# 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

We split our data into 2 datasets one with locations and other without. But we can't just split them, they have different sizes so we will get an overfitting. So we took 600,000 rows from $\text{ru\_geo\_dataset} \rightarrow \binom{\text{ru\_df\_loc}}{\text{ru\_df\_nloc}}$ for each, and $\text{uk\_geo\_dataset} \rightarrow \binom{\text{ua\_df\_loc}}{\text{ua\_df\_nloc}}$ we took 200,000 and 600,000. Than we added oversampling for `uk_geo_dataset`. And last step we concat 90% from bouth `df_loc` and `df_no_loc` to train_df and other 10% to val_df.

### Model

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

1. $\text{batch\_size} = 4$
2. $\text{epochs} = 2$
3. $\texttt{lr} = 0.00003$
4. $\text{num\_labels} = 3$
5. $\text{padding} = -100$
6. $\text{max\_len} = 512$