MMS - Massively Multilingual Sentiment Dataset and Benchmark

This is a static version of our online dataset, and benchmark documentation is available here:

- https://brand24-ai.github.io/mms_benchmark/

We advise you to look into online documentation that is much more interactive and easier to read and ingest.

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.

MMS Dataset Card

Dataset Card for https://huggingface.co/datasets/Brand24/mms

Easiness of using

One of the key ideas behind creating our library of datasets was to prioritize ease of use for researchers. Recognizing the importance of accessibility and convenience, we chose the HuggingFace platform as the storage and distribution platform for the datasets. HuggingFace provides a user-friendly interface and a wide range of tools and resources, making it easy for researchers to access and utilize the datasets.

To further enhance usability, we took the initiative to gather all the necessary citations for the datasets included in our library. By unifying the citations, we aimed to simplify and expedite the process of generating citations for researchers who utilize our datasets. This step reduces the time and effort required for researchers to acknowledge the datasets' sources properly.

However, it is essential to note that while we have taken steps to streamline the citation process, researchers should still independently verify the licenses of the datasets, especially if they intend to use them for purposes beyond strict academic research. Ensuring compliance with licensing requirements is crucial to maintaining ethical and legal data use standards.

Overall, our overarching goal in creating this unified corpus of datasets is accelerating academic sentiment analysis research. By providing a comprehensive collection of high-quality datasets and facilitating their accessibility, we aim to support researchers in exploring and advancing sentiment analysis techniques and methodologies.

Data ready to slice and dice and train a model

Our dataset is designed to be versatile and allows researchers to slice and dice the data for training and modeling according to their specific needs. Drawing from the field of linguistic typology, which examines the characteristics of languages, we have incorporated various linguistic features into our dataset selection process. These features include the text itself, sentiment labels, the original dataset source, domain, language, language family, genus, the presence or absence of definite and indefinite articles, the number of cases, word order, negative morphemes, polar questions, the position of negative morphemes, prefixing vs. suffixing, coding of nominal plurals, and grammatical genders. Researchers can easily access datasets that match their desired linguistic typology criteria by offering these features as filtering options in our library.

For instance, researchers can download datasets specific to Slavic languages with interrogative word order for polar questions or datasets from the Afro-Asiatic language family without morphological case-making. This flexibility empowers researchers to tailor their analyses and models to their linguistic interests and research questions.

```
import datasets

mms_dataset = datasets.load_dataset("Brand24/mms")
mms_dataset_df = mms_dataset["train"].to_pandas()
```

All features in dataset

```
mms_dataset_df.sample(5)
```

	_id	text	label	original_dataset	domain	language	Family	Genus	Defin articl
1117023	1117023	hlucnost mi prijde uplne v pohode, pere dobre,...	2	cs_mall_product_reviews	reviews	cs	Indo-European	Slavic	no art
824580	824580	"فندق جميل ولكن الخدمة جدا سيئه." .... الخدمة غير	0	ar_hard	reviews	ar	Afro-Asiatic	Semitic	defini
6014593	6014593	刚开始不习惯…之后还挺好用的…很轻便 很细…调节长度也很方便	2	zh_multilan_amazon	reviews	zh	Sino-Tibetan	Chinese	no art
5313872	5313872	Чемпионы. И этим все сказано.	2	ru_sentiment	social_media	ru	Indo-European	Slavic	no art
4290632	4290632	"@UnCharroDice: Y no ha de sobrar, quien con c...	1	es_twitter_sentiment	social_media	es	Indo-European	Romance	defini distin demc

Linguistic Typology

The field of language typology focuses on studying the similarities and differences among languages. These differences can be categorized into phonological (sounds), syntactic (structures), lexical (vocabulary), and theoretical aspects. Linguistic typology analyzes the current state of languages, contrasting with genealogical linguistics, which examines historical relationships between languages.

Genealogical linguistics studies language families and genera. A language family consists of languages that share a common ancestral language, while genera are branches within a language family. The Indo-European family, for example, includes genera such as Slavic, Romance, Germanic, and Indic. Over 7000 languages are categorized into approximately 150 language families, with Indo-European, Sino-Tibetan, Turkic, Afro-Asiatic, Nilo-Saharan, Niger-Congo, and Eskimo-Aleut being some of the largest families.

Within linguistic typology, languages are described using various linguistic features. Our work focuses on sentiment classification and selects ten relevant features:

- `text` : The feature text represents the actual text of the sentiment dataset. It is of type string and contains the text samples or sentences for sentiment analysis.
- `label` : The feature label corresponds to the sentiment labels of the text samples. It is of type ClassLabel and has three possible values: negative, neutral, and positive. These labels indicate the sentiment or emotional polarity associated with the text.
- `original_dataset` : The feature original_dataset refers to the name or identifier of the original dataset from which the text samples were extracted. It is of type string and provides information about the source dataset.
- `domain` : The feature domain represents the domain or topic of the sentiment dataset. It is of type string and provides context regarding the subject matter of the text samples.
- `language` : The feature language indicates the language of the text samples in the sentiment dataset. It is of type string and specifies the language in which the text is written.
- `Family` : The feature Family represents the language family to which a specific language belongs. It is of type string and provides information about the broader categorization of languages into language families.
- `Genus` : The feature Genus corresponds to the genus or branch within a language family. It is of type string and indicates the specific subgrouping of languages within a language family.
- `Definite article` : Half of the languages do not use the definite article, which signals uniqueness or definiteness of a concept.
- `Indefinite article` : Half of the languages do not use the indefinite article, with some languages using a separate article or the numeral “one.”
- `Number of cases` : Languages vary greatly in the number of morphological cases used.
- `Order of subject, verb, and object` : Different languages have different word orderings, with variations like SOV, SVO, VSO, VOS, OVS, and OSV.
- `Negative morphemes` : Negative morphemes indicate clausal negation in declarative sentences.
- `Polar questions` : Questions with yes/no answers, which can be formed using question particles, interrogative morphology, or intonation.
- `Position of the negative morpheme` : The position of the negative morpheme can vary in relation to subjects and objects.
- `Prefixing vs. suffixing` : Languages differ in their use of prefixes and suffixes in inflectional morphology.
- `Coding of nominal plurals` : Plurals can be expressed through morphological changes or the use of plurality indicator morphemes.
- `Grammatical genders` : Languages vary in the number of grammatical genders used, or may not use the concept at all.

These language features are available as filtering options in our library. Users can download specific facets of the collection, such as datasets in Slavic languages with interrogative word order for polar questions or datasets from the Afro-Asiatic language family without morphological case-making.

Datasheets for Datasets

The datasheets provide detailed information about the datasets, including data collection methods, annotation guidelines, and potential biases. They also specify the intended uses and potential limitations of

the datasets.

The initial pool of sentiment datasets was gathered through an extensive search using sources such as Google Scholar, GitHub repositories, and the HuggingFace datasets library. This search yielded a total of **345** datasets.

To ensure the quality of the datasets, a set of quality assurance criteria was applied to manually filter the initial pool of datasets. The following criteria were used:

1. **Strong Annotations**: Datasets containing weak annotations, such as labels based on emoji occurrence or automatically generated through classification by machine learning models, were rejected. This decision was made to minimize the presence of noise in the datasets, ensuring higher quality annotations.
2. **Well-Defined Annotation Protocol**: Datasets without sufficient information about the annotation protocol, including whether the annotation was done manually or automatically and the number of annotators involved, were rejected. This step aimed to avoid merging datasets with contradicting annotation instructions, ensuring consistency across the selected datasets.
3. **Numerical Ratings**: Datasets with numerical ratings were accepted. Specifically, Likert-type 5-point scales were mapped into three class sentiment labels. Ratings 1 and 2 were mapped to “negative,” rating 3 was mapped to “neutral,” and ratings 4 and 5 were mapped to “positive.” This mapping allowed for consistent sentiment labeling across the datasets.
4. **Three Classes Only**: Datasets annotated with binary sentiment labels were rejected. The decision to focus on datasets with three sentiment classes (negative, neutral, and positive) was made based on the unsatisfactory performance of binary sentiment labeling in three-class settings.
5. **Monolingual Datasets**: In cases where a dataset contained samples in multiple languages, it was divided into independent datasets for each constituent language. This approach ensured that the corpus includes separate datasets for different languages, allowing for targeted analysis and evaluation.

By applying these quality assurance criteria, we were able to filter the initial pool of sentiment datasets and select a final set of **79** datasets that met the specified standards for inclusion in the multilingual corpus.

```
f"We cover {mms_dataset_df.original_dataset.nunique()} datasets in {mms_dataset_df.language.nu
```

```
'We cover 79 datasets in 27 languages.'
```

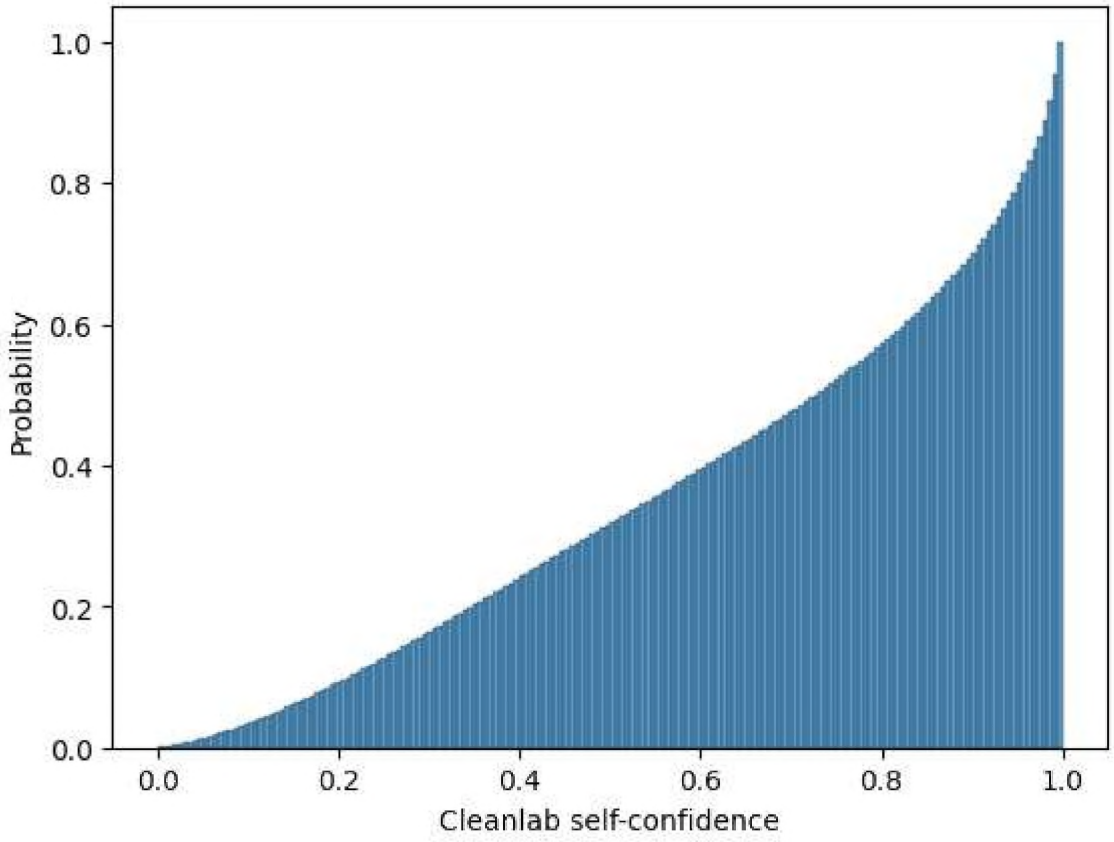
```
f"The classes that we cover: {mms_dataset_df.label_name.unique()}"
```

```
"The classes that we cover: ['positive' 'neutral' 'negative']"
```

Limitations

Despite the fact that our collection is the largest public collection of multilingual sentiment datasets, it still covers only 27 languages. The collection of datasets is highly biased towards the Indo-European family of languages, English in particular. We attribute this bias to the general culture of scientific publishing and its enforcement of English as the primary carrier of scientific discovery. Our work’s main potential negative social impact is that the models developed and trained using the provided datasets may still exhibit better performance for the major languages. This could further perpetuate the existing language disparities and inequality in sentiment analysis capabilities across different languages. Addressing this limitation and working towards more equitable representation and performance across languages is crucial to avoid reinforcing

language biases and the potential marginalization of underrepresented languages. The ethical implications of such disparities should be thoroughly discussed and considered.



Data Quality

An important limitation of our dataset collection is a significant variance in sample quality across all datasets and all languages. Above figure presents the distribution of self-confidence label-quality score for each data point computed by the **cleanlab** (Northcutt, Jiang, and Chuang 2021). The distribution of quality is skewed in favor of popular languages, with low-resource languages suffering from data quality issues. A related limitation is caused by an unequal distribution of data modalities across languages. For instance, our benchmark clearly shows that all models universally underperform when tested on Portuguese datasets. This is the direct result of the fact that data points for Portuguese almost exclusively represent the domain of social media. As a consequence, some combinations of filtering facets in our dataset collection produce very little data (i.e., asking for social media data in the Germanic genus of Indo-European languages will produce a significantly larger dataset than asking for news data representing Afro-Asiatic languages).

Finally, we acknowledge the lack of internal coherence of annotation protocols between datasets and languages. We have enforced strict quality criteria and rejected all datasets published without the annotation protocol, but we were unable, for obvious reasons, to unify annotation guidelines. The annotation of sentiment expressions and the assignment of sentiment labels are heavily subjective and, at the same time, influenced by cultural and linguistic features. Unfortunately, it is possible that semantically similar utterances will be assigned conflicting labels if they come from different datasets or modalities.

Filter examples by annotation qualitym

We know how imporant data quality is for the model training processes. Hence, we added cleanlab scores to each of 6M+ examples in all datasets. Now, it is enalbe to filter examples based on how good quality of data do you need for traning.

We can sort examples by top data quality. Cleanlab's `self confidence` is a function to compute label-quality scores for classification datasets, where lower scores indicate labels less likely to be correct. Hence, for the best quality we want to have the highest scores.

```
clean_labels_data = mms_dataset_df.sort_values(by="cleanlab_self_confidence", ascending=False)
```

```
clean_labels_data.head()
```

	_id	text	label	original_dataset	domain	language	Family	Genus	Definite articles	Indefinite articles
3075302	3075302	Great addition to any fan's yard! Show your te...	2	en_amazon	reviews	en	Indo-European	Germanic	definite word distinct from demonstrative	indefinite word distinct from one
629922	629922	مخيب للأمل. . بخ	0	ar_hard	reviews	ar	Afro-Asiatic	Semitic	definite affix	no article
2858237	2858237	This is a great flag to display your love of A...	2	en_amazon	reviews	en	Indo-European	Germanic	definite word distinct from demonstrative	indefinite word distinct from one
3110031	3110031	One of the best knives I now proudly own! Am a...	2	en_amazon	reviews	en	Indo-European	Germanic	definite word distinct from demonstrative	indefinite word distinct from one
2052971	2052971	Amen! My Savior Loves! Wonderful testimony!	2	en_amazon	reviews	en	Indo-European	Germanic	definite word distinct from demonstrative	indefinite word distinct from one

Datasets

We added all necessary citations to the HuggingFace datasets card. You can find them inside citation key. We added a helper fuinctions to parse them.

