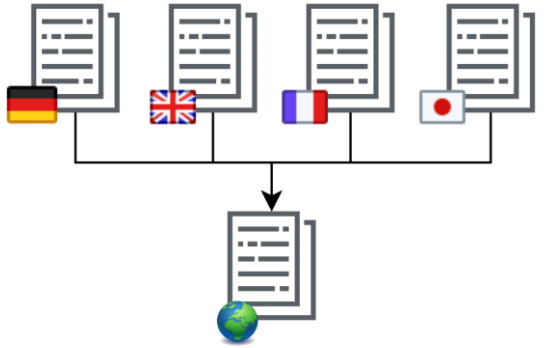
2. Data Augmentation



3. Ensemble Learning

1. Base Model





- Finetune multi-lingual XLM-RoBERTa model from Huggingface
- Hyperparameter tuning

- Multi-lingual approach: finetune on combined language datasets

2. Data Augmentation

- Generate artificial tweets using GPT-3.5-turbo model through OpenAI API
- Find appropriate prompt
- Use for low-frequency classes → *Bone Marrow Dysfunction*, *Interstitial Lung Disease*, *Pain*
- Generate in German, translate into other languages

Prompt:

"Generate 20 tweets that meet all of the following conditions:
The person writing talks about pain as a side effect of a medication they have been taking.
The specific name of the drug is mentioned.
The placeholder **[drug name]** is used for this.
No other side effects besides {symptom} are mentioned.
In a few sentences, it is mentioned what the medication was taken for."

Example for generated tweet:

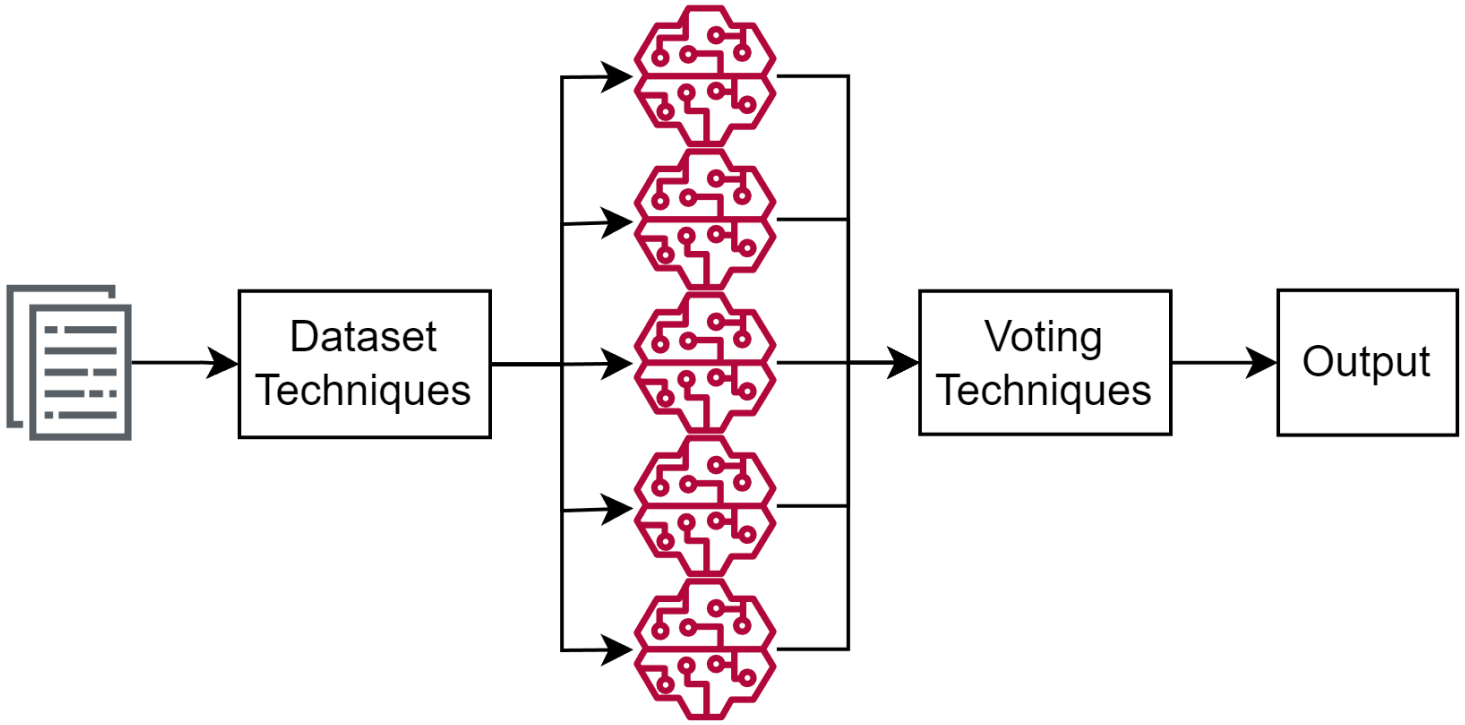
"My wrists are hurting after taking [drug name]. It's challenging to perform tasks that require wrist movements."

3. Ensemble Learning

Improves accuracy / robustness

Setup:

