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

Łukasz Augustyniak  
WUST (Wrocław University of Science and Technology)  
lukasz.augustyniak@pwr.edu.pl

Szymon Woźniak  
Brand24 AI

Marcin Gruza  
Brand24 AI, WUST

Piotr Gramacki  
Brand24 AI, WUST

Krzysztof Rajda  
Brand24 AI, WUST

Mikołaj Morzy  
Poznań University of Technology

Tomasz Kajdanowicz  
WUST

Abstract

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.

1 Introduction

Consider a hotel booking service that allows its customers to post reviews. You have found just the perfect accommodation to stay for a couple of days with your family, but you browse through the reviews section of the website to check the experiences of former guests. Suddenly, you encounter a review in Polish: "hotel jak hotel, mogło być gorzej." This review has the following sentiment scores:

\[ s_{neg} = 0.44, s_{neu} = 0.44, s_{pos} = 0.12 \]

Intrigued by the ambiguity of scores, you translate the review into English: "hotel like a hotel, all in all, it could have been worse," which is scored as

\[ s_{neg} = 0.80, s_{neu} = 0.16, s_{pos} = 0.04 \]

After translating into English ("ok, ale nic zajímavého") the sentiment classification model scores the review as negative

\[ s_{neg} = 0.50, s_{neu} = 0.37, s_{pos} = 0.13 \]

After your stay, you decide to add a very positive review of the hotel ("it was a killer place to stay"; \[ s_{neg} = 0.03, s_{neu} = 0.05, s_{pos} = 0.92 \]). You would be very surprised to learn that the Czechs would be quite confused about your opinion ("to bylo vražedné)

\[ s_{neg} = 0.80, s_{neu} = 0.16, s_{pos} = 0.04 \]

Sentiment scores in this paragraph are produced by the multilingual cardiffnlp/twitter-xlm-roberta-base-sentiment model [9].
multilingual services become ubiquitous in the modern global economy. As more websites begin to offer automatic translation of content, users do not bother to express themselves in the lingua franca of the Web, writing instead in their native languages. Despite impressive advancements in automatic translation, many NLP tasks are still difficult in the multilingual setting. And sentiment classification is among the most challenging. The expression of sentiment is highly culture-dependent \[31\]. The emotional valence of individual words, the presence of specific phrasemes, and the expectations around sentiment values make sentiment classification across languages a demanding task.

Models performing sentiment classifications have to cope with two independent sources of variance in the input data: cultural expressions of sentiment and errors in automatic translations. In addition, the productization of sentiment classification leads to several engineering choices which influence the efficiency of the model:

- **single multilingual model vs. dedicated monolingual models**: deploying a single model in a production environment is much easier than orchestrating an ensemble of models,
- **training vs. fine-tuning**: sentiment classification model can be trained from scratch, both in the multi- and monolingual regimes, and the alternative is the fine-tuning of a general-purpose pre-trained language model with sentiment data,
- **transfer learning between domains**: a sentiment classification model trained for a specific domain (e.g., book reviews) can be transferred to another domain (e.g., hotel reviews) under the assumption that sentiment expressions in a given language remain independent of the subject of sentiment,
- **transfer learning between languages**: one may hypothesize that related languages can utilize similar sentiment expression mechanisms regarding grammar, punctuation, and vocabulary.

In theory, it is possible to use training data in language \(L_1\), fine-tune a multilingual model with this data, and as a result of fine-tuning, improve the performance of the model on language \(L_2\), provided languages \(L_1\) and \(L_2\) are sufficiently similar.

Working with multilingual datasets and models opens another opportunity. Most NLP research focuses on 20 major languages. Many languages native to significant human populations are not adequately studied. Although the term low-resource language is not precisely defined and can be understood in terms of computerization, privilege, the abundance of resources, or density \[47\], the existence of a large chasm between languages with respect to linguistic resources is apparent. Our work aims at supporting low-resource languages in performant sentiment classification. Finally, there is a lively debate in the scientific community about the inherent ability of neural models to handle general linguistic phenomena. This paper tries to answer this question in the domain of sentiment classification.

Contribution presented in this paper is threefold:

- **multilingual corpus**: we present the largest multilingual collection of sentiment classification datasets consisting of 79 high-quality datasets covering 27 languages. The collection of the corpus and detailed descriptions of collected datasets are presented in Section\[3\].
- **multi-faceted benchmark**: the dataset is supplemented with a benchmark containing detailed run statistics of hundreds of experiments representing different training and testing scenarios. All details relevant to the benchmark are presented in Section\[4\].
- **library for dataset access**: all datasets in the multilingual corpus are publicly available via a library compatible with the HuggingFace library, along with the ability to filter and select datasets, verify their licenses, etc\[5\].