We can load citations as strings - easy adding to bibtex.

```
from mms_benchmark.citations import get_citations
```

```
print(get_citations(mms_dataset["train"], citation_as_dict=False)["pl_polemo"])
```

```
@inproceedings{dataset_pl_polemo,
    title = "Multi-Level Sentiment Analysis of {P}ol{E}mo 2.0: Extended Corpus of Multi-
Domain Consumer Reviews",
    author = "Koco{\'n}, Jan  and
        Mi{\l}kowski, Piotr  and
        Za{\'s}ko-Zieli{\'n}ska, Monika",
    booktitle = "Proceedings of the 23rd Conference on Computational Natural Language
Learning (CoNLL)",
    month = nov,
    year = "2019",
    address = "Hong Kong, China",
    publisher = "Association for Computational Linguistics",
    url = "https://aclanthology.org/K19-1092",
    doi = "10.18653/v1/K19-1092",
    pages = "980--991"
}
% ------------------------------------------------------------------------------------
```

Or as dictionary for working with them.

```
citations = get_citations(mms_dataset["train"], citation_as_dict=True)
```

```
citations["pl_polemo"]
```

```
{'pages': '980--991',
 'doi': '10.18653/v1/K19-1092',
 'url': 'https://aclanthology.org/K19-1092',
 'publisher': 'Association for Computational Linguistics',
 'address': 'Hong Kong, China',
 'year': '2019',
 'month': 'November',
 'booktitle': 'Proceedings of the 23rd Conference on Computational Natural Language Learning
(CoNLL)',
 'author': "Koco{\\'n}, Jan  and\nMi{\\l}kowski, Piotr  and\nZa{\\'s}ko-Zieli{\\'n}ska,
Monika",
 'title': 'Multi-Level Sentiment Analysis of {P}ol{E}mo 2.0: Extended Corpus of Multi-Domain
Consumer Reviews',
 'ENTRYTYPE': 'inproceedings',
 'ID': 'dataset_pl_polemo'}
```

Show all datasets with citations in a table

```
mms_dataset_df["citation"] = mms_dataset_df["original_dataset"].apply(lambda x: f'[@{citations
```

```
mms_dataset_df[DATASET_COLS].drop_duplicates().sort_values("language").reset_index(drop=True)
```

	language	original_dataset	domain	Family	Genus	Definite articles	Indefinite articles
0	ar	ar_arsentdl	social_media	Afro-Asiatic	Semitic	definite affix	no article
1	ar	ar_semeval_2017	mixed	Afro-Asiatic	Semitic	definite affix	no article
2	ar	ar_oclar	reviews	Afro-Asiatic	Semitic	definite affix	no article
3	ar	ar_labr	reviews	Afro-Asiatic	Semitic	definite affix	no article
4	ar	ar_syria_corpus	social_media	Afro-Asiatic	Semitic	definite affix	no article
5	ar	ar_brad	reviews	Afro-Asiatic	Semitic	definite affix	no article
6	ar	ar_bbn	social_media	Afro-Asiatic	Semitic	definite affix	no article
7	ar	ar_astd	social_media	Afro-Asiatic	Semitic	definite affix	no article
8	ar	ar_hard	reviews	Afro-Asiatic	Semitic	definite affix	no article
9	bg	bg_twitter_sentiment	social_media	Indo-European	Slavic	definite word distinct from demonstrative	no article
10	bs	bs_twitter_sentiment	social_media	Indo-European	Slavic	no article	no article
11	cs	cs_facebook	social_media	Indo-European	Slavic	no article	no article
12	cs	cs_mall_product_reviews	reviews	Indo-European	Slavic	no article	no article
13	cs	cs_movie_reviews	reviews	Indo-European	Slavic	no article	no article

	language	original_dataset	domain	Family	Genus	Definite articles	Indefinite articles
14	cs	cs_news_stance	social_media	Indo-European	Slavic	no article	no article
15	de	de_twitter_sentiment	social_media	Indo-European	Germanic	definite word distinct from demonstrative	indefinite word same as one
16	de	de_omp	social_media	Indo-European	Germanic	definite word distinct from demonstrative	indefinite word same as one
17	de	de_sb10k	social_media	Indo-European	Germanic	definite word distinct from demonstrative	indefinite word same as one
18	de	de_ifeel	social_media	Indo-European	Germanic	definite word distinct from demonstrative	indefinite word same as one
19	de	de_dai_labor	social_media	Indo-European	Germanic	definite word distinct from demonstrative	indefinite word same as one
20	de	de_multilan_amazon	reviews	Indo-European	Germanic	definite word distinct from demonstrative	indefinite word same as one
21	en	en_vader_twitter	social_media	Indo-European	Germanic	definite word distinct from demonstrative	indefinite word distinct from one
22	en	en_vader_nyt	news	Indo-European	Germanic	definite word distinct from demonstrative	indefinite word distinct from one
23	en	en_vader_movie_reviews	reviews	Indo-European	Germanic	definite word distinct from demonstrative	indefinite word distinct from one
24	en	en_vader_amazon	reviews	Indo-European	Germanic	definite word distinct from demonstrative	indefinite word distinct from one
25	en	en_twitter_sentiment	social_media	Indo-European	Germanic	definite word distinct from	indefinite word

						demonstrative	distinct from one
26	en **language**	en_tweets_sanders **original_dataset**	social_media **domain**	Indo-European **Family**	Germanic **Genus**	definite word distinct from demonstrative **Definite articles**	indefinite word distinct from one **Indefinite articles**
27	en	en_tweet_airlines	social_media	Indo-European	Germanic	definite word distinct from demonstrative	indefinite word distinct from one
28	en	en_silicone_sem	chats	Indo-European	Germanic	definite word distinct from demonstrative	indefinite word distinct from one
29	en	en_sentistrength	social_media	Indo-European	Germanic	definite word distinct from demonstrative	indefinite word distinct from one
30	en	en_semeval_2017	mixed	Indo-European	Germanic	definite word distinct from demonstrative	indefinite word distinct from one
31	en	en_poem_sentiment	poems	Indo-European	Germanic	definite word distinct from demonstrative	indefinite word distinct from one
32	en	en_per_sent	news	Indo-European	Germanic	definite word distinct from demonstrative	indefinite word distinct from one
33	en	en_multilan_amazon	reviews	Indo-European	Germanic	definite word distinct from demonstrative	indefinite word distinct from one
34	en	en_financial_phrasebank_sentences_75agree	news	Indo-European	Germanic	definite word distinct from demonstrative	indefinite word distinct from one
35	en	en_dai_labor	social_media	Indo-European	Germanic	definite word distinct from demonstrative	indefinite word distinct from one
36	en	en_amazon	reviews	Indo-European	Germanic	definite word distinct from demonstrative	indefinite word distinct from one
37	en	en_silicone_meld_s	chats	Indo-European	Germanic	definite word distinct from	indefinite word

	language	original_dataset	domain	Family	Genus	Definite articles	Indefinite articles
						demonstrative	distinct from one
38	es	es_twitter_sentiment	social_media	Indo-European	Romance	definite word distinct from demonstrative	indefinite word same as one
39	es	es_semeval2020	social_media	Indo-European	Romance	definite word distinct from demonstrative	indefinite word same as one
40	es	es_multilan_amazon	reviews	Indo-European	Romance	definite word distinct from demonstrative	indefinite word same as one
41	es	es_muchocine	reviews	Indo-European	Romance	definite word distinct from demonstrative	indefinite word same as one
42	es	es_paper_reviews	reviews	Indo-European	Romance	definite word distinct from demonstrative	indefinite word same as one
43	fa	fa_sentipers	reviews	Indo-European	Iranian	no article	indefinite word same as one
44	fr	fr_dai_labor	social_media	Indo-European	Romance	definite word distinct from demonstrative	indefinite word same as one
45	fr	fr_ifeel	social_media	Indo-European	Romance	definite word distinct from demonstrative	indefinite word same as one
46	fr	fr_multilan_amazon	reviews	Indo-European	Romance	definite word distinct from demonstrative	indefinite word same as one
47	he	he_hebrew_sentiment	social_media	Afro-Asiatic	Semitic	definite affix	indefinite word same as one
48	hi	hi_semeval2020	social_media	Indo-European	Indic	no article	no article

	language	original_dataset	domain	Family	Genus	Definite articles	Indefinite articles
49	hr	hr_sentiment_news_document	news	Indo-European	Slavic	no article	no article
50	hr	hr_twitter_sentiment	social_media	Indo-European	Slavic	no article	no article
51	hu	hu_twitter_sentiment	social_media	Uralic	Ugric	definite word distinct from demonstrative	indefinite word distinct from one
52	it	it_evalita2016	social_media	Indo-European	Romance	definite word distinct from demonstrative	indefinite word same as one
53	it	it_multiemotions	social_media	Indo-European	Romance	definite word distinct from demonstrative	indefinite word same as one
54	ja	ja_multilan_amazon	reviews	Japanese	Japanese	no article	indefinite word distinct from one
55	lv	lv_ltec_sentiment	social_media	Indo-European	Baltic	demonstrative word used as definite article	indefinite word same as one
56	pl	pl_twitter_sentiment	social_media	Indo-European	Slavic	no article	no article
57	pl	pl_polemo	reviews	Indo-European	Slavic	no article	no article
58	pl	pl_klej_allegro_reviews	reviews	Indo-European	Slavic	no article	no article
59	pl	pl_opi_lil_2012	social_media	Indo-European	Slavic	no article	no article
60	pt	pt_dai_labor	social_media	Indo-European	Romance	definite word distinct from demonstrative	indefinite word same as one
61	pt	pt_ifeel	social_media	Indo-European	Romance	definite word distinct from	indefinite word

	language	original_dataset	domain	Family	Genus	Definite articles	Indefinite articles
						demonstrative	same as one
62	pt	pt_tweet_sent_br	social_media	Indo-European	Romance	definite word distinct from demonstrative	indefinite word same as one
63	pt	pt_twitter_sentiment	social_media	Indo-European	Romance	definite word distinct from demonstrative	indefinite word same as one
64	ru	ru_sentiment	social_media	Indo-European	Slavic	no article	no article
65	ru	ru_twitter_sentiment	social_media	Indo-European	Slavic	no article	no article
66	sk	sk_twitter_sentiment	social_media	Indo-European	Slavic	no article	no article
67	sl	sl_sentinews	news	Indo-European	Slavic	no article	no article
68	sl	sl_twitter_sentiment	social_media	Indo-European	Slavic	no article	no article
69	sq	sq_twitter_sentiment	social_media	Indo-European	Albanian	definite affix	indefinite word distinct from one
70	sr	sr_movie_reviews	reviews	Indo-European	Slavic	no article	no article
71	sr	sr_senticomments	reviews	Indo-European	Slavic	no article	no article
72	sr	sr_twitter_sentiment	social_media	Indo-European	Slavic	no article	no article
73	sv	sv_twitter_sentiment	social_media	Indo-European	Germanic	definite affix	indefinite word same as one
74	th	th_wongnai_reviews	reviews	Tai-Kadai	Kam-Tai	no article	indefinite word

	language	original_dataset	domain	Family	Genus	Definite articles	Indefinite articles
							distinct from one
75	th	th_wisesight_sentiment	social_media	Tai-Kadai	Kam-Tai	no article	indefinite word distinct from one
76	ur	ur_roman_urdu	mixed	Indo-European	Indic	no article	no article
77	zh	zh_hotel_reviews	reviews	Sino-Tibetan	Chinese	no article	indefinite word same as one
78	zh	zh_multilan_amazon	reviews	Sino-Tibetan	Chinese	no article	indefinite word same as one

Dataset Stats

Datasets per language

```
pd.DataFrame(mms_dataset_df.groupby("language").original_dataset.nunique().sort_values(ascendi
```

language	original_dataset
en	17
ar	9
de	6
es	5
pl	4
cs	4
pt	4
sr	3
fr	3
th	2
sl	2
ru	2
it	2

	original_dataset
language	
hr	2
zh	2
bg	1
ja	1
lv	1
hu	1
hi	1
sk	1
he	1
sq	1
fa	1
sv	1
bs	1
ur	1

Labels per language

```
pd.DataFrame(mms_dataset_df.groupby(by=["language", "label_name"]).count()["text"])
```

		text
language	**label_name**	
ar	negative	138899
	neutral	192774
	positive	600402
bg	negative	13930
	neutral	28657
	positive	19563
bs	negative	11974
	neutral	11145
	positive	13064
cs	negative	39674
	neutral	59200
	positive	97413
de	negative	104667
	neutral	100071
	positive	111149
en	negative	304939

		text
language	**label_name**	
	neutral	290823
	positive	1734724
es	negative	108733
	neutral	122493
	positive	187486
fa	negative	1602
	neutral	5091
	positive	6832
fr	negative	84187
	neutral	43245
	positive	83199
he	negative	2279
	neutral	243
	positive	6097
hi	negative	4992
	neutral	6392
	positive	5615
hr	negative	19757
	neutral	19470
	positive	38367
hu	negative	8974
	neutral	17621
	positive	30087
it	negative	4043
	neutral	4193
	positive	3829
ja	negative	83982
	neutral	41979
	positive	83819
lv	negative	1378
	neutral	2618
	positive	1794
pl	negative	77422
	neutral	62074
	positive	97192
pt	negative	56827

	neutral	55165
language	**label_name**	**text**
	positive	45842
ru	negative	31770
	neutral	48106
	positive	31054
sk	negative	14431
	neutral	12842
	positive	29350
sl	negative	33694
	neutral	50553
	positive	29296
sq	negative	6889
	neutral	14757
	positive	22638
sr	negative	25089
	neutral	32283
	positive	18996
sv	negative	16266
	neutral	13342
	positive	11738
th	negative	9326
	neutral	28616
	positive	34377
ur	negative	5239
	neutral	8585
	positive	5836
zh	negative	117967
	neutral	69016
	positive	144719

Texts in Language Family and Genus

```
pd.DataFrame(mms_dataset_df.groupby(by=['Family', 'Genus',]).count()["text"])
```

		text
Family	**Genus**	
Afro-Asiatic	Semitic	940694
Indo-European	Albanian	44284
	Baltic	5790

		text
Family	**Genus**	
	Germanic	2687719
	Indic	36659
	Iranian	13525
	Romance	799242
	Slavic	966366
Japanese	Japanese	209780
Sino-Tibetan	Chinese	331702
Tai-Kadai	Kam-Tai	72319
Uralic	Ugric	56682

Examples per domain

```
pd.DataFrame(mms_dataset_df.groupby(by=["domain"]).count()["text"])
```

	text
domain	
chats	16781
mixed	94122
news	26413
poems	1052
reviews	4510893
social_media	1515501

Hosting, Licensing, and Maintenance Plan

- **Hosting**: The datasets and benchmark will be hosted on a reliable and scalable cloud infrastructure to ensure accessibility and availability (HuggingFace Hub). The choice of hosting platform will be based on factors such as reliability, performance, and cost-effectiveness.
- **Licensing**: We will clearly state the data license under which the datasets are released, ensuring that the terms of use are explicitly defined. We will consider licenses that facilitate research and allow for derivative works, while also addressing potential ethical considerations. See the license in repository.
- **Maintenance**: We (see Dataset Curators section) are committed to providing ongoing maintenance and support for the datasets and benchmark. This includes regular updates, bug fixes, and addressing any user feedback or inquiries. We will also establish a communication channel for users to report issues or request assistance.

References

Northcutt, Curtis, Lu Jiang, and Isaac Chuang. 2021. “Confident Learning: Estimating Uncertainty in Dataset Labels.” *Journal of Artificial Intelligence Research* 70: 1373–1411.

Benchmark results

The results of our benchmark for several Language Models using data from MMS.

Our preliminary results has been presented in (Rajda et al. 2022) and finally presented **in (Augustyniak et al. 2023) review at NeurIPS'23**.

