MMS Dataset and Benchmark

The most extensive open massively multilingual corpus of datasets for training sentiment models. The corpus consists of 79 manually selected from over 350 datasets reported in the scientific literature based on strict quality criteria and covers 27 languages.

Despite impressive advancements in multilingual corpora collection and model training, developing large-scale deployments of multilingual models still presents a significant challenge. This is particularly true for language tasks that are culture-dependent. One such example is the area of multilingual sentiment analysis, where affective markers can be subtle and deeply ensconced in culture.

This work presents the most extensive open massively multilingual corpus of datasets for training sentiment models. The corpus consists of 79 manually selected datasets from over 350 datasets reported in the scientific literature based on strict quality criteria. The corpus covers 27 languages representing 6 language families. Datasets can be queried using several linguistic and functional features. In addition, we present a multi-faceted sentiment classification benchmark summarizing hundreds of experiments conducted on different base models, training objectives, dataset collections, and fine-tuning strategies.

Dataset

Massively Multilingual Sentiment Datasets

Analysis and benchmarking

HuggingFace Spaces with Analysis and Benchmark

General statistics about the dataset

> It may take some time to download the dataset and generate train set inside HuggingFace dataset.
> Please be patient.

```
mms_dataset = datasets.load_dataset("Brand24/mms")
```

```
mms_dataset_df = mms_dataset["train"].to_pandas()
```

How many examples do we have?

```
mms_dataset.num_rows
```

```
{'train': 6164762}
```

Features

We provide not only texts and sentiment labels but we assigned many additional dimensions for datasets and languages, hence it is possible to splice and dice them as you want and need.

```
mms_dataset["train"].features
```

```
{'_id': Value(dtype='int32', id=None),
 'text': Value(dtype='string', id=None),
 'label': ClassLabel(names=['negative', 'neutral', 'positive'], id=None),
 'original_dataset': Value(dtype='string', id=None),
 'domain': Value(dtype='string', id=None),
 'language': Value(dtype='string', id=None),
 'Family': Value(dtype='string', id=None),
 'Genus': Value(dtype='string', id=None),
 'Definite articles': Value(dtype='string', id=None),
 'Indefinite articles': Value(dtype='string', id=None),
 'Number of cases': Value(dtype='string', id=None),
 'Order of subject, object, verb': Value(dtype='string', id=None),
 'Negative morphemes': Value(dtype='string', id=None),
 'Polar questions': Value(dtype='string', id=None),
 'Position of negative word wrt SOV': Value(dtype='string', id=None),
 'Prefixing vs suffixing': Value(dtype='string', id=None),
 'Coding of nominal plurality': Value(dtype='string', id=None),
 'Grammatical genders': Value(dtype='string', id=None),
 'cleanlab_self_confidence': Value(dtype='float32', id=None)}
```

Example

```
mms_dataset["train"][2001000]
```

```
{'_id': 2001000,
 'text': 'I was a tomboy and this has such great memories for me. They fit exactly how I
remember, PERFECTLY!!',
 'label': 2,
 'original_dataset': 'en_amazon',
 'domain': 'reviews',
 'language': 'en',
 'Family': 'Indo-European',
 'Genus': 'Germanic',
 'Definite articles': 'definite word distinct from demonstrative',
 'Indefinite articles': 'indefinite word distinct from one',
 'Number of cases': '2',
 'Order of subject, object, verb': 'SVO',
 'Negative morphemes': 'negative particle',
 'Polar questions': 'interrogative word order',
 'Position of negative word wrt SOV': 'SNegVO',
 'Prefixing vs suffixing': 'strongly suffixing',
```

```
'Coding of nominal plurality': 'plural suffix',
'Grammatical genders': 'no grammatical gender',
'cleanlab_self_confidence': 0.9978116750717163}
```

Classes

```
labels = mms_dataset["train"].features["label"].names
labels
```

```
['negative', 'neutral', 'positive']
```

```
mms_dataset_df["label_name"] = mms_dataset_df["label"].apply(lambda x: labels[x])
```

Classes distribution

```
labels_stats_df = pd.DataFrame(mms_dataset_df.label_name.value_counts())
labels_stats_df["percentage"] = (labels_stats_df["label_name"] / labels_stats_df["label_name"]
labels_stats_df
```

	label_name	percentage
positive	3494478	0.567
neutral	1341354	0.218
negative	1328930	0.216

Sentiment orientation for each language

```
cols = ['language', 'label_name']
mms_dataset_df[cols].value_counts().to_frame().reset_index().rename(columns={0: 'count'}).sort
```

	language	label_name	count
7	ar	negative	138899
4	ar	neutral	192774
1	ar	positive	600402
53	bg	negative	13930
41	bg	neutral	28657
...	...	...	...
62	ur	neutral	8585
67	ur	positive	5836
9	zh	negative	117967
21	zh	neutral	69016
6	zh	positive	144719