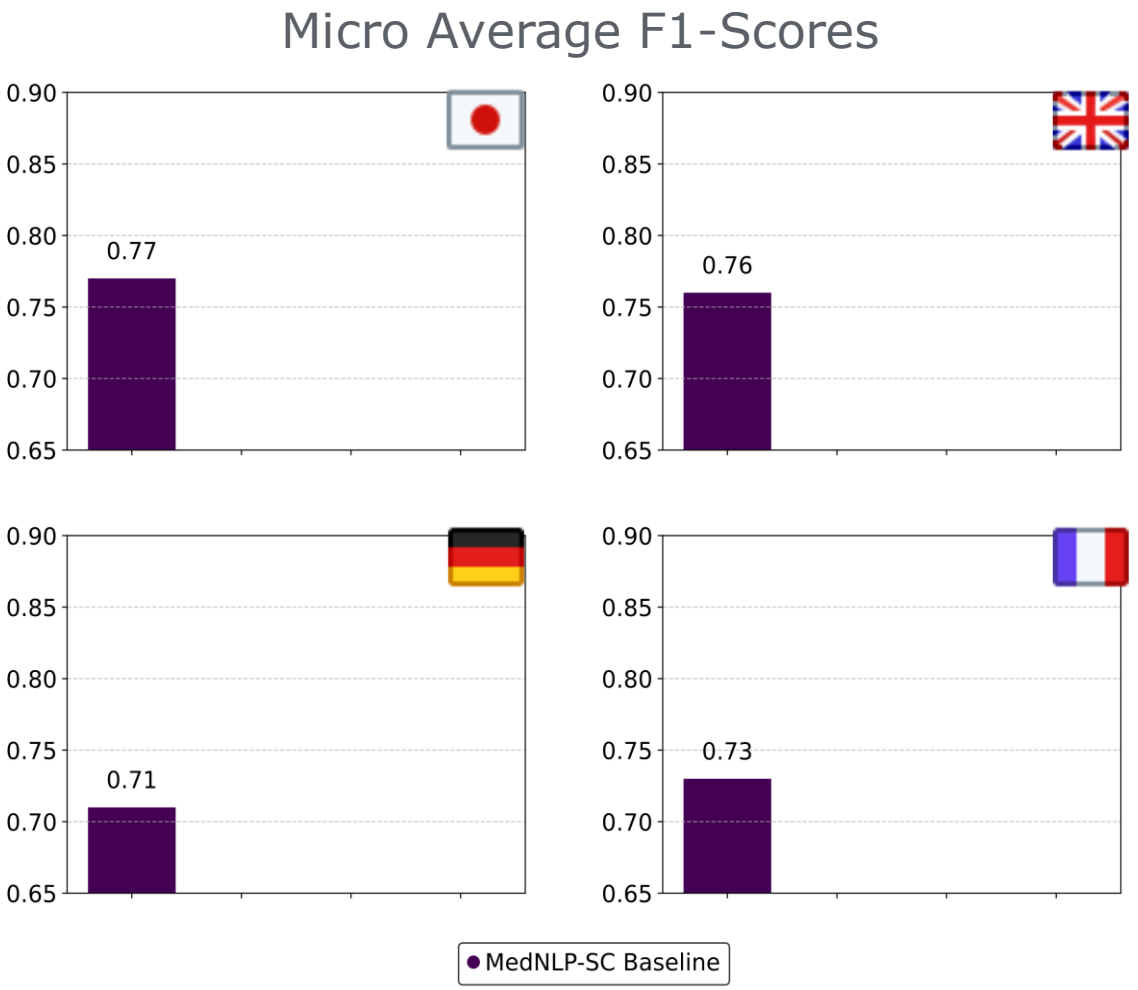
- Transform data with dataset techniques
- Finetune 5 models independently
- Combine outputs with voting techniques



Results

Compare to MedNLP-SC Baseline:

- XLM-R_{all} run of subtask overview



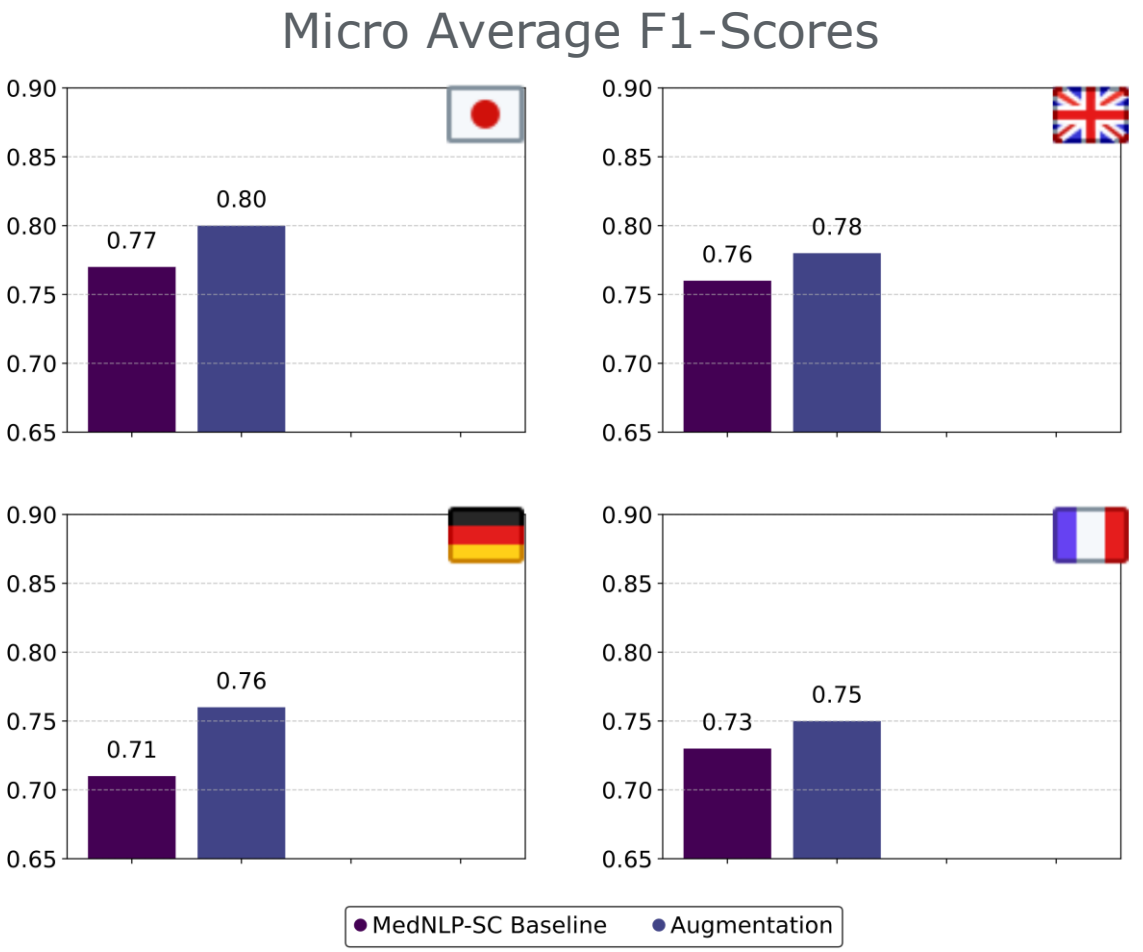
Results

Compare to MedNLP-SC Baseline:

- XLM-R_{all} run of subtask overview

1. Submission:

- Augmentation

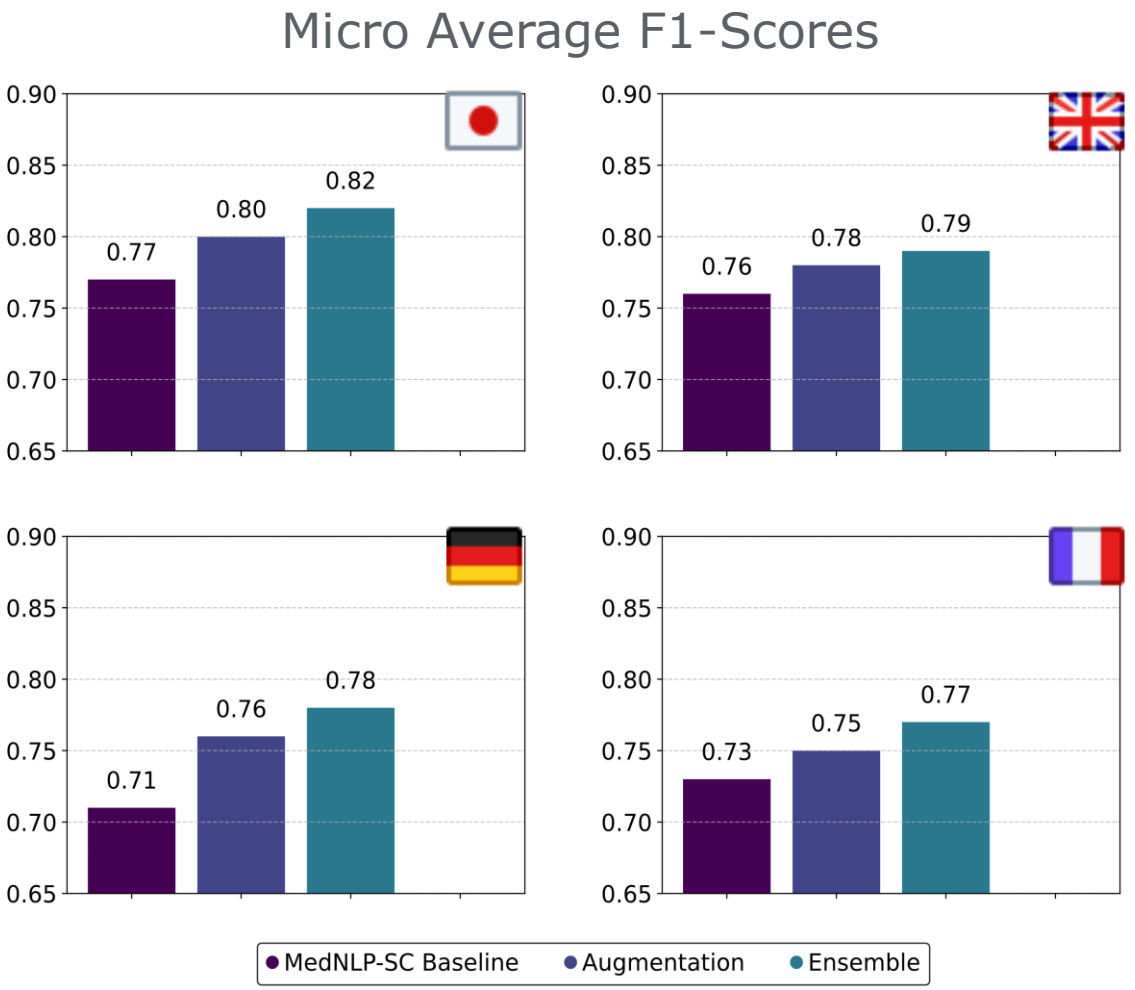


Results

Compare to MedNLP-SC Baseline:

- XLM-R_{all} run of subtask overview

1. Submission:
   - Augmentation
2. Submission:
   - Ensemble

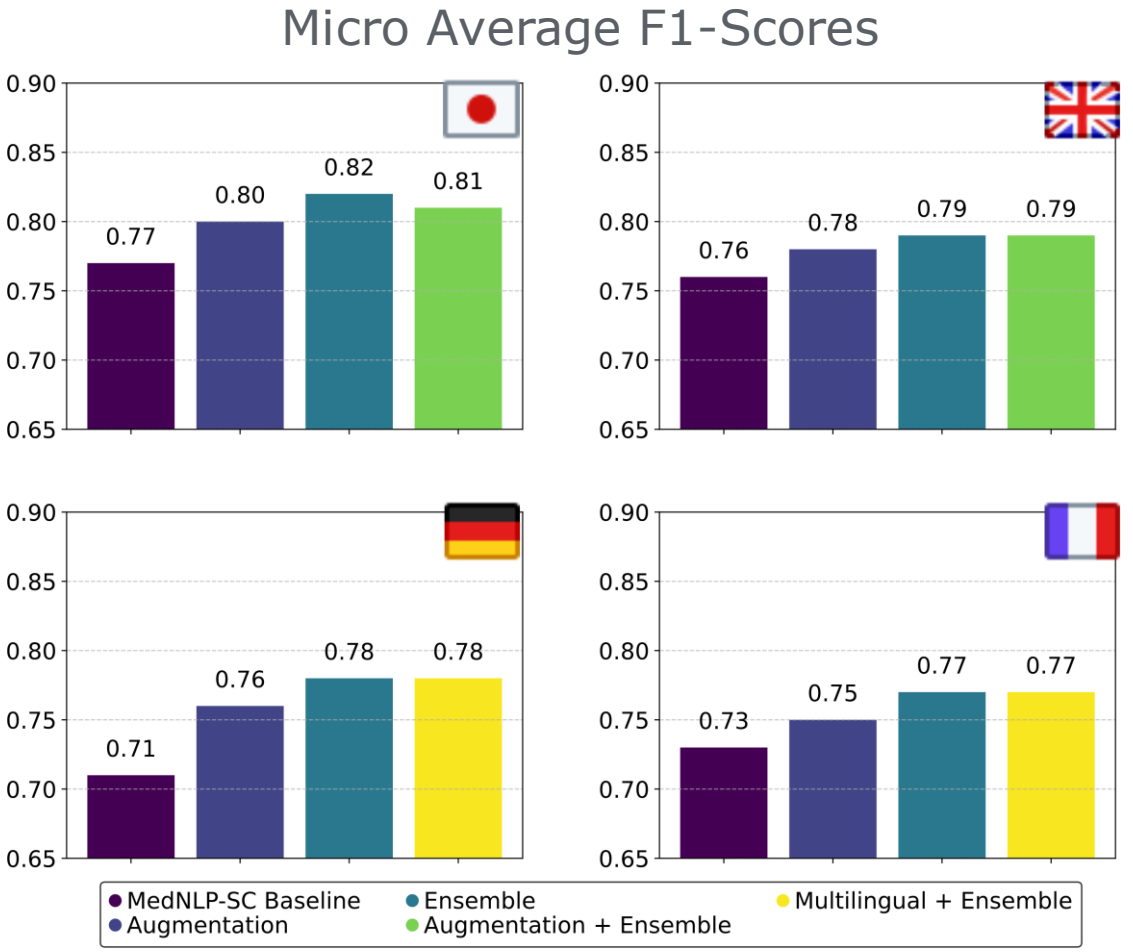


Results

Compare to MedNLP-SC Baseline:

- XLM-R_{all} run of subtask overview

1. Submission:
   - Augmentation
2. Submission:
   - Ensemble
3. Submission
   - Multi-lingual/Augmentation + Ensemble



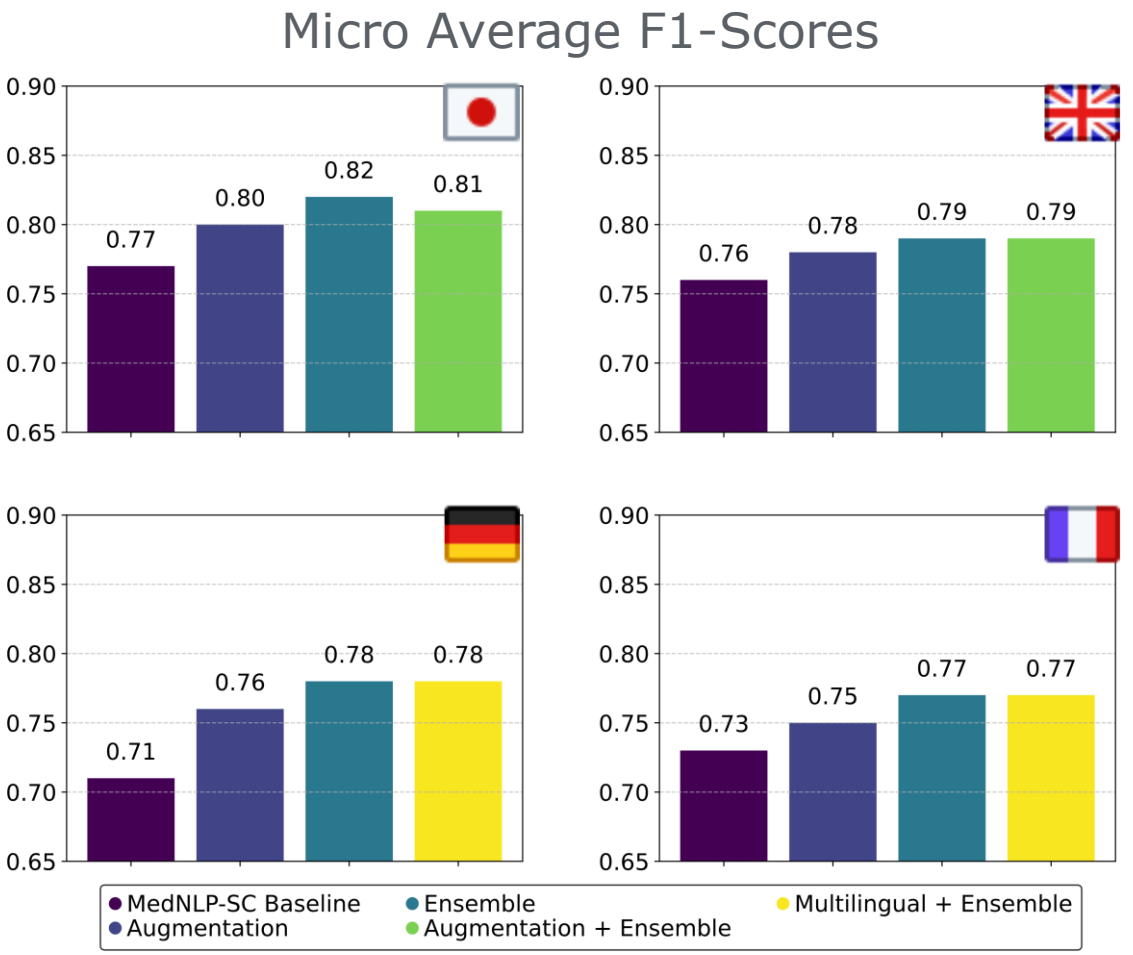
Results

Languages:

- Best overall results on Japanese data
- Best improvement to baseline on German data