Benchmark results - F1 Macro scores

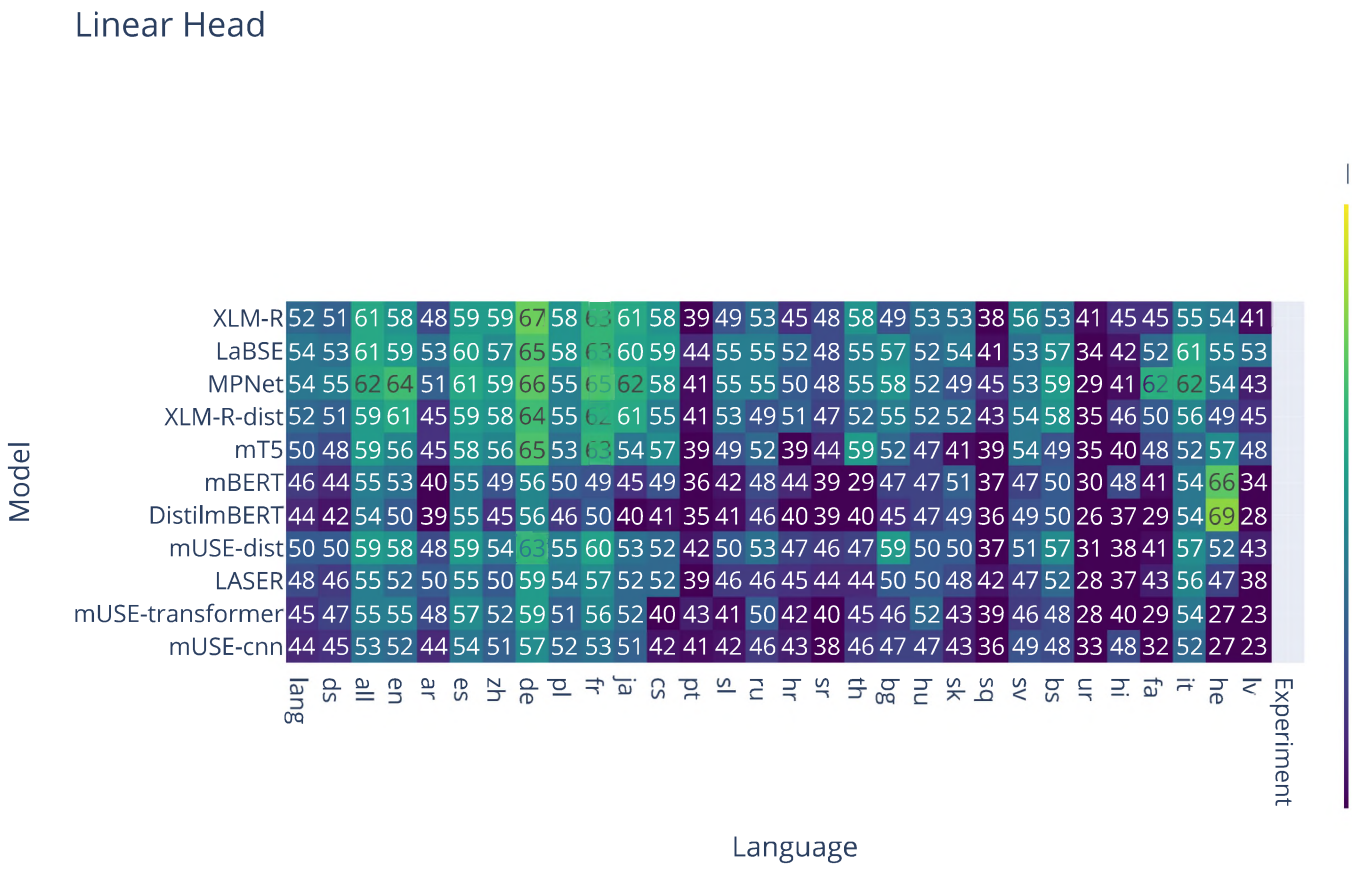
Models

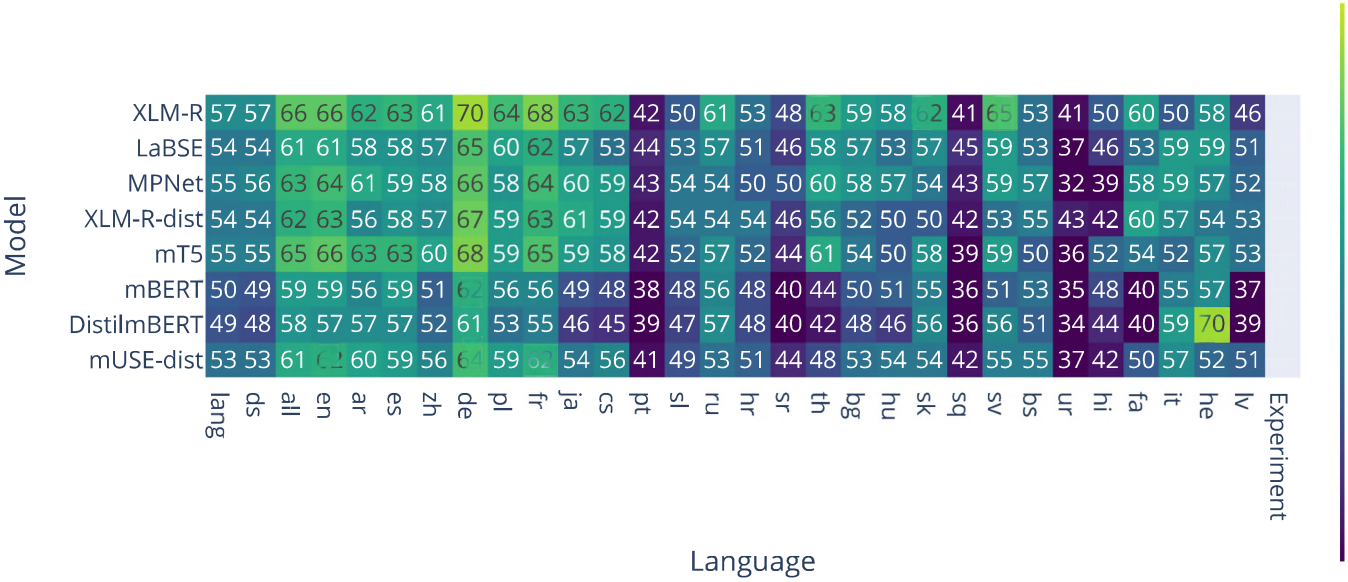
Model	Inf. time [s]	#params	#langs	base	data	reference
mT5	1.69	277M	101	T5	CC^b	(Xue et al. 2021)
LASER	1.64	52M	93	BiLSTM	$OPUS^c$	(Artetxe and Schwenk 2019)
mBERT	1.49	177M	104	BERT	Wiki	(Devlin et al. 2019)
MPNet**	1.38	278M	53	XLM-R	$OPUS^c$, $MUSE^d$, $Wikititles^e$	(Reimers and Gurevych 2020)
XLM-R-dist**	1.37	278M	53	XLM-R	$OPUS^c$, $MUSE^d$, $Wikititles^e$	(Reimers and Gurevych 2020)
XLM-R	1.37	278M	100	XLM-R	CC	(Conneau et al. 2020)
LaBSE	1.36	470M	109	BERT	CC, Wiki + mined bitexts	(Feng et al. 2020)
DistilmBERT	0.79	134M	104	BERT	Wiki	(Sanh et al. 2020)
mUSE-dist**	0.79	134M	53	DistilmBERT	$OPUS^c$, $MUSE^d$, $Wikititles^e$	(Reimers and Gurevych 2020)
mUSE-transformer*	0.65	85M	16	transformer	mined QA + bitexts, SNLI	(Yang et al. 2020)

Model	Inf. time [s]	#params	#langs	base	data	reference
mUSE-cnn*	0.12	68M	16	CNN	mined QA + bitexts, SNLI	(Yang et al. 2020)

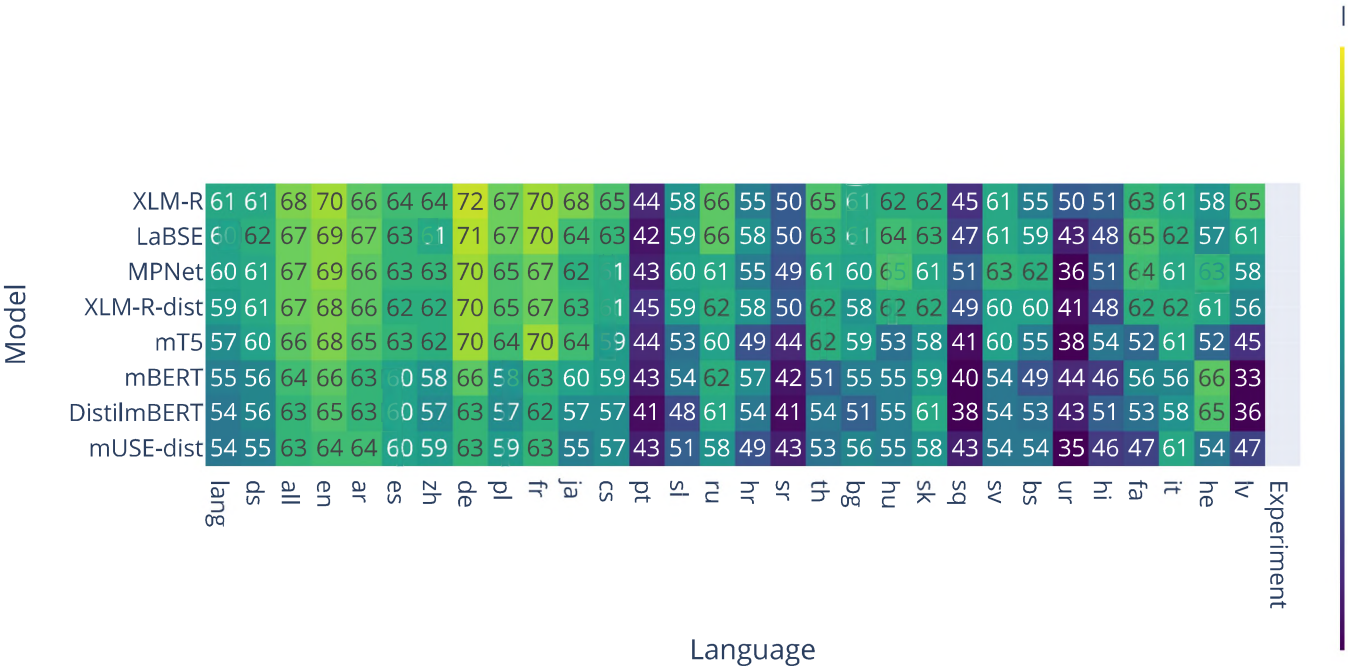
- * mUSE models were used in TensorFlow implementation in contrast to others in torch
- a Base model is either monolingual version on which it was based or another multilingual model which was used and adopted
- b Colossal Clean Crawled Corpus in multilingual version (mC4)
- c multiple datasets from OPUS website (https://opus.nlpl.eu)
- d bilingual dictionaries from MUSE (https://github.com/facebookresearch/MUSE)
- e just titles from wiki articles in multiple languages

Results





Fine-tuning



References

Artetxe, Mikel, and Holger Schwenk. 2019. "Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond." *Transactions of the Association for Computational Linguistics* 7 (September): 597–610. https://doi.org/10.1162/tacl_a_00288.

Augustyniak, Łukasz, Szymon Woźniak, Marcin Gruza, Piotr Gramacki, Krzysztof Rajda, Mikołaj Morzy, and Tomasz Kajdanowicz. 2023. "Massively Multilingual Corpus of Sentiment Datasets and Multi-Faceted Sentiment Classification Benchmark." https://arxiv.org/abs/2306.07902.

Conneau, Alexis, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. "Unsupervised Cross-Lingual Representation Learning at Scale." In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 8440–51. Online: Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.747.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding." In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–86. Minneapolis, Minnesota: Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423.

Feng, Fangxiaoyu, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020. "Language-agnostic BERT Sentence Embedding." *Computing Research Repository* arXiv:2007.01852. https://arxiv.org/abs/2007.01852.

Rajda, Krzysztof, Lukasz Augustyniak, Piotr Gramacki, Marcin Gruza, Szymon Woźniak, and Tomasz Kajdanowicz. 2022. "Assessment of Massively Multilingual Sentiment Classifiers." In *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*, 125–40. Dublin, Ireland: Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.wassa-1.13.

Reimers, Nils, and Iryna Gurevych. 2020. "Making Monolingual Sentence Embeddings Multilingual Using Knowledge Distillation." In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 4512–25. Online: Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.365.

Sanh, Victor, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter." *Computing Research Repository* arXiv:1910.01108. https://arxiv.org/abs/1910.01108.

Xue, Linting, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. "MT5: A Massively Multilingual Pre-Trained Text-to-Text Transformer." In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 483–98. Online: Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.naacl-main.41.

Yang, Yinfei, Daniel Cer, Amin Ahmad, Mandy Guo, Jax Law, Noah Constant, Gustavo Hernandez Abrego, et al. 2020. "Multilingual Universal Sentence Encoder for Semantic Retrieval." In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 87–94. Online: Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-demos.12.

MMS Dataset Citations

Citations for the MMS datasets

Citations

Dataset id: ar_arsentdl

- Domain: **social_media**
- Language: **ar**
- Language family: **Afro-Asiatic**
- Genus: **Semitic**
- Definite articles: **definite affix**
- Indefinite articles: **no article**
- Number of cases: **3**
- Order of subject, object, verb: **SVO**
- Negative morphemes: **negative particle**
- Polar questions: **interrogative intonation only**
- Position of negative word wrt SOV: **SNegVO**
- Prefixing vs suffixing: **weakly suffixing**
- Coding of nominal plurality: **mixed morphological plural**
- Grammatical genders: **masculine, feminine**

```
@InProceedings{dataset_ar_arsentdl,
    author = {Ramy Baly and
                Alaa Khaddaj and
                Hazem M. Hajj and
                Wassim El{-}Hajj and
                Khaled Bashir Shaban},
    title = {{ArSentD-LEV: A Multi-Topic Corpus for Target-based Sentiment Analysis in Arabic
    booktitle = {Proceedings of the Eleventh International Conference on Language Resources an
    year = {2018},
    month = {may},
    date = {7-12},
    location = {Miyazaki, Japan},
    editor = {Hend Al-Khalifa and King Saud University and KSA Walid Magdy and University of E
    publisher = {European Language Resources Association (ELRA)},
    address = {Paris, France},
    isbn = {979-10-95546-25-2},
    language = {english}
```

```
}
```

Dataset id: ar_astd

- Domain: **social_media**
- Language: **ar**
- Language family: **Afro-Asiatic**
- Genus: **Semitic**
- Definite articles: **definite affix**
- Indefinite articles: **no article**
- Number of cases: **3**
- Order of subject, object, verb: **SVO**
- Negative morphemes: **negative particle**
- Polar questions: **interrogative intonation only**
- Position of negative word wrt SOV: **SNegVO**
- Prefixing vs suffixing: **weakly suffixing**
- Coding of nominal plurality: **mixed morphological plural**
- Grammatical genders: **masculine, feminine**

```
@inproceedings{dataset_ar_astd,
    title = "{ASTD}: {A}rabic Sentiment Tweets Dataset",
    author = "Nabil, Mahmoud  and
        Aly, Mohamed  and
        Atiya, Amir",
    booktitle = "Proceedings of the 2015 Conference on Empirical Methods in Natural Language P
    month = sep,
    year = "2015",
    address = "Lisbon, Portugal",
    publisher = "Association for Computational Linguistics",
    url = "https://aclanthology.org/D15-1299",
    doi = "10.18653/v1/D15-1299",
    pages = "2515--2519",
}
```

Dataset id: ar_bbn

- Domain: **social_media**
- Language: **ar**
- Language family: **Afro-Asiatic**
- Genus: **Semitic**
- Definite articles: **definite affix**
- Indefinite articles: **no article**
- Number of cases: **3**
- Order of subject, object, verb: **SVO**
- Negative morphemes: **negative particle**
- Polar questions: **interrogative intonation only**

