# Assignment4-Transformers

**Hanna Halka** **Maryna Kosse**

github — https://github.com/HannaHalka/Assignment4-Transformers

## About data

In this assignment we have two ain datasets: uk_geo_dataset and ru_geo_dataset.

**uk_geo_dataset: size - 1,000,000**

| text | loc_markers | org_markers | per_markers | is_valid |
|---|---|---|---|---|
| Чим довше мають скачки тиску гіпертензією, тим... | [ ] | [ ] | [ ] | 0 |
| ... | ... | ... | ... | ... |
| Україна, як і раніше, посідає в рейтингу 24-у ... | [(0, 7)] | [ ] | [ ] | 0 |

**ru_geo_dataset: size - 10,000,000**

| text | loc_markers | org_markers | per_markers | doc_id | sent_id |
|---|---|---|---|---|---|
| Вице-премьер по социальным вопросам Татьяна Го... | [(82, 88)] | [(149, 160)] | [(36, 52)] | 0 | 0 |
| ... | ... | ... | ... | ... | ... |
| — Никогда не думала, что это возможно». | [ ] | [ ] | [ ] | 16979 | 13 |

### Data pre-processing

The data has imbalanced classes, both in terms of language and in terms of the prevalence of the target variable. The best results were achieved when sampling equally from all 4 categories:

1. Ukrainian data with location tokens
2. Ukrainian data without location tokens
3. russian data with location tokens
4. russian data without location tokens

### Models

We decided to choose `microsoft/mdeberta-v3-base` because this model supports Ukrainian, and russian languages. **params:**

1. $\text{batch\_size} = 4$
3. $\texttt{lr} = 0.00003$
5. $\text{padding} = -100$

2. $\text{epochs} = 2$ 4. $\text{num\_labels} = 3$ 6. $\text{max\_len} = 512$

We've tested 4 variations of this model with different data samples:

Note: all models achieved similar high F1 scores during training, so they're not reported here.

1. A simple model trained on 1000 samples, all from the Ukrainian dataset with locations.

Performance:

```
Score: 0.57988
Private score: 0.63673
```

2. An ensemble of 3 simple models, each trained on a different 1000 samples from the same source.

Performance:

```
Score: 0.55507
Private score: 0.61030
```

3. An ensemble of 3 models with the same architecture, but trained on 10000 different samples each. Each of the 4 data "categories" were represented equally.

Performance:

```
Score: 0.63283
Private score: 0.66392
```

Note: We also tested this architecture on 1000 equally distributed samples, but the results were bad and weren't submitted.

4. An ensemble of 3 models with the same architecture, but trained on 70000 different samples each. Each of the 4 data "categories" were represented equally.

Performance:

```
Score: 0.59705
Private score: 0.63561
```

## Conclusion

The pre-trained mDeBERTa model is quite robust on its own, and requires minimal finetuning. While it does benefit from balanced, diverse datasets, it does not require them to achieve good performance. To the contrary, training the model on few datapoints with high representation of the target class produces better results than training on the same amount of more balanced data.