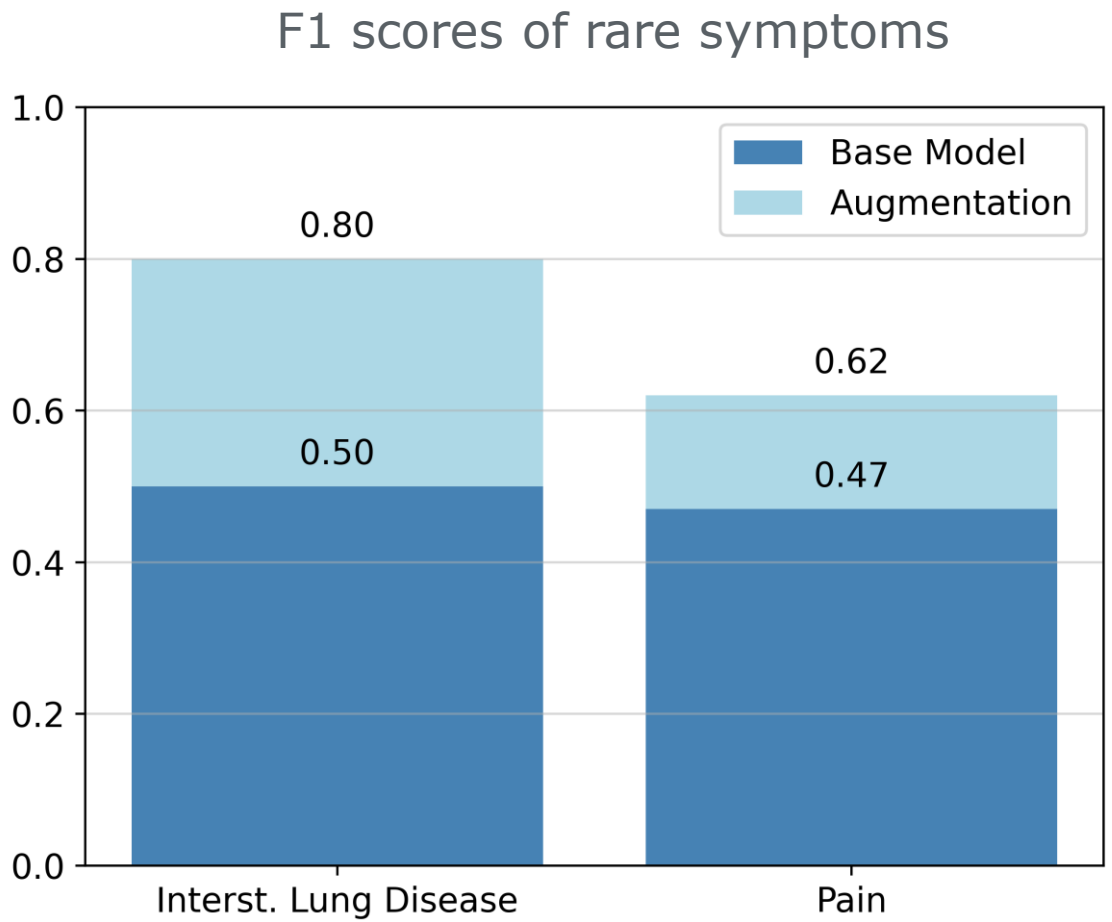
Methods:

- All submissions outperform baseline
- Ensembles perform best
- Effectiveness of augmentation shown



Results

Data Augmentation improves F1 scores of rare symptoms on test data



Conclusion

- Approach generalizes well to held-out test data

- Ensemble learning boosts overall accuracy of results
- Data augmentation improves low-frequency classes

- Second-best results in NTCIR-17 SM-ADE subtask across all languages

Conclusion

- Approach generalizes well to held-out test data

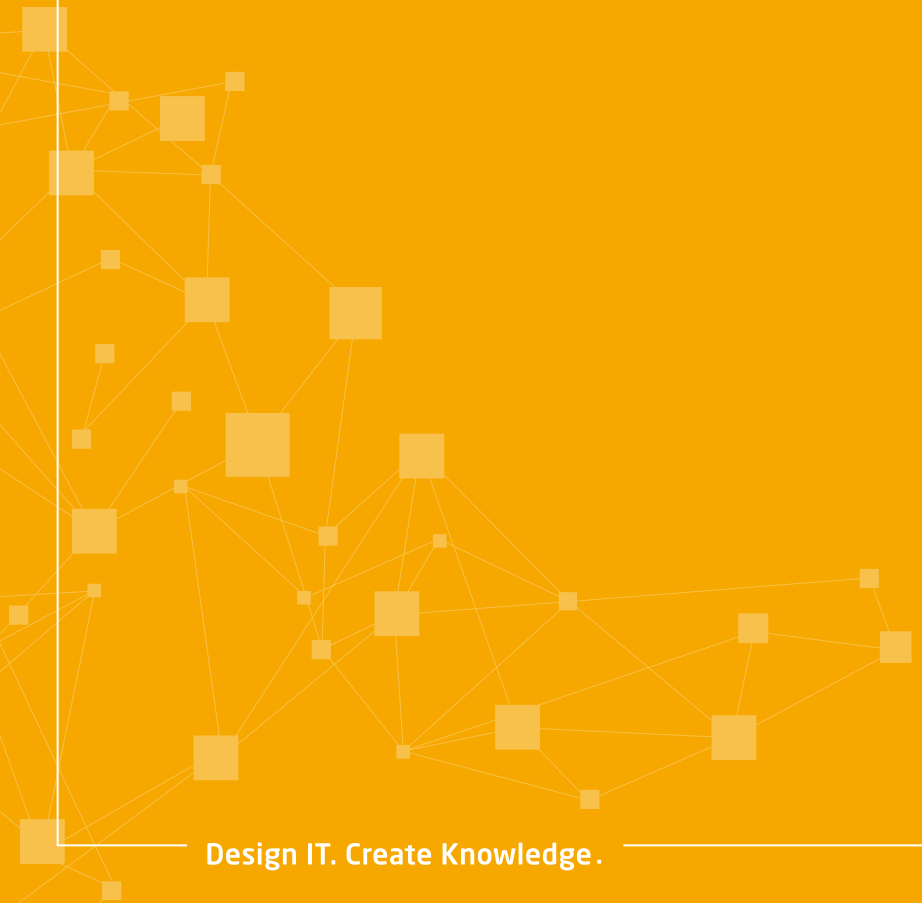
- Ensemble learning boosts overall accuracy of results
- Data augmentation improves low-frequency classes