- Position of negative word wrt SOV: **SNegVO**
- Prefixing vs suffixing: **weakly suffixing**
- Coding of nominal plurality: **mixed morphological plural**
- Grammatical genders: **masculine, feminine**

```
@inproceedings{dataset_ar_bbn,
    title = "Sentiment after Translation: A Case-Study on {A}rabic Social Media Posts",
    author = "Salameh, Mohammad  and
        Mohammad, Saif  and
        Kiritchenko, Svetlana",
    booktitle = "Proceedings of the 2015 Conference of the North {A}merican Chapter of the Ass
    month = may # "{--}" # jun,
    year = "2015",
    address = "Denver, Colorado",
    publisher = "Association for Computational Linguistics",
    url = "https://aclanthology.org/N15-1078",
    doi = "10.3115/v1/N15-1078",
    pages = "767--777",
}
```

Dataset id: ar_brad

- Domain: **reviews**
- Language: **ar**
- Language family: **Afro-Asiatic**
- Genus: **Semitic**
- Definite articles: **definite affix**
- Indefinite articles: **no article**
- Number of cases: **3**
- Order of subject, object, verb: **SVO**
- Negative morphemes: **negative particle**
- Polar questions: **interrogative intonation only**
- Position of negative word wrt SOV: **SNegVO**
- Prefixing vs suffixing: **weakly suffixing**
- Coding of nominal plurality: **mixed morphological plural**
- Grammatical genders: **masculine, feminine**

```
@INPROCEEDINGS{dataset_ar_brad,
    author={Elnagar, Ashraf and Einea, Omar},
    booktitle={2016 IEEE/ACS 13th International Conference of Computer Systems and Application
    title={{BRAD} 1.0: Book reviews in Arabic dataset},
    year={2016},
    volume={},
    number={},
    pages={1-8},
    doi={10.1109/AICCSA.2016.7945800}
}
```

Dataset id: ar_hard

- Domain: **reviews**
- Language: **ar**
- Language family: **Afro-Asiatic**
- Genus: **Semitic**
- Definite articles: **definite affix**
- Indefinite articles: **no article**
- Number of cases: **3**
- Order of subject, object, verb: **SVO**
- Negative morphemes: **negative particle**
- Polar questions: **interrogative intonation only**
- Position of negative word wrt SOV: **SNegVO**
- Prefixing vs suffixing: **weakly suffixing**
- Coding of nominal plurality: **mixed morphological plural**
- Grammatical genders: **masculine, feminine**

```
@Book{dataset_ar_hard,
    author="Elnagar, Ashraf
    and Khalifa, Yasmin S.
    and Einea, Anas",
    title={Hotel Arabic-Reviews Dataset Construction for Sentiment Analysis Applications},
    bookTitle="Intelligent Natural Language Processing: Trends and Applications",
    year="2018",
    publisher="Springer International Publishing",
    address="Cham",
    pages="35--52",
    isbn="978-3-319-67056-0",
    doi="10.1007/978-3-319-67056-0_3",
    url="https://doi.org/10.1007/978-3-319-67056-0_3"
}
```

Dataset id: ar_labr

- Domain: **reviews**
- Language: **ar**
- Language family: **Afro-Asiatic**
- Genus: **Semitic**
- Definite articles: **definite affix**
- Indefinite articles: **no article**
- Number of cases: **3**
- Order of subject, object, verb: **SVO**
- Negative morphemes: **negative particle**
- Polar questions: **interrogative intonation only**
- Position of negative word wrt SOV: **SNegVO**
- Prefixing vs suffixing: **weakly suffixing**
- Coding of nominal plurality: **mixed morphological plural**
- Grammatical genders: **masculine, feminine**

```
@inproceedings{dataset_ar_labr,
    title = "{LABR}: A Large Scale {A}rabic Book Reviews Dataset",
    author = "Aly, Mohamed  and
        Atiya, Amir",
    booktitle = "Proceedings of the 51st Annual Meeting of the Association for Computational L
    month = aug,
    year = "2013",
    address = "Sofia, Bulgaria",
    publisher = "Association for Computational Linguistics",
    url = "https://aclanthology.org/P13-2088",
    pages = "494--498",
}
```

Dataset id: ar_oclar

- Domain: **reviews**
- Language: **ar**
- Language family: **Afro-Asiatic**
- Genus: **Semitic**
- Definite articles: **definite affix**
- Indefinite articles: **no article**
- Number of cases: **3**
- Order of subject, object, verb: **SVO**
- Negative morphemes: **negative particle**
- Polar questions: **interrogative intonation only**
- Position of negative word wrt SOV: **SNegVO**
- Prefixing vs suffixing: **weakly suffixing**
- Coding of nominal plurality: **mixed morphological plural**
- Grammatical genders: **masculine, feminine**

```
@inproceedings{dataset_ar_oclar,
    author={Al Omari, Marwan and Al-Hajj, Moustafa and Hammami, Nacereddine and Sabra, Amani},
    booktitle={2019 International Conference on Computer and Information Sciences (ICCIS)},
    title={Sentiment Classifier: Logistic Regression for Arabic Services' Reviews in Lebanon},
    year={2019},
    volume={},
    number={},
    pages={1-5},
    doi={10.1109/ICCISci.2019.8716394}
}
```

Dataset id: ar_semeval_2017

- Domain: **mixed**
- Language: **ar**
- Language family: **Afro-Asiatic**
- Genus: **Semitic**

- Definite articles: **definite affix**
- Indefinite articles: **no article**
- Number of cases: **3**
- Order of subject, object, verb: **SVO**
- Negative morphemes: **negative particle**
- Polar questions: **interrogative intonation only**
- Position of negative word wrt SOV: **SNegVO**
- Prefixing vs suffixing: **weakly suffixing**
- Coding of nominal plurality: **mixed morphological plural**
- Grammatical genders: **masculine, feminine**

```
@inproceedings{dataset_semeval_2017,
    title = "{S}em{E}val-2017 Task 4: Sentiment Analysis in {T}witter",
    author = "Rosenthal, Sara  and
        Farra, Noura  and
        Nakov, Preslav",
    booktitle = "Proceedings of the 11th International Workshop on Semantic Evaluation ({S}em{
    month = aug,
    year = "2017",
    address = "Vancouver, Canada",
    publisher = "Association for Computational Linguistics",
    url = "https://aclanthology.org/S17-2088",
    doi = "10.18653/v1/S17-2088",
    pages = "502--518",
    abstract = "This paper describes the fifth year of the Sentiment Analysis in Twitter task.
}
```

Dataset id: ar_syria_corpus

- Domain: **social_media**
- Language: **ar**
- Language family: **Afro-Asiatic**
- Genus: **Semitic**
- Definite articles: **definite affix**
- Indefinite articles: **no article**
- Number of cases: **3**
- Order of subject, object, verb: **SVO**
- Negative morphemes: **negative particle**
- Polar questions: **interrogative intonation only**
- Position of negative word wrt SOV: **SNegVO**
- Prefixing vs suffixing: **weakly suffixing**
- Coding of nominal plurality: **mixed morphological plural**
- Grammatical genders: **masculine, feminine**

```
@inproceedings{dataset_ar_bbn,
    title = "Sentiment after Translation: A Case-Study on {A}rabic Social Media Posts",
    author = "Salameh, Mohammad  and
        Mohammad, Saif  and
```

```
        Kiritchenko, Svetlana",
    booktitle = "Proceedings of the 2015 Conference of the North {A}merican Chapter of the Ass
    month = may # "{--}" # jun,
    year = "2015",
    address = "Denver, Colorado",
    publisher = "Association for Computational Linguistics",
    url = "https://aclanthology.org/N15-1078",
    doi = "10.3115/v1/N15-1078",
    pages = "767--777",
}
```

Dataset id: bg_twitter_sentiment

- Domain: **social_media**
- Language: **bg**
- Language family: **Indo-European**
- Genus: **Slavic**
- Definite articles: **definite word distinct from demonstrative**
- Indefinite articles: **no article**
- Number of cases: **no morphological case-making**
- Order of subject, object, verb: **SVO**
- Negative morphemes: **negative particle**
- Polar questions: **question particle**
- Position of negative word wrt SOV: **SNegVO**
- Prefixing vs suffixing: **strongly suffixing**
- Coding of nominal plurality: **plural suffix**
- Grammatical genders: **masculine, feminine, neuter**

```
@article{dataset_twitter_sentiment,
    doi = {10.1371/journal.pone.0155036},
    author = {Mozetič, Igor AND Grčar, Miha AND Smailović, Jasmina},
    journal = {PLOS ONE},
    publisher = {Public Library of Science},
    title = {Multilingual Twitter Sentiment Classification: The Role of Human Annotators},
    year = {2016},
    month = {05},
    volume = {11},
    url = {https://doi.org/10.1371/journal.pone.0155036},
    pages = {1-26},
    number = {5},
}
```

Dataset id: bs_twitter_sentiment

- Domain: **social_media**
- Language: **bs**
- Language family: **Indo-European**
- Genus: **Slavic**

- Definite articles: **no article**
- Indefinite articles: **no article**
- Number of cases: **5**
- Order of subject, object, verb: **SVO**
- Negative morphemes: **negative particle**
- Polar questions: **question particle**
- Position of negative word wrt SOV: **other**
- Prefixing vs suffixing: **strongly suffixing**
- Coding of nominal plurality: **plural suffix**
- Grammatical genders: **masculine, feminine, neuter**

```
@article{dataset_twitter_sentiment,
    doi = {10.1371/journal.pone.0155036},
    author = {Mozetič, Igor AND Grčar, Miha AND Smailović, Jasmina},
    journal = {PLOS ONE},
    publisher = {Public Library of Science},
    title = {Multilingual Twitter Sentiment Classification: The Role of Human Annotators},
    year = {2016},
    month = {05},
    volume = {11},
    url = {https://doi.org/10.1371/journal.pone.0155036},
    pages = {1-26},
    number = {5},
}
```

Dataset id: cs_facebook

- Domain: **social_media**
- Language: **cs**
- Language family: **Indo-European**
- Genus: **Slavic**
- Definite articles: **no article**
- Indefinite articles: **no article**
- Number of cases: **6-7**
- Order of subject, object, verb: **SVO**
- Negative morphemes: **negative affix**
- Polar questions: **interrogative word order**
- Position of negative word wrt SOV: **MorphNeg**
- Prefixing vs suffixing: **weakly suffixing**
- Coding of nominal plurality: **plural suffix**
- Grammatical genders: **masculine, feminine, neuter**

```
@inproceedings{dataset_cs_social_media,
    title = "Sentiment Analysis in {C}zech Social Media Using Supervised Machine Learning",
    author = "Habernal, Ivan  and
      Pt{\'a}{\v{c}}ek, Tom{\'a}{\v{s}}  and
      Steinberger, Josef",
    booktitle = "Proceedings of the 4th Workshop on Computational Approaches to Subjectivity,
    month = jun,
```

```
    year = "2013",
    address = "Atlanta, Georgia",
    publisher = "Association for Computational Linguistics",
    url = "https://aclanthology.org/W13-1609",
    pages = "65--74",
}
```

Dataset id: cs_mall_product_reviews

- Domain: **reviews**
- Language: **cs**
- Language family: **Indo-European**
- Genus: **Slavic**
- Definite articles: **no article**
- Indefinite articles: **no article**
- Number of cases: **6-7**
- Order of subject, object, verb: **SVO**
- Negative morphemes: **negative affix**
- Polar questions: **interrogative word order**
- Position of negative word wrt SOV: **MorphNeg**
- Prefixing vs suffixing: **weakly suffixing**
- Coding of nominal plurality: **plural suffix**
- Grammatical genders: **masculine, feminine, neuter**

```
@inproceedings{dataset_cs_social_media,
    title = "Sentiment Analysis in {C}zech Social Media Using Supervised Machine Learning",
    author = "Habernal, Ivan  and
      Pt{\'a}{\v{c}}ek, Tom{\'a}{\v{s}}  and
      Steinberger, Josef",
    booktitle = "Proceedings of the 4th Workshop on Computational Approaches to Subjectivity,
    month = jun,
    year = "2013",
    address = "Atlanta, Georgia",
    publisher = "Association for Computational Linguistics",
    url = "https://aclanthology.org/W13-1609",
    pages = "65--74",
}
```

Dataset id: cs_movie_reviews

- Domain: **reviews**
- Language: **cs**
- Language family: **Indo-European**
- Genus: **Slavic**
- Definite articles: **no article**
- Indefinite articles: **no article**
- Number of cases: **6-7**

- Order of subject, object, verb: **SVO**
- Negative morphemes: **negative affix**
- Polar questions: **interrogative word order**
- Position of negative word wrt SOV: **MorphNeg**
- Prefixing vs suffixing: **weakly suffixing**
- Coding of nominal plurality: **plural suffix**
- Grammatical genders: **masculine, feminine, neuter**

```
@inproceedings{dataset_cs_social_media,
    title = "Sentiment Analysis in {C}zech Social Media Using Supervised Machine Learning",
    author = "Habernal, Ivan  and
      Pt{\'a}{\v{c}}ek, Tom{\'a}{\v{s}}  and
      Steinberger, Josef",
    booktitle = "Proceedings of the 4th Workshop on Computational Approaches to Subjectivity,
    month = jun,
    year = "2013",
    address = "Atlanta, Georgia",
    publisher = "Association for Computational Linguistics",
    url = "https://aclanthology.org/W13-1609",
    pages = "65--74",
}
```

Dataset id: cs_news_stance

- Domain: **social_media**
- Language: **cs**
- Language family: **Indo-European**
- Genus: **Slavic**
- Definite articles: **no article**
- Indefinite articles: **no article**
- Number of cases: **6-7**
- Order of subject, object, verb: **SVO**
- Negative morphemes: **negative affix**
- Polar questions: **interrogative word order**
- Position of negative word wrt SOV: **MorphNeg**
- Prefixing vs suffixing: **weakly suffixing**
- Coding of nominal plurality: **plural suffix**
- Grammatical genders: **masculine, feminine, neuter**

```
@inproceedings{dataset_cs_social_media,
    title = "Sentiment Analysis in {C}zech Social Media Using Supervised Machine Learning",
    author = "Habernal, Ivan  and
      Pt{\'a}{\v{c}}ek, Tom{\'a}{\v{s}}  and
      Steinberger, Josef",
    booktitle = "Proceedings of the 4th Workshop on Computational Approaches to Subjectivity,
    month = jun,
    year = "2013",
    address = "Atlanta, Georgia",
```

```
    publisher = "Association for Computational Linguistics",
    url = "https://aclanthology.org/W13-1609",
    pages = "65--74",
}
```

Dataset id: de_dai_labor

- Domain: **social_media**
- Language: **de**
- Language family: **Indo-European**
- Genus: **Germanic**
- Definite articles: **definite word distinct from demonstrative**
- Indefinite articles: **indefinite word same as one**
- Number of cases: **4**
- Order of subject, object, verb: **no dominant order**
- Negative morphemes: **negative particle**
- Polar questions: **interrogative word order**
- Position of negative word wrt SOV: **more than one position**
- Prefixing vs suffixing: **strongly suffixing**
- Coding of nominal plurality: **plural suffix**
- Grammatical genders: **masculine, feminine, neuter**

```
@inproceedings{dataset_dai_labor,
    author = {Narr, Sascha  and Michael Hülfenhaus and  Albayrak, Sahin},
    title = {Language-Independent Twitter Sentiment Analysis},
    booktitle = {Workshop on Knowledge Discovery, Data Mining and Machine Learning (KDML-2012)
    year = {2012},
    location = {Dortmund, Germany},
}
```

