# Massively Multilingual Corpus of Sentiment Datasets and Multi-faceted Sentiment Classification Benchmark

Łukasz Augustyniak, Szymon Woźniak, Marcin Gruza, Piotr Gramacki, Krzysztof Rajda, Mikołaj Morzy, Tomasz Kajdanowicz

WUST, Brand24, PUT

## INTRO

How to build a multilingual sentiment model?

- single multi-lingual model vs. dedicated monolingual models
- training vs. fine-tuning
- transfer learning between domains
- transfer learning between languages

## METHOD

- comprehensive linguistic typology
- manual selection of 79 datasets from the pool of 350 published datasets
- benchmarking 11 models (training mode, domain language, data modality, knowledge transfer)

## RESULTS

- multilingual corpus of 79 high-quality sentiment datasets in 27 languages
- multi-faceted benchmark
- open library for dataset access

## DISCUSSION

- large single multilingual model can perform approximately equally well for all languages
- all benchmarked models performed better when fine-tuned rather than trained from scratch
- bigger is better, but fine-tuning helps smaller models to be competitive
- large variability of performance across data splits

The **most extensive** open

**massively multilingual**

corpus of **sentiment** datasets

**80 datasets**

**27 languages**

**11 models**


**Take a picture** to visit HF dataset and download the paper, dataset, code examples, code snippets, and benchmark

BRAND24

Wrocław University of Science and Technology

POZNAN UNIVERSITY OF TECHNOLOGY

Fundusze Europejskie Program Regionalny

Rzeczpospolita Polska

DOLNY ŚLĄSK

Unia Europejska Europejski Fundusz Rozwoju Regionalnego

Table 1: Summary of the corpus. Categories: N - news, R - reviews, SM - social media, O - other

| | #datasets | category | | | | #samples | | | mean | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | N | R | SM | O | NEG | NEU | POS | #words | #chars |
| English | 17 | 3 | 4 | 6 | 4 | 304,939 | 290,823 | 1,734,724 | 62 | 339 |
| Arabic | 9 | 0 | 4 | 4 | 1 | 138,899 | 192,774 | 600,402 | 52 | 289 |
| Spanish | 5 | 0 | 3 | 2 | 0 | 108,733 | 122,493 | 187,486 | 26 | 150 |
| Chinese | 2 | 0 | 2 | 0 | 0 | 117,967 | 69,016 | 144,719 | 60 | 80 |
| German | 6 | 0 | 1 | 5 | 0 | 104,667 | 100,071 | 111,149 | 26 | 171 |
| Polish | 4 | 0 | 2 | 2 | 0 | 77,422 | 62,074 | 97,192 | 19 | 123 |
| French | 3 | 0 | 1 | 2 | 0 | 84,187 | 43,245 | 83,199 | 28 | 159 |
| Japanese | 1 | 0 | 1 | 0 | 0 | 83,982 | 41,979 | 83,819 | 61 | 101 |
| Czech | 4 | 0 | 2 | 2 | 0 | 39,674 | 59,200 | 97,413 | 34 | 212 |
| Portuguese | 4 | 0 | 0 | 4 | 0 | 56,827 | 55,165 | 45,842 | 11 | 63 |
| Slovenian | 2 | 1 | 0 | 1 | 0 | 33,694 | 50,553 | 29,296 | 41 | 269 |
| Russian | 2 | 0 | 0 | 2 | 0 | 31,770 | 48,106 | 31,054 | 11 | 70 |
| Croatian | 2 | 1 | 0 | 1 | 0 | 19,757 | 19,470 | 38,367 | 17 | 116 |
| Serbian | 3 | 0 | 2 | 1 | 0 | 25,089 | 32,283 | 18,996 | 44 | 269 |
| Thai | 2 | 0 | 1 | 1 | 0 | 9,326 | 28,616 | 34,377 | 22 | 381 |
| Bulgarian | 1 | 0 | 0 | 1 | 0 | 13,930 | 28,657 | 19,563 | 12 | 86 |
| Hungarian | 1 | 0 | 0 | 1 | 0 | 8,974 | 17,621 | 30,087 | 11 | 83 |
| Slovak | 1 | 0 | 0 | 1 | 0 | 14,431 | 12,842 | 29,350 | 13 | 98 |
| Albanian | 1 | 0 | 0 | 1 | 0 | 6,889 | 14,757 | 22,638 | 13 | 91 |
| Swedish | 1 | 0 | 0 | 1 | 0 | 16,266 | 13,342 | 11,738 | 14 | 94 |
| Bosnian | 1 | 0 | 0 | 1 | 0 | 11,974 | 11,145 | 13,064 | 12 | 76 |
| Urdu | 1 | 0 | 0 | 0 | 1 | 5,239 | 8,585 | 5,836 | 13 | 69 |
| Hindi | 1 | 0 | 0 | 1 | 0 | 4,992 | 6,392 | 5,615 | 26 | 128 |
| Persian | 1 | 0 | 1 | 0 | 0 | 1,602 | 5,091 | 6,832 | 21 | 104 |
| Italian | 2 | 0 | 0 | 2 | 0 | 4,043 | 4,193 | 3,829 | 16 | 103 |
| Hebrew | 1 | 0 | 0 | 1 | 0 | 2,279 | 243 | 6,097 | 22 | 110 |
| Latvian | 1 | 0 | 0 | 1 | 0 | 1,378 | 2,618 | 1,794 | 20 | 138 |