- Second-best results in NTCIR-17 SM-ADE subtask across all languages

Future Work

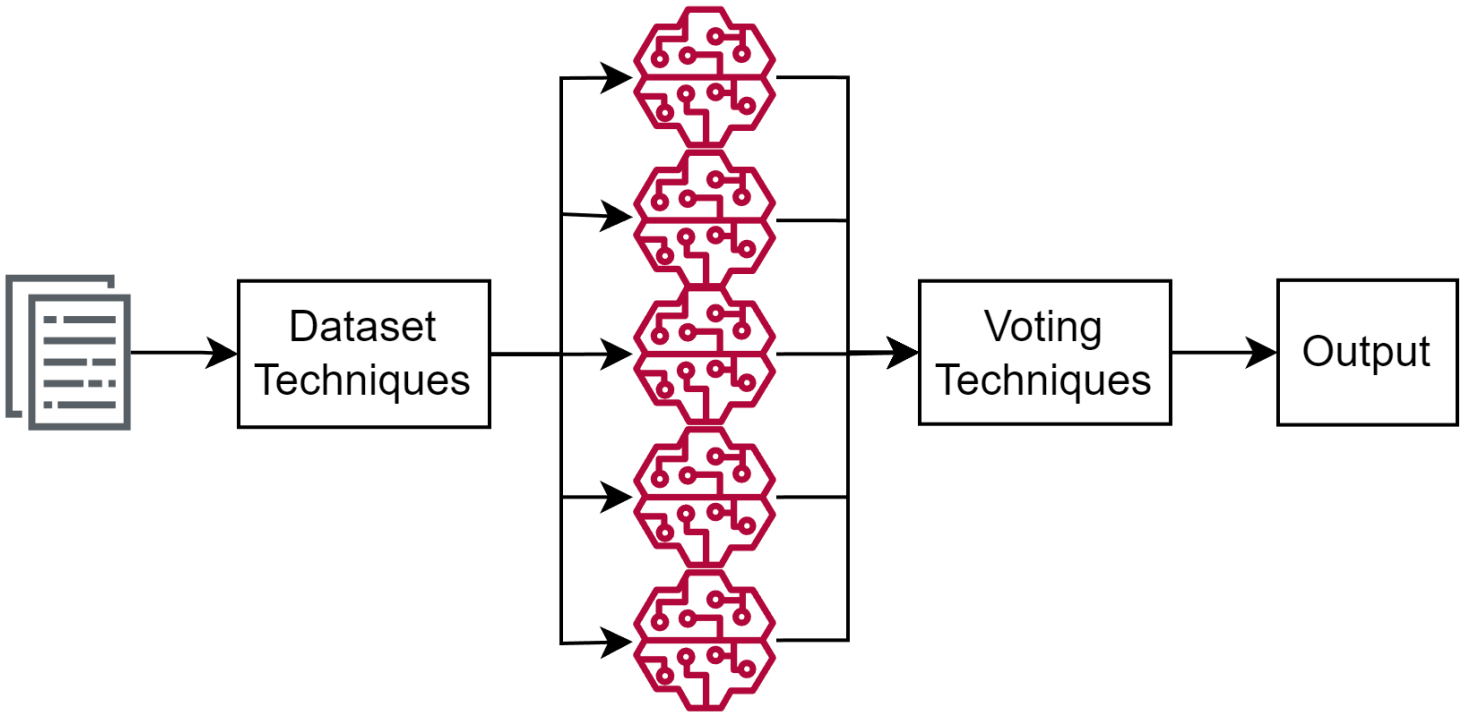
- Generate augmentation data with multiple ADE-labels
- Combine multi-lingual ensembles with data augmentation