Dataset id: de_ifeel

- Domain: **social_media**
- Language: **de**
- Language family: **Indo-European**
- Genus: **Germanic**
- Definite articles: **definite word distinct from demonstrative**
- Indefinite articles: **indefinite word same as one**
- Number of cases: **4**
- Order of subject, object, verb: **no dominant order**
- Negative morphemes: **negative particle**
- Polar questions: **interrogative word order**
- Position of negative word wrt SOV: **more than one position**
- Prefixing vs suffixing: **strongly suffixing**
- Coding of nominal plurality: **plural suffix**

- Grammatical genders: **masculine, feminine, neuter**

```
@inproceedings{dataset_dai_labor,
    author = {Narr, Sascha  and Michael Hülfenhaus and  Albayrak, Sahin},
    title = {Language-Independent Twitter Sentiment Analysis},
    booktitle = {Workshop on Knowledge Discovery, Data Mining and Machine Learning (KDML-2012)
    year = {2012},
    location = {Dortmund, Germany},
}
```

Dataset id: de_multilan_amazon

- Domain: **reviews**
- Language: **de**
- Language family: **Indo-European**
- Genus: **Germanic**
- Definite articles: **definite word distinct from demonstrative**
- Indefinite articles: **indefinite word same as one**
- Number of cases: **4**
- Order of subject, object, verb: **no dominant order**
- Negative morphemes: **negative particle**
- Polar questions: **interrogative word order**
- Position of negative word wrt SOV: **more than one position**
- Prefixing vs suffixing: **strongly suffixing**
- Coding of nominal plurality: **plural suffix**
- Grammatical genders: **masculine, feminine, neuter**

```
@inproceedings{dataset_multilan_amazon,
    title = "The Multilingual {A}mazon Reviews Corpus",
    author = {Keung, Phillip  and
        Lu, Yichao  and
        Szarvas, Gy{\"o}rgy  and
        Smith, Noah A.},
    booktitle = "Proceedings of the 2020 Conference on Empirical Methods in Natural Language P
    month = nov,
    year = "2020",
    address = "Online",
    publisher = "Association for Computational Linguistics",
    url = "https://aclanthology.org/2020.emnlp-main.369",
    doi = "10.18653/v1/2020.emnlp-main.369",
    pages = "4563--4568",
}
```

Dataset id: de_omp

- Domain: **social_media**
- Language: **de**

- Language family: **Indo-European**
- Genus: **Germanic**
- Definite articles: **definite word distinct from demonstrative**
- Indefinite articles: **indefinite word same as one**
- Number of cases: **4**
- Order of subject, object, verb: **no dominant order**
- Negative morphemes: **negative particle**
- Polar questions: **interrogative word order**
- Position of negative word wrt SOV: **more than one position**
- Prefixing vs suffixing: **strongly suffixing**
- Coding of nominal plurality: **plural suffix**
- Grammatical genders: **masculine, feminine, neuter**

```
@inproceedings{dataset_de_omp,
    title = "Academic-Industrial Perspective on the Development and Deployment of a Moderation
    author = "Schabus, Dietmar  and
        Skowron, Marcin",
    booktitle = "Proceedings of the Eleventh International Conference on Language Resources an
    month = may,
    year = "2018",
    address = "Miyazaki, Japan",
    publisher = "European Language Resources Association (ELRA)",
    url = "https://aclanthology.org/L18-1253",
}
```

Dataset id: de_sb10k

- Domain: **social_media**
- Language: **de**
- Language family: **Indo-European**
- Genus: **Germanic**
- Definite articles: **definite word distinct from demonstrative**
- Indefinite articles: **indefinite word same as one**
- Number of cases: **4**
- Order of subject, object, verb: **no dominant order**
- Negative morphemes: **negative particle**
- Polar questions: **interrogative word order**
- Position of negative word wrt SOV: **more than one position**
- Prefixing vs suffixing: **strongly suffixing**
- Coding of nominal plurality: **plural suffix**
- Grammatical genders: **masculine, feminine, neuter**

```
@inproceedings{dataset_de_sb10k,
    title = "A {T}witter Corpus and Benchmark Resources for {G}erman Sentiment Analysis",
    author = "Cieliebak, Mark  and
        Deriu, Jan Milan  and
        Egger, Dominic  and
        Uzdilli, Fatih",
```

```
    booktitle = "Proceedings of the Fifth International Workshop on Natural Language Processin
    month = apr,
    year = "2017",
    address = "Valencia, Spain",
    publisher = "Association for Computational Linguistics",
    url = "https://aclanthology.org/W17-1106",
    doi = "10.18653/v1/W17-1106",
    pages = "45--51",
    abstract = "In this paper we present SB10k, a new corpus for sentiment analysis with appro
}
```

Dataset id: de_twitter_sentiment

- Domain: **social_media**
- Language: **de**
- Language family: **Indo-European**
- Genus: **Germanic**
- Definite articles: **definite word distinct from demonstrative**
- Indefinite articles: **indefinite word same as one**
- Number of cases: **4**
- Order of subject, object, verb: **no dominant order**
- Negative morphemes: **negative particle**
- Polar questions: **interrogative word order**
- Position of negative word wrt SOV: **more than one position**
- Prefixing vs suffixing: **strongly suffixing**
- Coding of nominal plurality: **plural suffix**
- Grammatical genders: **masculine, feminine, neuter**

```
@article{dataset_twitter_sentiment,
    doi = {10.1371/journal.pone.0155036},
    author = {Mozetič, Igor AND Grčar, Miha AND Smailović, Jasmina},
    journal = {PLOS ONE},
    publisher = {Public Library of Science},
    title = {Multilingual Twitter Sentiment Classification: The Role of Human Annotators},
    year = {2016},
    month = {05},
    volume = {11},
    url = {https://doi.org/10.1371/journal.pone.0155036},
    pages = {1-26},
    number = {5},
}
```

Dataset id: en_amazon

- Domain: **reviews**
- Language: **en**
- Language 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**

```
@inproceedings{dataset_en_amazon,
    title = "Justifying Recommendations using Distantly-Labeled Reviews and Fine-Grained Aspec
    author = "Ni, Jianmo  and
        Li, Jiacheng  and
        McAuley, Julian",
    booktitle = "Proceedings of the 2019 Conference on Empirical Methods in Natural Language P
    month = nov,
    year = "2019",
    address = "Hong Kong, China",
    publisher = "Association for Computational Linguistics",
    url = "https://aclanthology.org/D19-1018",
    doi = "10.18653/v1/D19-1018",
    pages = "188--197",
}
```

Dataset id: en_dai_labor

- Domain: **social_media**
- Language: **en**
- Language 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**

```
@inproceedings{dataset_dai_labor,
    author = {Narr, Sascha  and Michael Hülfenhaus and  Albayrak, Sahin},
    title = {Language-Independent Twitter Sentiment Analysis},
    booktitle = {Workshop on Knowledge Discovery, Data Mining and Machine Learning (KDML-2012)
    year = {2012},
```

```
    location = {Dortmund, Germany},
}
```

Dataset id: en_financial_phrasebank_sentences_75agree

- Domain: **news**
- Language: **en**
- Language 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**

```
@article{dataset_en_financial_phrasebank_sentences_75agree,
    author = {Malo, Pekka and Sinha, Ankur and Korhonen, Pekka and Wallenius, Jyrki and Takala
    title = {Good Debt or Bad Debt: Detecting Semantic Orientations in Economic Texts},
    year = {2014},
    issue_date = {April 2014},
    publisher = {John Wiley &amp;amp; Sons, Inc.},
    address = {USA},
    volume = {65},
    number = {4},
    issn = {2330-1635},
    url = {https://doi.org/10.1002/asi.23062},
    doi = {10.1002/asi.23062},
    journal = {Journal of the Association for Information Science and Technology},
    month = {apr},
    pages = {782–796},
    numpages = {15},
    keywords = {economics, automatic classification, linguistic analysis}
}
```

Dataset id: en_multilan_amazon

- Domain: **reviews**
- Language: **en**
- Language 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**

```
@inproceedings{dataset_multilan_amazon,
    title = "The Multilingual {A}mazon Reviews Corpus",
    author = {Keung, Phillip  and
        Lu, Yichao  and
        Szarvas, Gy{\"o}rgy  and
        Smith, Noah A.},
    booktitle = "Proceedings of the 2020 Conference on Empirical Methods in Natural Language P
    month = nov,
    year = "2020",
    address = "Online",
    publisher = "Association for Computational Linguistics",
    url = "https://aclanthology.org/2020.emnlp-main.369",
    doi = "10.18653/v1/2020.emnlp-main.369",
    pages = "4563--4568",
}
```

Dataset id: en_per_sent

- Domain: **news**
- Language: **en**
- Language 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**

```
@inproceedings{dataset_en_per_sent,
    title = "Author{'}s Sentiment Prediction",
    author = "Bastan, Mohaddeseh  and
        Koupaee, Mahnaz  and
        Son, Youngseo  and
        Sicoli, Richard  and
```

```
        Balasubramanian, Niranjan",
    booktitle = "Proceedings of the 28th International Conference on Computational Linguistics
    month = dec,
    year = "2020",
    address = "Barcelona, Spain (Online)",
    publisher = "International Committee on Computational Linguistics",
    url = "https://aclanthology.org/2020.coling-main.52",
    doi = "10.18653/v1/2020.coling-main.52",
    pages = "604--615",
}
```

Dataset id: en_poem_sentiment

- Domain: **poems**
- Language: **en**
- Language 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**

```
@inproceedings{dataset_en_poem_sentiment,
    title = "Investigating Societal Biases in a Poetry Composition System",
    author = "Sheng, Emily  and
        Uthus, David",
    booktitle = "Proceedings of the Second Workshop on Gender Bias in Natural Language Process
    month = dec,
    year = "2020",
    address = "Barcelona, Spain (Online)",
    publisher = "Association for Computational Linguistics",
    url = "https://aclanthology.org/2020.gebnlp-1.9",
    pages = "93--106",
}
```

Dataset id: en_semeval_2017

- Domain: **mixed**
- Language: **en**
- Language 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**

```
@inproceedings{dataset_semeval_2017,
    title = "{S}em{E}val-2017 Task 4: Sentiment Analysis in {T}witter",
    author = "Rosenthal, Sara  and
        Farra, Noura  and
        Nakov, Preslav",
    booktitle = "Proceedings of the 11th International Workshop on Semantic Evaluation ({S}em{
    month = aug,
    year = "2017",
    address = "Vancouver, Canada",
    publisher = "Association for Computational Linguistics",
    url = "https://aclanthology.org/S17-2088",
    doi = "10.18653/v1/S17-2088",
    pages = "502--518",
    abstract = "This paper describes the fifth year of the Sentiment Analysis in Twitter task.
}
```

Dataset id: en_sentistrength

- Domain: **social_media**
- Language: **en**
- Language 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**

```
@article{dataset_en_sentistrength,
    author = {Thelwall, Mike and Buckley, Kevan and Paltoglou, Georgios},
    title = {Sentiment Strength Detection for the Social Web},
    year = {2012},
```

```
    issue_date = {January 2012},
    publisher = {John Wiley \& Sons, Inc.},
    address = {USA},
    volume = {63},
    number = {1},
    issn = {1532-2882},
    url = {https://doi.org/10.1002/asi.21662},
    doi = {10.1002/asi.21662},
    abstract = {Sentiment analysis is concerned with the automatic extraction of sentiment-rel
    information from text. Although most sentiment analysis addresses commercial tass,
    such as extracting opinions from product reviews, there is increasing interest in
    the affective dimension of the social web, and Twitter in particular. Most sentiment
    analysis algorithms are not ideally suited to this task because they exploit indirect
    indicators of sentiment that can reflect genre or topic instead. Hence, such algorithms
    used to process social web texts can identify spurious sentiment patterns caused by
    topics rather than affective phenomena. This article assesses an improved version
    of the algorithm SentiStrength for sentiment strength detection across the social
    web that primarily uses direct indications of sentiment. The results from six diverse
    social web data sets (MySpace, Twitter, YouTube, Digg, RunnersWorld, BBCForums) indicate
    that SentiStrength 2 is successful in the sense of performing better than a baseline
    approach for all data sets in both supervised and unsupervised cases. SentiStrength
    is not always better than machine-learning approaches that exploit indirect indicators
    of sentiment, however, and is particularly weaker for positive sentiment in news-related
    discussions. Overall, the results suggest that, even unsupervised, SentiStrength is
    robust enough to be applied to a wide variety of different social web contexts.},
    journal = {J. Am. Soc. Inf. Sci. Technol.},
    month = jan,
    pages = {163–173},
    numpages = {11}
}
```

Dataset id: en_silicone_meld_s

- Domain: **chats**
- Language: **en**
- Language 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**

```
@inproceedings{dataset_en_silicone,
    title = "Hierarchical Pre-training for Sequence Labelling in Spoken Dialog",
```

```
    author = "Chapuis, Emile  and
        Colombo, Pierre  and
        Manica, Matteo  and
        Labeau, Matthieu  and
        Clavel, Chlo{\'e}",
    booktitle = "Findings of the Association for Computational Linguistics: EMNLP 2020",
    month = nov,
    year = "2020",
    address = "Online",
    publisher = "Association for Computational Linguistics",
    url = "https://aclanthology.org/2020.findings-emnlp.239",
    doi = "10.18653/v1/2020.findings-emnlp.239",
    pages = "2636--2648",
}
```

Dataset id: en_silicone_sem

- Domain: **chats**
- Language: **en**
- Language 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**

```
@inproceedings{dataset_en_silicone,
    title = "Hierarchical Pre-training for Sequence Labelling in Spoken Dialog",
    author = "Chapuis, Emile  and
        Colombo, Pierre  and
        Manica, Matteo  and
        Labeau, Matthieu  and
        Clavel, Chlo{\'e}",
    booktitle = "Findings of the Association for Computational Linguistics: EMNLP 2020",
    month = nov,
    year = "2020",
    address = "Online",
    publisher = "Association for Computational Linguistics",
    url = "https://aclanthology.org/2020.findings-emnlp.239",
    doi = "10.18653/v1/2020.findings-emnlp.239",
    pages = "2636--2648",
}
```

Dataset id: en_tweet_airlines

- Domain: **social_media**
- Language: **en**
- Language 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**

```
@misc{dataset_en_tweet_airlines,
    url={https://www.kaggle.com/crowdflower/twitter-airline-sentiment},
    author={Crowdflower Inc.},
    title={Twitter US Airline Sentiment},
    year={2015}
}
```

Dataset id: en_tweets_sanders

- Domain: **social_media**
- Language: **en**
- Language 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**

```
@article{dataset_en_tweets_sanders,
    title={{Sanders-Twitter Sentiment Corpus}},
    author={Sanders, Niek J},
    journal={Sanders Analytics LLC},
    year={2011}
}
```

Dataset id: en_twitter_sentiment

- Domain: **social_media**
- Language: **en**
- Language 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**

```
@article{dataset_twitter_sentiment,
    doi = {10.1371/journal.pone.0155036},
    author = {Mozetič, Igor AND Grčar, Miha AND Smailović, Jasmina},
    journal = {PLOS ONE},
    publisher = {Public Library of Science},
    title = {Multilingual Twitter Sentiment Classification: The Role of Human Annotators},
    year = {2016},
    month = {05},
    volume = {11},
    url = {https://doi.org/10.1371/journal.pone.0155036},
    pages = {1-26},
    number = {5},
}
```

Dataset id: en_vader_amazon

- Domain: **reviews**
- Language: **en**
- Language 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**

```
@inproceedings{dataset_en_vader,
    title={{VADER}: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Tex
    author={Clayton J. Hutto and Eric Gilbert},
    booktitle={Proceedings of the International AAAI Conference on Web and Social Media},
    year={2014},
    url={https://ojs.aaai.org/index.php/ICWSM/article/view/14550},
    month={May},
    pages={216-225},
    volume=8,
}
```