81 rows × 3 columns

Per language

```
cols = ['language']
mms_dataset_df[cols].value_counts().to_frame().reset_index().rename(columns={0: 'count'}).sort
```

	language	count
1	ar	932075
15	bg	62150
20	bs	36183
8	cs	196287
4	de	315887
0	en	2330486
2	es	418712
23	fa	13525
6	fr	210631
25	he	8619
22	hi	16999
12	hr	77594
16	hu	56682
24	it	12065
7	ja	209780
26	lv	5790
5	pl	236688
9	pt	157834
11	ru	110930
17	sk	56623
10	sl	113543
18	sq	44284
13	sr	76368
19	sv	41346
14	th	72319
21	ur	19660
3	zh	331702

Example of filtering datasets

Choose only Polish

```
pl = mms_dataset.filter(lambda row: row['language'] == 'pl')
```

```
pl["train"].to_pandas().sample(5)
```

	_id	text	label	original_dataset	domain	language	Family	Genus	Definite articles	Indef artic
215921	5119386	Typujcie jaki dziś będzie wynik St.Pats - Legi...	2	pl_twitter_sentiment	social_media	pl	Indo-European	Slavic	no article	no ar
86525	4989990	@KaczmarSF Przyjemne ciarki mam, gdy patrzę na...	2	pl_twitter_sentiment	social_media	pl	Indo-European	Slavic	no article	no ar
66031	4969496	szkoda bylo czasu i kasy .	0	pl_polemo	reviews	pl	Indo-European	Slavic	no article	no ar
137768	5041233	@shinyvalentine mam ja w dupie lecz bylo to kr...	0	pl_twitter_sentiment	social_media	pl	Indo-European	Slavic	no article	no ar
118766	5022231	@itiNieWracaj pokazują to gdzieś?	2	pl_twitter_sentiment	social_media	pl	Indo-European	Slavic	no article	no ar

Use cases

Case 1

Thus, when training a sentiment classifier using our dataset, one may download different facets of the collection. For instance, one can download all datasets in `Slavic` languages in which polar questions are formed using the interrogative word order or download all datasets from the `Afro-Asiatic` language family with no morphological case-making.

```
slavic = mms_dataset.filter(lambda row: row["Genus"] == "Slavic" and row["Polar questions"] ==
```

```
slavic
```

```
DatasetDict({
    train: Dataset({
        features: ['_id', 'text', 'label', 'original_dataset', 'domain', 'language',
```

```
'Family', 'Genus', 'Definite articles', 'Indefinite articles', 'Number of cases', 'Order of subject, object, verb', 'Negative morphemes', 'Polar questions', 'Position of negative word wrt SOV', 'Prefixing vs suffixing', 'Coding of nominal plurality', 'Grammatical genders', 'cleanlab_self_confidence'],
        num_rows: 252910
    })
})
```

Case 2

```
afro_asiatic = mms_dataset.filter(lambda row: row["Family"] == "Afro-Asiatic" and row["Number
```

```
afro_asiatic
```

```
DatasetDict({
    train: Dataset({
        features: ['_id', 'text', 'label', 'original_dataset', 'domain', 'language', 'Family', 'Genus', 'Definite articles', 'Indefinite articles', 'Number of cases', 'Order of subject, object, verb', 'Negative morphemes', 'Polar questions', 'Position of negative word wrt SOV', 'Prefixing vs suffixing', 'Coding of nominal plurality', 'Grammatical genders', 'cleanlab_self_confidence'],
        num_rows: 8619
    })
})
```

Dataset Curators

The corpus was put together by

- @laugustyniak
- @swozniak
- @mgruza
- @pgramacki
- @krajda
- @mmorzy
- @tkajdanowicz

Citation

```
@misc{augustyniak2023massively,
      title={Massively Multilingual Corpus of Sentiment Datasets and Multi-faceted Sentiment C
      author={Łukasz Augustyniak and Szymon Woźniak and Marcin Gruza and Piotr Gramacki and Kr
      year={2023},
      eprint={2306.07902},
      archivePrefix={arXiv},
      primaryClass={cs.CL}
}
```

Licensing Information

These data are released under this licensing scheme. We do not own any text from which these data and datasets have been extracted.

We license the actual packaging of these data under the Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) https://creativecommons.org/licenses/by-nc/4.0/

This work is published from Poland.

Should you consider that our data contains material that is owned by you and should, therefore not be reproduced here, please: * Clearly identify yourself with detailed contact data such as an address, telephone number, or email address at which you can be contacted. * Clearly identify the copyrighted work claimed to be infringed. * Clearly identify the material claimed to be infringing and the information reasonably sufficient to allow us to locate the material.

We will comply with legitimate requests by removing the affected sources from the next release of the corpus.