Backup Slides

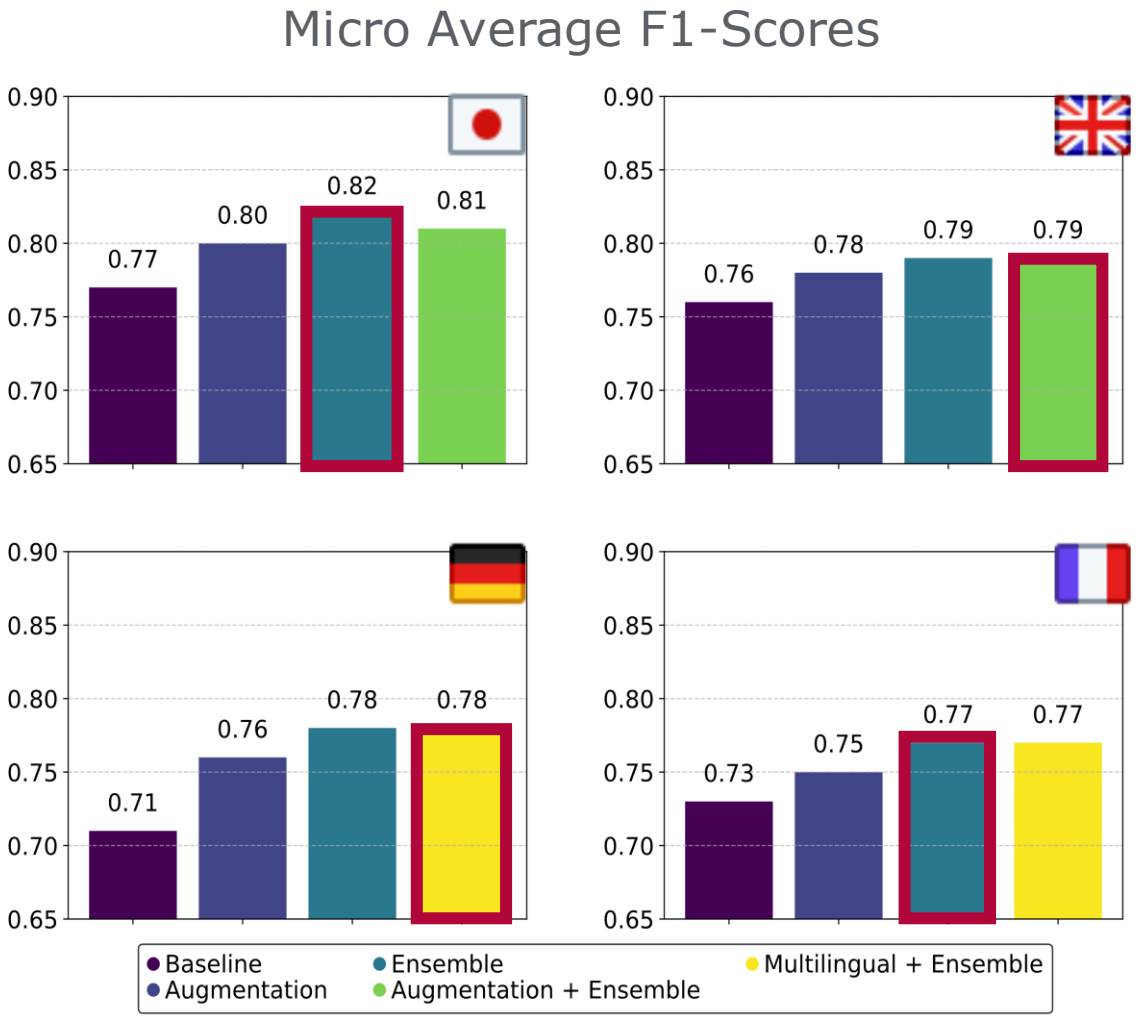


Ensemble Learning

- Dataset techniques:
  - KFOLD
  - Shuffle
- Voting techniques:
  - MajorityVote
  - WeightedVote
  - AverageProbabilityVote



Best Result


Micro Average F1-Scores
0.90
0.85
0.80
0.75
0.70
0.65
0.77
0.80
0.82
0.81
0.76
0.78
0.79
0.79
0.71
0.76
0.78
0.78
0.73
0.75
0.77
0.77
Baseline
Augmentation
Ensemble
Augmentation + Ensemble
Multilingual + Ensemble

Results

Baseline:

- XLM-R_{all} run of Subtask Overview

Augmentation:

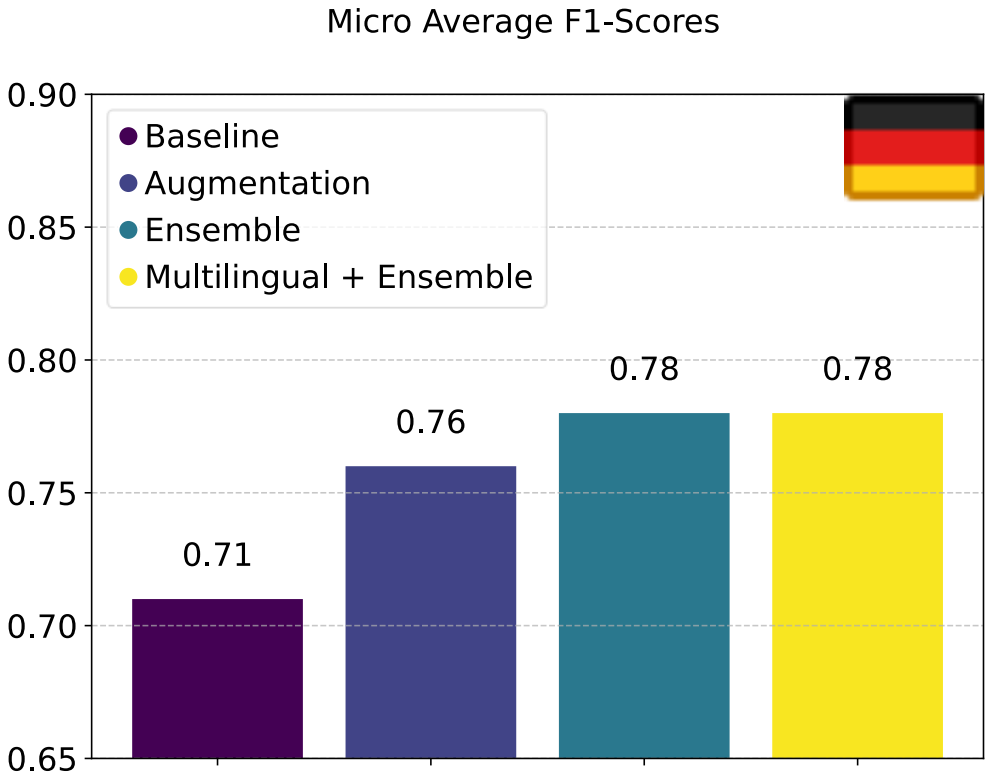
- Generate examples for 3 classes

Ensemble:

- KFOLD + MajorityVote

Multilingual + Ensemble:

- KFOLD + AverageProbabiltiyVote



Ensemble: Experimental Results

Table 7: Micro F1 Scores for Ensemble Learning Experiments on Japanese Test Set as described above.

	BL	MV	APV	MaxPV	MedPV	WV
Shuffle	.880	.888	**.891**	.859	.888	.881
K-fold	.884	.886	**.888**	.868	.886	.884
Shuffle+Aug	.888	.888	.889	.871	.888	**.891**
K-fold+Aug	.881	.883	.879	.870	.883	**.886**
Shuffle+Multi	.880	**.885**	.882	.857	**.885**	.884
K-fold+Multi	.875	.877	**.879**	.860	.877	.870

Table 6: Micro F1 Scores for Ensemble Learning Experiments on English Test Set as described above.