Dataset id: en_vader_movie_reviews

- Domain: **reviews**
- Language: **en**
- Language 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**

```
@inproceedings{dataset_en_vader,
    title={{VADER}: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Tex
    author={Clayton J. Hutto and Eric Gilbert},
    booktitle={Proceedings of the International AAAI Conference on Web and Social Media},
    year={2014},
    url={https://ojs.aaai.org/index.php/ICWSM/article/view/14550},
    month={May},
    pages={216-225},
    volume=8,
}
```

Dataset id: en_vader_nyt

- Domain: **news**
- Language: **en**
- Language 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**

```
@inproceedings{dataset_en_vader,
    title={{VADER}: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Tex
    author={Clayton J. Hutto and Eric Gilbert},
    booktitle={Proceedings of the International AAAI Conference on Web and Social Media},
    year={2014},
    url={https://ojs.aaai.org/index.php/ICWSM/article/view/14550},
    month={May},
    pages={216-225},
    volume=8,
}
```

Dataset id: en_vader_twitter

- Domain: **social_media**
- Language: **en**
- Language 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**

```
@inproceedings{dataset_en_vader,
    title={{VADER}: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Tex
    author={Clayton J. Hutto and Eric Gilbert},
    booktitle={Proceedings of the International AAAI Conference on Web and Social Media},
    year={2014},
    url={https://ojs.aaai.org/index.php/ICWSM/article/view/14550},
    month={May},
    pages={216-225},
    volume=8,
```

```
}
```

Dataset id: es_muchocine

- Domain: **reviews**
- Language: **es**
- Language family: **Indo-European**
- Genus: **Romance**
- Definite articles: **definite word distinct from demonstrative**
- Indefinite articles: **indefinite word same as one**
- Number of cases: **no morphological case-making**
- 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: **masculine, feminine**

```
@article{dataset_es_muchocine,
    title={Experiments in sentiment classification of movie reviews in Spanish},
    author={Cruz, Fermin L and Troyano, Jose A and Enriquez, Fernando and Ortega, Javier},
    journal={Procesamiento del Lenguaje Natural},
    volume={41},
    pages={73--80},
    year={2008},
    publisher={SOC ESPANOLA PROCESAMIENTO LENGUAJE NATURAL-SEPLN DEPT LENGUAJES \& SISTEMAS~…}
}
```

Dataset id: es_multilan_amazon

- Domain: **reviews**
- Language: **es**
- Language family: **Indo-European**
- Genus: **Romance**
- Definite articles: **definite word distinct from demonstrative**
- Indefinite articles: **indefinite word same as one**
- Number of cases: **no morphological case-making**
- 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: **masculine, feminine**

```
@inproceedings{dataset_multilan_amazon,
    title = "The Multilingual {A}mazon Reviews Corpus",
    author = {Keung, Phillip  and
        Lu, Yichao  and
        Szarvas, Gy{\"o}rgy  and
        Smith, Noah A.},
    booktitle = "Proceedings of the 2020 Conference on Empirical Methods in Natural Language P
    month = nov,
    year = "2020",
    address = "Online",
    publisher = "Association for Computational Linguistics",
    url = "https://aclanthology.org/2020.emnlp-main.369",
    doi = "10.18653/v1/2020.emnlp-main.369",
    pages = "4563--4568",
}
```

Dataset id: es_paper_reviews

- Domain: **reviews**
- Language: **es**
- Language family: **Indo-European**
- Genus: **Romance**
- Definite articles: **definite word distinct from demonstrative**
- Indefinite articles: **indefinite word same as one**
- Number of cases: **no morphological case-making**
- 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: **masculine, feminine**

```
@article{dataset_es_paper_reviews,
    author = {Keith Norambuena, Brian and Lettura, Exequiel and Villegas, Claudio},
    year = {2019},
    month = {02},
    pages = {191-214},
    title = {Sentiment analysis and opinion mining applied to scientific paper reviews},
    volume = {23},
    journal = {Intelligent Data Analysis},
    doi = {10.3233/IDA-173807}
}
```

Dataset id: es_semeval2020

- Domain: **social_media**
- Language: **es**

- Language family: **Indo-European**
- Genus: **Romance**
- Definite articles: **definite word distinct from demonstrative**
- Indefinite articles: **indefinite word same as one**
- Number of cases: **no morphological case-making**
- 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: **masculine, feminine**

```
@inproceedings{dataset_semeval_2020,
    title = "{S}em{E}val-2020 Task 9: Overview of Sentiment Analysis of Code-Mixed Tweets",
    author = {Patwa, Parth  and
        Aguilar, Gustavo  and
        Kar, Sudipta  and
        Pandey, Suraj  and
        PYKL, Srinivas  and
        Gamb{\"a}ck, Bj{\"o}rn  and
        Chakraborty, Tanmoy  and
        Solorio, Thamar  and
        Das, Amitava},
    booktitle = "Proceedings of the Fourteenth Workshop on Semantic Evaluation",
    month = dec,
    year = "2020",
    address = "Barcelona (online)",
    publisher = "International Committee for Computational Linguistics",
    url = "https://aclanthology.org/2020.semeval-1.100",
    doi = "10.18653/v1/2020.semeval-1.100",
    pages = "774--790",
}
```

Dataset id: es_twitter_sentiment

- Domain: **social_media**
- Language: **es**
- Language family: **Indo-European**
- Genus: **Romance**
- Definite articles: **definite word distinct from demonstrative**
- Indefinite articles: **indefinite word same as one**
- Number of cases: **no morphological case-making**
- 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: **masculine, feminine**

```
@article{dataset_twitter_sentiment,
    doi = {10.1371/journal.pone.0155036},
    author = {Mozetič, Igor AND Grčar, Miha AND Smailović, Jasmina},
    journal = {PLOS ONE},
    publisher = {Public Library of Science},
    title = {Multilingual Twitter Sentiment Classification: The Role of Human Annotators},
    year = {2016},
    month = {05},
    volume = {11},
    url = {https://doi.org/10.1371/journal.pone.0155036},
    pages = {1-26},
    number = {5},
}
```

Dataset id: fa_sentipers

- Domain: **reviews**
- Language: **fa**
- Language family: **Indo-European**
- Genus: **Iranian**
- Definite articles: **no article**
- Indefinite articles: **indefinite word same as one**
- Number of cases: **2**
- Order of subject, object, verb: **SOV**
- Negative morphemes: **negative affix**
- Polar questions: **question particle**
- Position of negative word wrt SOV: **MorphNeg**
- Prefixing vs suffixing: **weakly suffixing**
- Coding of nominal plurality: **plural suffix**
- Grammatical genders: **no grammatical gender**

```
@article{dataset_fa_sentipers,
    author    = {Pedram Hosseini and
                 Ali Ahmadian Ramaki and
                 Hassan Maleki and
                 Mansoureh Anvari and
                 Seyed Abolghasem Mirroshandel},
    title     = {{SentiPers}: {A} Sentiment Analysis Corpus for Persian},
    journal   = {Computing Research Repository},
    volume    = {arXiv:1801.07737},
    note = {Version 2},
    year      = {2018},
    url       = {http://arxiv.org/abs/1801.07737},
    eprinttype = {arXiv},
    eprint    = {1801.07737},
    timestamp = {Mon, 13 Aug 2018 16:47:47 +0200},
    biburl    = {https://dblp.org/rec/journals/corr/abs-1801-07737.bib},
    bibsource = {dblp computer science bibliography, https://dblp.org}
```

```
}
```

Dataset id: fr_dai_labor

- Domain: **social_media**
- Language: **fr**
- Language family: **Indo-European**
- Genus: **Romance**
- Definite articles: **definite word distinct from demonstrative**
- Indefinite articles: **indefinite word same as one**
- Number of cases: **no morphological case-making**
- Order of subject, object, verb: **SVO**
- Negative morphemes: **negative particle**
- Polar questions: **question particle**
- Position of negative word wrt SOV: **OptDoubleNeg**
- Prefixing vs suffixing: **strongly suffixing**
- Coding of nominal plurality: **plural suffix**
- Grammatical genders: **masculine, feminine**

```
@inproceedings{dataset_dai_labor,
    author = {Narr, Sascha  and Michael Hülfenhaus and  Albayrak, Sahin},
    title = {Language-Independent Twitter Sentiment Analysis},
    booktitle = {Workshop on Knowledge Discovery, Data Mining and Machine Learning (KDML-2012)
    year = {2012},
    location = {Dortmund, Germany},
}
```

Dataset id: fr_ifeel

- Domain: **social_media**
- Language: **fr**
- Language family: **Indo-European**
- Genus: **Romance**
- Definite articles: **definite word distinct from demonstrative**
- Indefinite articles: **indefinite word same as one**
- Number of cases: **no morphological case-making**
- Order of subject, object, verb: **SVO**
- Negative morphemes: **negative particle**
- Polar questions: **question particle**
- Position of negative word wrt SOV: **OptDoubleNeg**
- Prefixing vs suffixing: **strongly suffixing**
- Coding of nominal plurality: **plural suffix**
- Grammatical genders: **masculine, feminine**

```
@inproceedings{dataset_dai_labor,
    author = {Narr, Sascha  and Michael Hülfenhaus and  Albayrak, Sahin},
```

```
    title = {Language-Independent Twitter Sentiment Analysis},
    booktitle = {Workshop on Knowledge Discovery, Data Mining and Machine Learning (KDML-2012)
    year = {2012},
    location = {Dortmund, Germany},
}
```

Dataset id: fr_multilan_amazon

- Domain: **reviews**
- Language: **fr**
- Language family: **Indo-European**
- Genus: **Romance**
- Definite articles: **definite word distinct from demonstrative**
- Indefinite articles: **indefinite word same as one**
- Number of cases: **no morphological case-making**
- Order of subject, object, verb: **SVO**
- Negative morphemes: **negative particle**
- Polar questions: **question particle**
- Position of negative word wrt SOV: **OptDoubleNeg**
- Prefixing vs suffixing: **strongly suffixing**
- Coding of nominal plurality: **plural suffix**
- Grammatical genders: **masculine, feminine**

```
@inproceedings{dataset_multilan_amazon,
    title = "The Multilingual {A}mazon Reviews Corpus",
    author = {Keung, Phillip  and
        Lu, Yichao  and
        Szarvas, Gy{\"o}rgy  and
        Smith, Noah A.},
    booktitle = "Proceedings of the 2020 Conference on Empirical Methods in Natural Language P
    month = nov,
    year = "2020",
    address = "Online",
    publisher = "Association for Computational Linguistics",
    url = "https://aclanthology.org/2020.emnlp-main.369",
    doi = "10.18653/v1/2020.emnlp-main.369",
    pages = "4563--4568",
}
```

Dataset id: he_hebrew_sentiment

- Domain: **social_media**
- Language: **he**
- Language family: **Afro-Asiatic**
- Genus: **Semitic**
- Definite articles: **definite affix**
- Indefinite articles: **indefinite word same as one**

- Number of cases: **no morphological case-making**
- Order of subject, object, verb: **SVO**
- Negative morphemes: **negative particle**
- Polar questions: **question particle**
- Position of negative word wrt SOV: **SNegVO**
- Prefixing vs suffixing: **weakly suffixing**
- Coding of nominal plurality: **plural suffix**
- Grammatical genders: **masculine, feminine**

```
@inproceedings{dataset_he_hebrew_sentiment,
    title = "Representations and Architectures in Neural Sentiment Analysis for Morphologicall
    author = "Amram, Adam  and
        Ben David, Anat  and
        Tsarfaty, Reut",
    booktitle = "Proceedings of the 27th International Conference on Computational Linguistics
    month = aug,
    year = "2018",
    address = "Santa Fe, New Mexico, USA",
    publisher = "Association for Computational Linguistics",
    url = "https://aclanthology.org/C18-1190",
    pages = "2242--2252",
    abstract = "This paper empirically studies the effects of representation choices on neural
}
```

Dataset id: hi_semeval2020

- Domain: **social_media**
- Language: **hi**
- Language family: **Indo-European**
- Genus: **Indic**
- Definite articles: **no article**
- Indefinite articles: **no article**
- Number of cases: **3**
- Order of subject, object, verb: **SOV**
- Negative morphemes: **negative particle**
- Polar questions: **question particle**
- Position of negative word wrt SOV: **SONegV**
- Prefixing vs suffixing: **strongly suffixing**
- Coding of nominal plurality: **plural suffix**
- Grammatical genders: **masculine, feminine**

```
@inproceedings{dataset_semeval_2020,
    title = "{S}em{E}val-2020 Task 9: Overview of Sentiment Analysis of Code-Mixed Tweets",
    author = {Patwa, Parth  and
        Aguilar, Gustavo  and
        Kar, Sudipta  and
        Pandey, Suraj  and
        PYKL, Srinivas  and
```

```
        Gamb{\"a}ck, Bj{\"o}rn  and
        Chakraborty, Tanmoy  and
        Solorio, Thamar  and
        Das, Amitava},
    booktitle = "Proceedings of the Fourteenth Workshop on Semantic Evaluation",
    month = dec,
    year = "2020",
    address = "Barcelona (online)",
    publisher = "International Committee for Computational Linguistics",
    url = "https://aclanthology.org/2020.semeval-1.100",
    doi = "10.18653/v1/2020.semeval-1.100",
    pages = "774--790",
}
```

Dataset id: hr_sentiment_news_document

- Domain: **news**
- Language: **hr**
- Language family: **Indo-European**
- Genus: **Slavic**
- Definite articles: **no article**
- Indefinite articles: **no article**
- Number of cases: **5**
- Order of subject, object, verb: **SVO**
- Negative morphemes: **negative particle**
- Polar questions: **question particle**
- Position of negative word wrt SOV: **other**
- Prefixing vs suffixing: **strongly suffixing**
- Coding of nominal plurality: **plural suffix**
- Grammatical genders: **masculine, feminine, neuter**

```
@Article{dataset_hr_sentiment_news_document,
    AUTHOR = {Pelicon, Andraž and Pranjić, Marko and Miljković, Dragana and Škrlj, Blaž and Po
    TITLE = {Zero-Shot Learning for Cross-Lingual News Sentiment Classification},
    JOURNAL = {Applied Sciences},
    VOLUME = {10},
    YEAR = {2020},
    NUMBER = {17},
    ARTICLE-NUMBER = {5993},
    URL = {https://www.mdpi.com/2076-3417/10/17/5993},
    ISSN = {2076-3417},
    ABSTRACT = {In this paper, we address the task of zero-shot cross-lingual news sentiment c
    DOI = {10.3390/app10175993}
}
```

Dataset id: hr_twitter_sentiment

- Domain: **social_media**

- Language: **hr**
- Language family: **Indo-European**
- Genus: **Slavic**
- Definite articles: **no article**
- Indefinite articles: **no article**
- Number of cases: **5**
- Order of subject, object, verb: **SVO**
- Negative morphemes: **negative particle**
- Polar questions: **question particle**
- Position of negative word wrt SOV: **other**
- Prefixing vs suffixing: **strongly suffixing**
- Coding of nominal plurality: **plural suffix**
- Grammatical genders: **masculine, feminine, neuter**

```
@article{dataset_twitter_sentiment,
    doi = {10.1371/journal.pone.0155036},
    author = {Mozetič, Igor AND Grčar, Miha AND Smailović, Jasmina},
    journal = {PLOS ONE},
    publisher = {Public Library of Science},
    title = {Multilingual Twitter Sentiment Classification: The Role of Human Annotators},
    year = {2016},
    month = {05},
    volume = {11},
    url = {https://doi.org/10.1371/journal.pone.0155036},
    pages = {1-26},
    number = {5},
}
```