Table 2: Models included in the benchmark

| Model | #params | #langs | base | reference |
|---|---|---|---|---|
| mT5 | 277M | 101 | T5 | Xue et al. [93] |
| LASER | 52M | 93 | BiLSTM | Artetxe and Schwenk [5] |
| mBERT | 177M | 104 | BERT | Devlin et al. [26] |
| MPNet | 278M | 53 | XLM-R | Reimers and Gurevych [64] |
| XLM-R-dist | 278M | 53 | XLM-R | Reimers and Gurevych [64] |
| XLM-R | 278M | 100 | XLM-R | Conneau et al. [22] |
| LaBSE | 470M | 109 | BERT | Feng et al. [30] |
| DistilmBERT | 134M | 104 | BERT | Sanh et al. [73] |
| mUSE-dist | 134M | 53 | DistilmBERT | Reimers and Gurevych [64] |
| mUSE-transformer | 85M | 16 | transformer | Yang et al. [95] |
| mUSE-cnn | 68M | 16 | CNN | Yang et al. [95] |

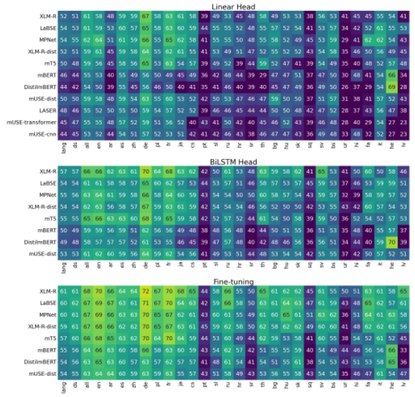


Figure 4: Detailed results of models' comparison. Legend: lang - averaged by all languages, ds - averaged by dataset, ar - Arabic, bg - Bulgarian, bs - Bosnian, cs - Czech, de - German, en - English, es - Spanish, fa - Persian, fr - French, he - Hebrew, hi - Hindi, hr - Croatian, hu - Hungarian, it - Italian, ja - Japanese, lv - Latvian, pl - Polish, pt - Portuguese, ru - Russian, sk - Slovak, sl - Slovenian, sq - Albanian, sr - Serbian, sv - Swedish, th - Thai, ur - Urdu, zh - Chinese.

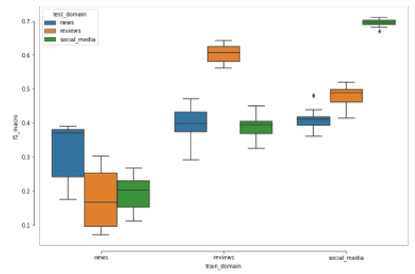


Figure 5: Transfer learning between data modalities