	BL	MV	APV	MaxPV	MedPV	WV
Shuffle	.862	.861	.858	.842	.861	**.866**
K-fold	.857	**.869**	.868	.842	**.869**	**.869**
Shuffle+Aug	.861	.861	**.866**	.849	.861	.860
K-fold+Aug	.866	.872	**.873**	.854	.872	.871
Shuffle+Multi	.858	**.862**	.860	.838	**.862**	.855
K-fold+Multi	.860	.863	**.865**	.840	.863	.864

	BL	MV	APV	MaxPV	MedPV	WV
Shuffle	.836	.835	**.838**	.831	.835	.833
K-fold	.841	**.846**	.843	.830	**.846**	.844
Shuffle+Aug	.837	.834	**.838**	.827	.834	.830
K-fold+Aug	.823	.830	**.831**	.819	.830	.826
Shuffle+Multi	.841	**.844**	.843	.830	**.844**	.841
K-fold+Multi	.838	.846	**.849**	.823	.846	.844

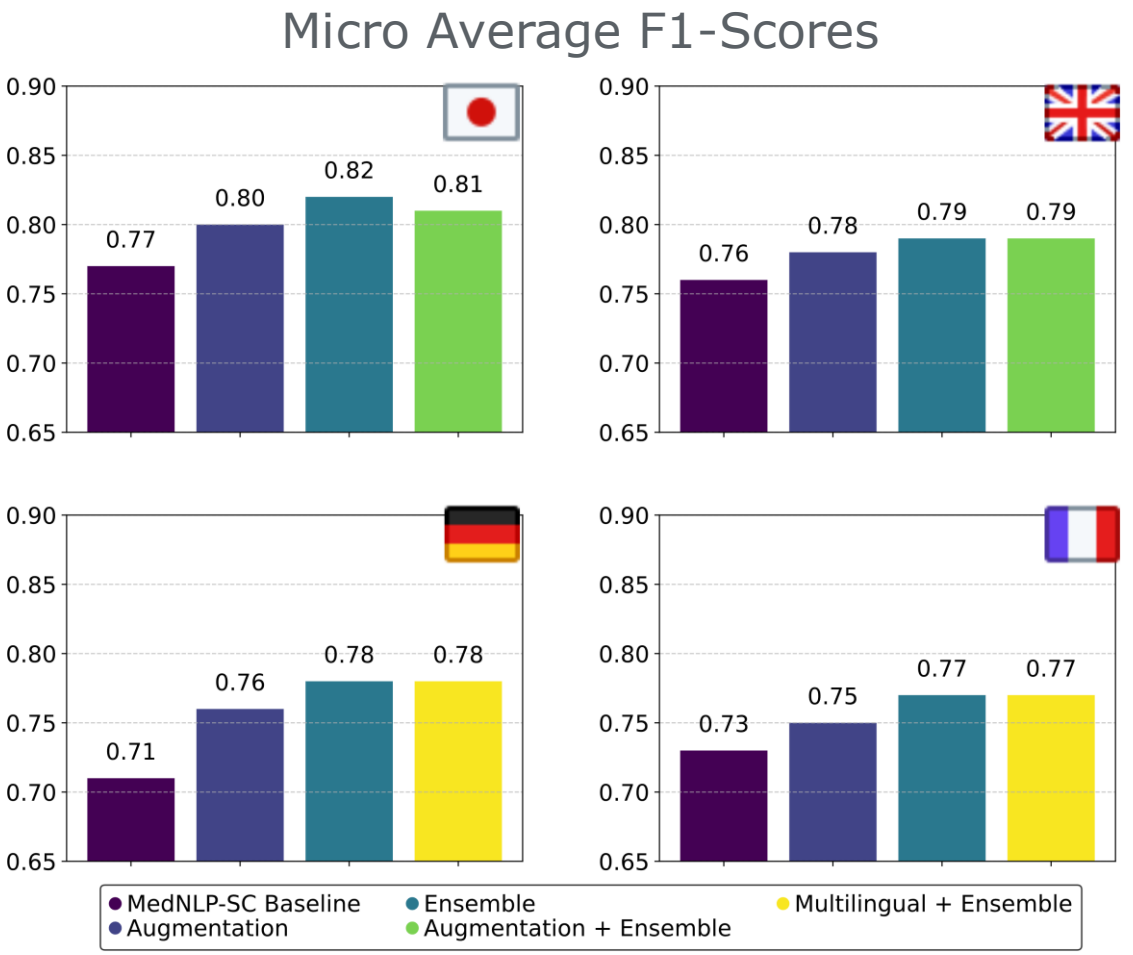
Table 5: Micro F1 Scores for Ensemble Learning Experiments on French Test Set as described above.

	BL	MV	APV	MaxPV	MedPV	WV
Shuffle	.842	.853	**.856**	.819	.853	.848
K-fold	.837	.841	**.844**	.825	.841	.840
Shuffle+Aug	.843	.847	**.852**	.814	.847	.851
K-fold+Aug	.844	.842	**.846**	.821	.842	.839
Shuffle+Multi	.843	**.844**	**.844**	.832	**.844**	.843
K-fold+Multi	.852	.863	**.864**	.841	.863	.859

Results

3. Submission

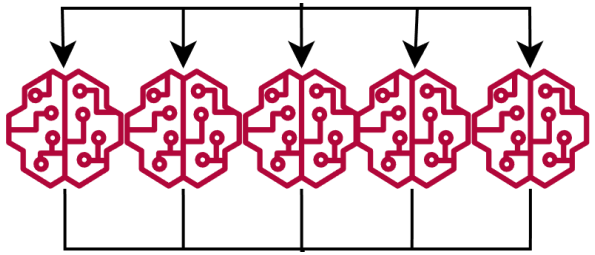
- Augmentation on mono-lingual dataset + ensemble better for Japanese and English
- Multi-lingual dataset + ensemble better for German and French



Submissions

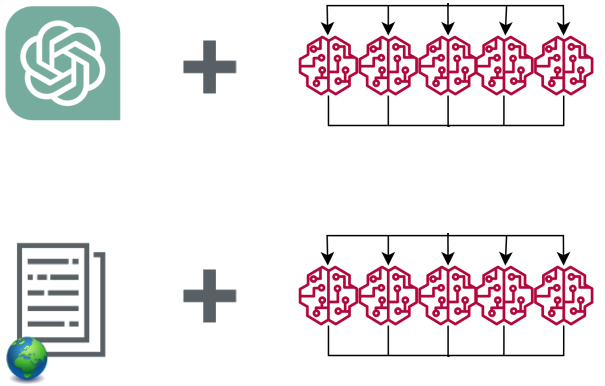
1. Submission:

Augmentation



2. Submission:

Ensemble



3. Submission:

Ensemble + Augmentation
OR
Ensemble + Multilingual