Dataset id: hu_twitter_sentiment

- Domain: **social_media**
- Language: **hu**
- Language family: **Uralic**
- Genus: **Ugric**
- Definite articles: **definite word distinct from demonstrative**
- Indefinite articles: **indefinite word distinct from one**
- Number of cases: **10 or more**
- Order of subject, object, verb: **no dominant order**
- Negative morphemes: **negative particle**
- Polar questions: **question particle**
- Position of negative word wrt SOV: **SNegVO**
- Prefixing vs suffixing: **strongly suffixing**
- Coding of nominal plurality: **plural suffix**
- Grammatical genders: **no grammatical gender**

```
@article{dataset_twitter_sentiment,
    doi = {10.1371/journal.pone.0155036},
    author = {Mozetič, Igor AND Grčar, Miha AND Smailović, Jasmina},
```

```
    journal = {PLOS ONE},
    publisher = {Public Library of Science},
    title = {Multilingual Twitter Sentiment Classification: The Role of Human Annotators},
    year = {2016},
    month = {05},
    volume = {11},
    url = {https://doi.org/10.1371/journal.pone.0155036},
    pages = {1-26},
    number = {5},
}
```

Dataset id: it_evalita2016

- Domain: **social_media**
- Language: **it**
- Language family: **Indo-European**
- Genus: **Romance**
- Definite articles: **definite word distinct from demonstrative**
- Indefinite articles: **indefinite word same as one**
- Number of cases: **no morphological case-making**
- Order of subject, object, verb: **SVO**
- Negative morphemes: **negative particle**
- Polar questions: **interrogative intonation only**
- Position of negative word wrt SOV: **SNegVO**
- Prefixing vs suffixing: **strongly suffixing**
- Coding of nominal plurality: **plural suffix**
- Grammatical genders: **masculine, feminine**

```
@inproceedings{dataset_it_evalita2016,
    TITLE = {{Overview of the Evalita 2016 SENTIment POLarity Classification Task}},
    AUTHOR = {Barbieri, Francesco and Basile, Valerio and Croce, Danilo and Nissim, Malvina an
    URL = {https://hal.inria.fr/hal-01414731},
    BOOKTITLE = {{Proceedings of Third Italian Conference on Computational Linguistics (CLiC-i
    ADDRESS = {Naples, Italy},
    YEAR = {2016},
    MONTH = Dec,
    KEYWORDS = {Natural language processing and web ; Social media analysis ; Sentiment analys
    PDF = {https://hal.inria.fr/hal-01414731/file/paper_026.pdf},
    HAL_ID = {hal-01414731},
    HAL_VERSION = {v1},
}
```

Dataset id: it_multiemotions

- Domain: **social_media**
- Language: **it**
- Language family: **Indo-European**
- Genus: **Romance**

- Definite articles: **definite word distinct from demonstrative**
- Indefinite articles: **indefinite word same as one**
- Number of cases: **no morphological case-making**
- Order of subject, object, verb: **SVO**
- Negative morphemes: **negative particle**
- Polar questions: **interrogative intonation only**
- Position of negative word wrt SOV: **SNegVO**
- Prefixing vs suffixing: **strongly suffixing**
- Coding of nominal plurality: **plural suffix**
- Grammatical genders: **masculine, feminine**

```
@inproceedings{dataset_it_multiemotions,
    author = {Sprugnoli, Rachele},
    year = {2020},
    month = {12},
    pages = {},
    title = {MultiEmotions-It: a New Dataset for Opinion Polarity and Emotion Analysis for Ita
    booktitle = {Proceedings of the Seventh Italian Conference on Computational Linguistics},
}
```

Dataset id: ja_multilan_amazon

- Domain: **reviews**
- Language: **ja**
- Language family: **Japanese**
- Genus: **Japanese**
- Definite articles: **no article**
- Indefinite articles: **indefinite word distinct from one**
- Number of cases: **8-9**
- Order of subject, object, verb: **SOV**
- Negative morphemes: **negative affix**
- Polar questions: **question particle**
- Position of negative word wrt SOV: **MorphNeg**
- Prefixing vs suffixing: **strongly suffixing**
- Coding of nominal plurality: **plural suffix**
- Grammatical genders: **no grammatical gender**

```
@inproceedings{dataset_multilan_amazon,
    title = "The Multilingual {A}mazon Reviews Corpus",
    author = {Keung, Phillip  and
        Lu, Yichao  and
        Szarvas, Gy{\"o}rgy  and
        Smith, Noah A.},
    booktitle = "Proceedings of the 2020 Conference on Empirical Methods in Natural Language P
    month = nov,
    year = "2020",
    address = "Online",
    publisher = "Association for Computational Linguistics",
```

```
    url = "https://aclanthology.org/2020.emnlp-main.369",
    doi = "10.18653/v1/2020.emnlp-main.369",
    pages = "4563--4568",
}
```

Dataset id: lv_ltec_sentiment

- Domain: **social_media**
- Language: **lv**
- Language family: **Indo-European**
- Genus: **Baltic**
- Definite articles: **demonstrative word used as definite article**
- Indefinite articles: **indefinite word same as one**
- Number of cases: **5**
- Order of subject, object, verb: **SVO**
- Negative morphemes: **negative affix**
- Polar questions: **question particle**
- Position of negative word wrt SOV: **MorphNeg**
- Prefixing vs suffixing: **weakly suffixing**
- Coding of nominal plurality: **plural suffix**
- Grammatical genders: **masculine, feminine**

```
@article{dataset_lv_ltec_sentiment,
    author    = {Uga Sprogis and
                 Matiss Rikters},
    title     = {What Can We Learn From Almost a Decade of Food Tweets},
    journal   = {Computing Research Repository},
    volume    = {arXiv:2007.05194},
    note = {Version 2},
    year      = {2020},
    url       = {https://arxiv.org/abs/2007.05194},
    eprinttype = {arXiv},
    eprint    = {2007.05194},
    timestamp = {Mon, 20 Jul 2020 14:20:39 +0200},
    biburl    = {https://dblp.org/rec/journals/corr/abs-2007-05194.bib},
    bibsource = {dblp computer science bibliography, https://dblp.org}
}
```

Dataset id: pl_klej_allegro_reviews

- Domain: **reviews**
- Language: **pl**
- Language family: **Indo-European**
- Genus: **Slavic**
- Definite articles: **no article**
- Indefinite articles: **no article**
- Number of cases: **6-7**

- Order of subject, object, verb: **SVO**
- Negative morphemes: **negative particle**
- Polar questions: **question particle**
- Position of negative word wrt SOV: **SNegVO**
- Prefixing vs suffixing: **strongly suffixing**
- Coding of nominal plurality: **plural suffix**
- Grammatical genders: **masculine, feminine, neuter**

```
@inproceedings{dataset_pl_klej_allegro_reviews,
    title = "{KLEJ}: Comprehensive Benchmark for {P}olish Language Understanding",
    author = "Rybak, Piotr  and
        Mroczkowski, Robert  and
        Tracz, Janusz  and
        Gawlik, Ireneusz",
    booktitle = "Proceedings of the 58th Annual Meeting of the Association for Computational L
    month = jul,
    year = "2020",
    address = "Online",
    publisher = "Association for Computational Linguistics",
    url = "https://aclanthology.org/2020.acl-main.111",
    doi = "10.18653/v1/2020.acl-main.111",
    pages = "1191--1201",
}
```

Dataset id: pl_opi_lil_2012

- Domain: **social_media**
- Language: **pl**
- Language family: **Indo-European**
- Genus: **Slavic**
- Definite articles: **no article**
- Indefinite articles: **no article**
- Number of cases: **6-7**
- Order of subject, object, verb: **SVO**
- Negative morphemes: **negative particle**
- Polar questions: **question particle**
- Position of negative word wrt SOV: **SNegVO**
- Prefixing vs suffixing: **strongly suffixing**
- Coding of nominal plurality: **plural suffix**
- Grammatical genders: **masculine, feminine, neuter**

```
@article{dataset_pl_opi_lil_2012,
    author = {Pawel Sobkowicz and Antoni Sobkowicz},
    title ={Two-Year Study of Emotion and Communication Patterns in a Highly Polarized Politic
    journal = {Social Science Computer Review},
    volume = {30},
    number = {4},
    pages = {448-469},
```

```
    year = {2012},
    doi = {10.1177/0894439312436512}
}
```

Dataset id: pl_polemo

- Domain: **reviews**
- Language: **pl**
- Language family: **Indo-European**
- Genus: **Slavic**
- Definite articles: **no article**
- Indefinite articles: **no article**
- Number of cases: **6-7**
- Order of subject, object, verb: **SVO**
- Negative morphemes: **negative particle**
- Polar questions: **question particle**
- Position of negative word wrt SOV: **SNegVO**
- Prefixing vs suffixing: **strongly suffixing**
- Coding of nominal plurality: **plural suffix**
- Grammatical genders: **masculine, feminine, neuter**

```
@inproceedings{dataset_pl_polemo,
    title = "Multi-Level Sentiment Analysis of {P}ol{E}mo 2.0: Extended Corpus of Multi-Domain
    author = "Koco{\'n}, Jan  and
        Mi{\l}kowski, Piotr  and
        Za{\'s}ko-Zieli{\'n}ska, Monika",
    booktitle = "Proceedings of the 23rd Conference on Computational Natural Language Learning
    month = nov,
    year = "2019",
    address = "Hong Kong, China",
    publisher = "Association for Computational Linguistics",
    url = "https://aclanthology.org/K19-1092",
    doi = "10.18653/v1/K19-1092",
    pages = "980--991"
}
```

Dataset id: pl_twitter_sentiment

- Domain: **social_media**
- Language: **pl**
- Language family: **Indo-European**
- Genus: **Slavic**
- Definite articles: **no article**
- Indefinite articles: **no article**
- Number of cases: **6-7**
- Order of subject, object, verb: **SVO**

- Negative morphemes: **negative particle**
- Polar questions: **question particle**
- Position of negative word wrt SOV: **SNegVO**
- Prefixing vs suffixing: **strongly suffixing**
- Coding of nominal plurality: **plural suffix**
- Grammatical genders: **masculine, feminine, neuter**

```
@article{dataset_twitter_sentiment,
    doi = {10.1371/journal.pone.0155036},
    author = {Mozetič, Igor AND Grčar, Miha AND Smailović, Jasmina},
    journal = {PLOS ONE},
    publisher = {Public Library of Science},
    title = {Multilingual Twitter Sentiment Classification: The Role of Human Annotators},
    year = {2016},
    month = {05},
    volume = {11},
    url = {https://doi.org/10.1371/journal.pone.0155036},
    pages = {1-26},
    number = {5},
}
```

Dataset id: pt_dai_labor

- Domain: **social_media**
- Language: **pt**
- Language family: **Indo-European**
- Genus: **Romance**
- Definite articles: **definite word distinct from demonstrative**
- Indefinite articles: **indefinite word same as one**
- Number of cases: **no morphological case-making**
- Order of subject, object, verb: **SVO**
- Negative morphemes: **negative particle**
- Polar questions: **question particle**
- Position of negative word wrt SOV: **SNegVO**
- Prefixing vs suffixing: **strongly suffixing**
- Coding of nominal plurality: **plural suffix**
- Grammatical genders: **masculine, feminine**

```
@inproceedings{dataset_dai_labor,
    author = {Narr, Sascha  and Michael Hülfenhaus and  Albayrak, Sahin},
    title = {Language-Independent Twitter Sentiment Analysis},
    booktitle = {Workshop on Knowledge Discovery, Data Mining and Machine Learning (KDML-2012)
    year = {2012},
    location = {Dortmund, Germany},
}
```

Dataset id: pt_ifeel

- Domain: **social_media**
- Language: **pt**
- Language family: **Indo-European**
- Genus: **Romance**
- Definite articles: **definite word distinct from demonstrative**
- Indefinite articles: **indefinite word same as one**
- Number of cases: **no morphological case-making**
- Order of subject, object, verb: **SVO**
- Negative morphemes: **negative particle**
- Polar questions: **question particle**
- Position of negative word wrt SOV: **SNegVO**
- Prefixing vs suffixing: **strongly suffixing**
- Coding of nominal plurality: **plural suffix**
- Grammatical genders: **masculine, feminine**

```
@inproceedings{dataset_dai_labor,
    author = {Narr, Sascha  and Michael Hülfenhaus and  Albayrak, Sahin},
    title = {Language-Independent Twitter Sentiment Analysis},
    booktitle = {Workshop on Knowledge Discovery, Data Mining and Machine Learning (KDML-2012)
    year = {2012},
    location = {Dortmund, Germany},
}
```

Dataset id: pt_tweet_sent_br

- Domain: **social_media**
- Language: **pt**
- Language family: **Indo-European**
- Genus: **Romance**
- Definite articles: **definite word distinct from demonstrative**
- Indefinite articles: **indefinite word same as one**
- Number of cases: **no morphological case-making**
- Order of subject, object, verb: **SVO**
- Negative morphemes: **negative particle**
- Polar questions: **question particle**
- Position of negative word wrt SOV: **SNegVO**
- Prefixing vs suffixing: **strongly suffixing**
- Coding of nominal plurality: **plural suffix**
- Grammatical genders: **masculine, feminine**

```
@inproceedings{dataset_pt_tweet_sent_br,
    title = "Building a Sentiment Corpus of Tweets in {B}razilian {P}ortuguese",
    author = "Brum, Henrico  and
        Volpe Nunes, Maria das Gra{\c{c}}as",
    booktitle = "Proceedings of the Eleventh International Conference on Language Resources an
    month = may,
    year = "2018",
```

```
    address = "Miyazaki, Japan",
    publisher = "European Language Resources Association (ELRA)",
    url = "https://aclanthology.org/L18-1658",
}
```

Dataset id: pt_twitter_sentiment

- Domain: **social_media**
- Language: **pt**
- Language family: **Indo-European**
- Genus: **Romance**
- Definite articles: **definite word distinct from demonstrative**
- Indefinite articles: **indefinite word same as one**
- Number of cases: **no morphological case-making**
- Order of subject, object, verb: **SVO**
- Negative morphemes: **negative particle**
- Polar questions: **question particle**
- Position of negative word wrt SOV: **SNegVO**
- Prefixing vs suffixing: **strongly suffixing**
- Coding of nominal plurality: **plural suffix**
- Grammatical genders: **masculine, feminine**

```
@article{dataset_twitter_sentiment,
    doi = {10.1371/journal.pone.0155036},
    author = {Mozetič, Igor AND Grčar, Miha AND Smailović, Jasmina},
    journal = {PLOS ONE},
    publisher = {Public Library of Science},
    title = {Multilingual Twitter Sentiment Classification: The Role of Human Annotators},
    year = {2016},
    month = {05},
    volume = {11},
    url = {https://doi.org/10.1371/journal.pone.0155036},
    pages = {1-26},
    number = {5},
}
```

Dataset id: ru_sentiment

- Domain: **social_media**
- Language: **ru**
- Language family: **Indo-European**
- Genus: **Slavic**
- Definite articles: **no article**
- Indefinite articles: **no article**
- Number of cases: **6-7**
- Order of subject, object, verb: **SVO**
- Negative morphemes: **negative particle**

- Polar questions: **question particle**
- Position of negative word wrt SOV: **SNegVO**
- Prefixing vs suffixing: **strongly suffixing**
- Coding of nominal plurality: **plural suffix**
- Grammatical genders: **masculine, feminine, neuter**

```
@inproceedings{dataset_ru_sentiment,
    title = "{R}u{S}entiment: An Enriched Sentiment Analysis Dataset for Social Media in {R}us
    author = "Rogers, Anna  and
        Romanov, Alexey  and
        Rumshisky, Anna  and
        Volkova, Svitlana  and
        Gronas, Mikhail  and
        Gribov, Alex",
    booktitle = "Proceedings of the 27th International Conference on Computational Linguistics
    month = aug,
    year = "2018",
    address = "Santa Fe, New Mexico, USA",
    publisher = "Association for Computational Linguistics",
    url = "https://aclanthology.org/C18-1064",
    pages = "755--763",
    abstract = "This paper presents RuSentiment, a new dataset for sentiment analysis of socia
}
```

Dataset id: ru_twitter_sentiment

- Domain: **social_media**
- Language: **ru**
- Language family: **Indo-European**
- Genus: **Slavic**
- Definite articles: **no article**
- Indefinite articles: **no article**
- Number of cases: **6-7**
- Order of subject, object, verb: **SVO**
- Negative morphemes: **negative particle**
- Polar questions: **question particle**
- Position of negative word wrt SOV: **SNegVO**
- Prefixing vs suffixing: **strongly suffixing**
- Coding of nominal plurality: **plural suffix**
- Grammatical genders: **masculine, feminine, neuter**

```
@article{dataset_twitter_sentiment,
    doi = {10.1371/journal.pone.0155036},
    author = {Mozetič, Igor AND Grčar, Miha AND Smailović, Jasmina},
    journal = {PLOS ONE},
    publisher = {Public Library of Science},
    title = {Multilingual Twitter Sentiment Classification: The Role of Human Annotators},
    year = {2016},
    month = {05},
```

```
    volume = {11},
    url = {https://doi.org/10.1371/journal.pone.0155036},
    pages = {1-26},
    number = {5},
}
```

Dataset id: sk_twitter_sentiment

- Domain: **social_media**
- Language: **sk**
- Language family: **Indo-European**
- Genus: **Slavic**
- Definite articles: **no article**
- Indefinite articles: **no article**
- Number of cases: **6-7**
- Order of subject, object, verb: **SVO**
- Negative morphemes: **negative affix**
- Polar questions: **interrogative word order**
- Position of negative word wrt SOV: **MorphNeg**
- Prefixing vs suffixing: **weakly suffixing**
- Coding of nominal plurality: **plural suffix**
- Grammatical genders: **masculine, feminine, neuter**

```
@article{dataset_twitter_sentiment,
    doi = {10.1371/journal.pone.0155036},
    author = {Mozetič, Igor AND Grčar, Miha AND Smailović, Jasmina},
    journal = {PLOS ONE},
    publisher = {Public Library of Science},
    title = {Multilingual Twitter Sentiment Classification: The Role of Human Annotators},
    year = {2016},
    month = {05},
    volume = {11},
    url = {https://doi.org/10.1371/journal.pone.0155036},
    pages = {1-26},
    number = {5},
}
```

Dataset id: sl_sentinews

- Domain: **news**
- Language: **sl**
- Language family: **Indo-European**
- Genus: **Slavic**
- Definite articles: **no article**
- Indefinite articles: **no article**
- Number of cases: **6-7**
- Order of subject, object, verb: **SVO**
- Negative morphemes: **negative particle**

- Polar questions: **question particle**
- Position of negative word wrt SOV: **SNegVO**
- Prefixing vs suffixing: **strongly suffixing**
- Coding of nominal plurality: **plural suffix**
- Grammatical genders: **masculine, feminine, neuter**

```
@Article{Bučar2018,
    author={Bu{\v{c}}ar, Jo{\v{z}}e
    and {\v{Z}}nidar{\v{s}}i{\v{c}}, Martin
    and Povh, Janez},
    title={Annotated news corpora and a lexicon for sentiment analysis in Slovene},
    journal={Language Resources and Evaluation},
    year={2018},
    month={Sep},
    day={01},
    volume={52},
    number={3},
    pages={895-919},
    abstract={In this study, we introduce Slovene web-crawled news corpora with sentiment anno
    issn={1574-0218},
    doi={10.1007/s10579-018-9413-3},
    url={https://doi.org/10.1007/s10579-018-9413-3}
}
```

Dataset id: sl_twitter_sentiment

- Domain: **social_media**
- Language: **sl**
- Language family: **Indo-European**
- Genus: **Slavic**
- Definite articles: **no article**
- Indefinite articles: **no article**
- Number of cases: **6-7**
- Order of subject, object, verb: **SVO**
- Negative morphemes: **negative particle**
- Polar questions: **question particle**
- Position of negative word wrt SOV: **SNegVO**
- Prefixing vs suffixing: **strongly suffixing**
- Coding of nominal plurality: **plural suffix**
- Grammatical genders: **masculine, feminine, neuter**

```
@article{dataset_twitter_sentiment,
    doi = {10.1371/journal.pone.0155036},
    author = {Mozetič, Igor AND Grčar, Miha AND Smailović, Jasmina},
    journal = {PLOS ONE},
    publisher = {Public Library of Science},
    title = {Multilingual Twitter Sentiment Classification: The Role of Human Annotators},
    year = {2016},
    month = {05},
```

```
    volume = {11},
    url = {https://doi.org/10.1371/journal.pone.0155036},
    pages = {1-26},
    number = {5},
}
```

Dataset id: sq_twitter_sentiment

- Domain: **social_media**
- Language: **sq**
- Language family: **Indo-European**
- Genus: **Albanian**
- Definite articles: **definite affix**
- Indefinite articles: **indefinite word distinct from one**
- Number of cases: **4**
- Order of subject, object, verb: **SVO**
- Negative morphemes: **negative particle**
- Polar questions: **question particle**
- Position of negative word wrt SOV: **SNegVO**
- Prefixing vs suffixing: **strongly suffixing**
- Coding of nominal plurality: **plural suffix**
- Grammatical genders: **masculine, feminine**

```
@article{dataset_twitter_sentiment,
    doi = {10.1371/journal.pone.0155036},
    author = {Mozetič, Igor AND Grčar, Miha AND Smailović, Jasmina},
    journal = {PLOS ONE},
    publisher = {Public Library of Science},
    title = {Multilingual Twitter Sentiment Classification: The Role of Human Annotators},
    year = {2016},
    month = {05},
    volume = {11},
    url = {https://doi.org/10.1371/journal.pone.0155036},
    pages = {1-26},
    number = {5},
}
```

Dataset id: sr_movie_reviews

- Domain: **reviews**
- Language: **sr**
- Language family: **Indo-European**
- Genus: **Slavic**
- Definite articles: **no article**
- Indefinite articles: **no article**
- Number of cases: **5**
- Order of subject, object, verb: **SVO**
- Negative morphemes: **negative particle**

- Polar questions: **question particle**
- Position of negative word wrt SOV: **other**
- Prefixing vs suffixing: **strongly suffixing**
- Coding of nominal plurality: **plural suffix**
- Grammatical genders: **masculine, feminine, neuter**

```
@inproceedings{dataset_sr_serb_movie_reviews,
    title = "Reliable Baselines for Sentiment Analysis in Resource-Limited Languages: The {S}e
    author = "Batanovi{\'c}, Vuk  and
        Nikoli{\'c}, Bo{\v{s}}ko  and
        Milosavljevi{\'c}, Milan",
    booktitle = "Proceedings of the Tenth International Conference on Language Resources and E
    month = may,
    year = "2016",
    address = "Portoro{\v{z}}, Slovenia",
    publisher = "European Language Resources Association (ELRA)",
    url = "https://aclanthology.org/L16-1427",
    pages = "2688--2696",
    abstract = "Collecting data for sentiment analysis in resource-limited languages carries a
}
```

Dataset id: sr_senticomments

- Domain: **reviews**
- Language: **sr**
- Language family: **Indo-European**
- Genus: **Slavic**
- Definite articles: **no article**
- Indefinite articles: **no article**
- Number of cases: **5**
- Order of subject, object, verb: **SVO**
- Negative morphemes: **negative particle**
- Polar questions: **question particle**
- Position of negative word wrt SOV: **other**
- Prefixing vs suffixing: **strongly suffixing**
- Coding of nominal plurality: **plural suffix**
- Grammatical genders: **masculine, feminine, neuter**

```
@article{dataset_sr_senticomments,
    doi = {10.1371/journal.pone.0242050},
    author = {Batanović, Vuk AND Cvetanović, Miloš AND Nikolić, Boško},
    journal = {PLOS ONE},
    publisher = {Public Library of Science},
    title = {A versatile framework for resource-limited sentiment articulation, annotation, an
    year = {2020},
    month = {11},
    volume = {15},
    url = {https://doi.org/10.1371/journal.pone.0242050},
    pages = {1-30},
```

```
    abstract = {Choosing a comprehensive and cost-effective way of articulating and annotating
    number = {11},
}
```

Dataset id: sr_twitter_sentiment

- Domain: **social_media**
- Language: **sr**
- Language family: **Indo-European**
- Genus: **Slavic**
- Definite articles: **no article**
- Indefinite articles: **no article**
- Number of cases: **5**
- Order of subject, object, verb: **SVO**
- Negative morphemes: **negative particle**
- Polar questions: **question particle**
- Position of negative word wrt SOV: **other**
- Prefixing vs suffixing: **strongly suffixing**
- Coding of nominal plurality: **plural suffix**
- Grammatical genders: **masculine, feminine, neuter**

```
@article{dataset_twitter_sentiment,
    doi = {10.1371/journal.pone.0155036},
    author = {Mozetič, Igor AND Grčar, Miha AND Smailović, Jasmina},
    journal = {PLOS ONE},
    publisher = {Public Library of Science},
    title = {Multilingual Twitter Sentiment Classification: The Role of Human Annotators},
    year = {2016},
    month = {05},
    volume = {11},
    url = {https://doi.org/10.1371/journal.pone.0155036},
    pages = {1-26},
    number = {5},
}
```

Dataset id: sv_twitter_sentiment

- Domain: **social_media**
- Language: **sv**
- Language family: **Indo-European**
- Genus: **Germanic**
- Definite articles: **definite affix**
- Indefinite articles: **indefinite word same as 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: **more than one position**
- Prefixing vs suffixing: **strongly suffixing**
- Coding of nominal plurality: **plural suffix**
- Grammatical genders: **common, neuter**

```
@article{dataset_twitter_sentiment,
    doi = {10.1371/journal.pone.0155036},
    author = {Mozetič, Igor AND Grčar, Miha AND Smailović, Jasmina},
    journal = {PLOS ONE},
    publisher = {Public Library of Science},
    title = {Multilingual Twitter Sentiment Classification: The Role of Human Annotators},
    year = {2016},
    month = {05},
    volume = {11},
    url = {https://doi.org/10.1371/journal.pone.0155036},
    pages = {1-26},
    number = {5},
}
```

Dataset id: th_wisesight_sentiment

- Domain: **social_media**
- Language: **th**
- Language family: **Tai-Kadai**
- Genus: **Kam-Tai**
- Definite articles: **no article**
- Indefinite articles: **indefinite word distinct from one**
- Number of cases: **no morphological case-making**
- Order of subject, object, verb: **SVO**
- Negative morphemes: **negative auxiliary verb**
- Polar questions: **question particle**
- Position of negative word wrt SOV: **SNegVO**
- Prefixing vs suffixing: **little affixation**
- Coding of nominal plurality: **mixed morphological plural**
- Grammatical genders: **noun classifiers**

```
@misc{dataset_th_wisesight_sentiment,
    author       = {Suriyawongkul, Arthit and
                    Chuangsuwanich, Ekapol and
                    Chormai, Pattarawat and
                    Polpanumas, Charin},
    title        = {PyThaiNLP/wisesight-sentiment: First release (v1.0)},
    month        = sep,
    year         = 2019,
    publisher    = {Zenodo},
    version      = {v1.0},
    doi          = {10.5281/zenodo.3457447},
    url          = {https://doi.org/10.5281/zenodo.3457447},
    note = {Zenodo}
```

```
}
```

Dataset id: th_wongnai_reviews

- Domain: **reviews**
- Language: **th**
- Language family: **Tai-Kadai**
- Genus: **Kam-Tai**
- Definite articles: **no article**
- Indefinite articles: **indefinite word distinct from one**
- Number of cases: **no morphological case-making**
- Order of subject, object, verb: **SVO**
- Negative morphemes: **negative auxiliary verb**
- Polar questions: **question particle**
- Position of negative word wrt SOV: **SNegVO**
- Prefixing vs suffixing: **little affixation**
- Coding of nominal plurality: **mixed morphological plural**
- Grammatical genders: **noun classifiers**

```
@misc{dataset_th_wongnai_reviews,
    author = {Ekkalak Thongthanomkul and Tanapol Nearunchorn and Yuwat Chuesathuchon},
    title = {wongnai-corpus},
    year = {2019},
    publisher = {GitHub},
    journal = {GitHub repository},
    howpublished = {\url{https://github.com/wongnai/wongnai-corpus}}
}
```

Dataset id: ur_roman_urdu

- Domain: **mixed**
- Language: **ur**
- Language family: **Indo-European**
- Genus: **Indic**
- Definite articles: **no article**
- Indefinite articles: **no article**
- Number of cases: **2**
- Order of subject, object, verb: **SOV**
- Negative morphemes: **negative affix**
- Polar questions: **question particle**
- Position of negative word wrt SOV: **SONegV**
- Prefixing vs suffixing: **strongly suffixing**
- Coding of nominal plurality: **plural suffix**
- Grammatical genders: **masculine, feminine**

```
@InProceedings{dataset_ur_roman_urdu,
    title     = "Performing Natural Language Processing on Roman Urdu Datasets",
```

```
    author    = "Zareen Sharf and Saif Ur Rahman",
    booktitle = "International Journal of Computer Science and Network Security",
    volume    = "18",
    pages     = "141-148",
    year      = "2018",
    url = {http://paper.ijcsns.org/07_book/201801/20180117.pdf}
}
```

Dataset id: zh_hotel_reviews

- Domain: **reviews**
- Language: **zh**
- Language family: **Sino-Tibetan**
- Genus: **Chinese**
- Definite articles: **no article**
- Indefinite articles: **indefinite word same as one**
- Number of cases: **no morphological case-making**
- Order of subject, object, verb: **SVO**
- Negative morphemes: **negative particle**
- Polar questions: **question particle**
- Position of negative word wrt SOV: **SNegVO**
- Prefixing vs suffixing: **little affixation**
- Coding of nominal plurality: **no plural**
- Grammatical genders: **noun classifiers**

```
@inproceedings{dataset_zh_hotel_reviews,
    title = "An Empirical Study on Sentiment Classification of {C}hinese Review using Word Emb
    author = "Lin, Yiou  and
        Lei, Hang  and
        Wu, Jia  and
        Li, Xiaoyu",
    booktitle = "Proceedings of the 29th Pacific Asia Conference on Language, Information and
    month = oct,
    year = "2015",
    address = "Shanghai, China",
    url = "https://aclanthology.org/Y15-2030",
    pages = "258--266",
}
```

Dataset id: zh_multilan_amazon

- Domain: **reviews**
- Language: **zh**
- Language family: **Sino-Tibetan**
- Genus: **Chinese**
- Definite articles: **no article**
- Indefinite articles: **indefinite word same as one**
- Number of cases: **no morphological case-making**

- Order of subject, object, verb: **SVO**
- Negative morphemes: **negative particle**
- Polar questions: **question particle**
- Position of negative word wrt SOV: **SNegVO**
- Prefixing vs suffixing: **little affixation**
- Coding of nominal plurality: **no plural**
- Grammatical genders: **noun classifiers**

```
@inproceedings{dataset_multilan_amazon,
    title = "The Multilingual {A}mazon Reviews Corpus",
    author = {Keung, Phillip  and
        Lu, Yichao  and
        Szarvas, Gy{\"o}rgy  and
        Smith, Noah A.},
    booktitle = "Proceedings of the 2020 Conference on Empirical Methods in Natural Language P
    month = nov,
    year = "2020",
    address = "Online",
    publisher = "Association for Computational Linguistics",
    url = "https://aclanthology.org/2020.emnlp-main.369",
    doi = "10.18653/v1/2020.emnlp-main.369",
    pages = "4563--4568",
}
```