2 Linguistic typology

The similarity of languages and the main aspects of their differences is the field of study of language typology. The differences between languages can be phonological (differences in sounds used

\[https://huggingface.co/datasets/Brand24/mms\]
by languages), syntactic (differences in language structures), lexical (differences in vocabulary), and theoretical (differences characterized as general properties of languages). Linguistic typology analyzes the current state of languages and is often contrasted with genealogical linguistics. The latter is concerned with historical relationships between languages established via historical records or with the help of comparative linguistics. The main focus of genealogical linguistics resolves around language families and language genera. The term language family refers to a group of languages sharing pedigree from a common ancestral language (the proto-language). As of today, linguists define over 7000 languages categorized into 150 families [15]. The largest families of languages include Indo-European, Sino-Tibetan, Turkic, Afro-Asiatic, Nilo-Saharan, Niger-Congo, and Eskimo-Aleut [24]. Main families are further divided into branches called genera. Examples of genera within the Indo-European family of languages include Slavic, Romance, Germanic, and Indic. This division of language families into genera closely follows the genetic family of humankind as attested by DNA similarity [43]. Some language families do not produce distinctive genera but form dialect continua defined by mutual intelligibility. Finally, an important concept is that of a sprachbund, a geographical area occupied by languages sharing linguistic features. The linguistic similarities within a sprachbund result from cultural exchange and geographical contact rather than by chance or common origin, as described by Thomason and Kaufman [85].

Languages can be described using hundreds of linguistic features. World Atlas of Language Structures [27] lists almost 200 different features. Since our work focuses on sentiment classification, we select 10 features that seem to be the most relevant to the task of sentiment expression. These features are:

1. **definite article**: the morpheme associated with nouns and used to code the uniqueness or definiteness of a concept; almost half of the languages do not use the definite article.

2. **indefinite article**: the morpheme used together with nouns to signal that the related concept is unknown to the hearer; half of the languages do not use the indefinite article, some languages use a separate article, and some use the numeral "one" as the indefinite article.

3. **number of cases**: morphological cases are a common way to express various relationships between words; human languages vary greatly in the number of cases used by a language.

4. **order of subject, verb, and object**: some languages have the strict ordering of words, while languages that convey semantics through inflection may be much looser with the ordering; half of all languages use the SOV (subject-object-verb) ordering, one third uses the SVO (subject-verb-object) ordering, and a small fraction of languages use the remaining VSO, VOS, OVS, and OSV orderings. Interestingly, around 13% of world languages do not have any fixed word order.

5. **negative morphemes**: negative morpheme is used to signal clausal negation in declarative sentences; this is usually achieved using a negative affix or a negative particle.

6. **polar questions**: a polar question is a question with only yes/no answers; these questions can be built using question particles, interrogative morphology, or intonation only.

7. **position of the negative morpheme**: languages differ by the position of the negative morpheme in relation to subjects and objects, with many variants such as SNegVO, NegSVO, SVNegO, obligatory and optional double negations, etc.

8. **prefixing vs. suffixing**: languages differ significantly in their use of prefixes versus suffixes in inflectional morphology.

9. **coding of nominal plurals**: two major types of plural coding are present in languages, either by changing the morphological form of the noun or by using a plurality indicator morpheme somewhere in the noun phrase.

10. **grammatical genders**: there is significant variability among languages with respect to the number of grammatical genders, some languages do not use the concept at all, and some languages may have 5 or more grammatical genders.

All language features mentioned above are available as filtering features in our library. 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 (Listing 1) or download all datasets from the Afro-Asiatic language family with no morphological case-making (Listing 2).
3 Datasets

3.1 Quality criteria

The initial pool of sentiment datasets has been gathered using extensive search and consisted of 345 datasets found through Google Scholar, GitHub repositories, and the HuggingFace datasets library. This initial pool of datasets has been manually filtered based on the following set of quality assurance criteria:

1. **strong annotations**: we have rejected datasets containing weak annotations (e.g., datasets with labels based on emoji occurrence or generated automatically through classification by machine learning models) due to an extensive amount of noise [58].

2. **well-defined annotation protocol**: we have rejected datasets without sufficient information about the annotation protocol (e.g., whether annotation was manual or automatic, number of annotators) to avoid merging datasets with contradicting annotation instructions.

3. **numerical ratings**: we have accepted datasets with numerical ratings, mapping Likert-type 5-point scales into three class sentiment labels as follows: ratings 1 and 2 were mapped to **negative**, rating 3 was mapped to **neutral**, and ratings 4 and 5 were mapped to **positive**.

4. **three classes only**: we have rejected datasets annotated with binary sentiment labels as their performance in three class settings was unsatisfactory.

5. **monolingual datasets**: when a dataset contained samples in multiple languages, we opted to divide it into independent datasets in constituent languages.

3.2 Pre-processing of datasets

Despite quality assurance criteria described in Section 3.1, the datasets still contained conflicting entries, i.e., duplicated records with different sentiment labels. We have cross-referenced all datasets to identify conflicts and have made the data coherent using majority-label voting. Finally, labeling and rating schemes of all datasets have been mapped to a 3-class scheme with **negative**, **neutral**, and **positive** labels only. For datasets with emotional annotations, we mapped positive emotions (joy, happiness) into positive sentiment and negative emotions (fear, sadness, disgust, anger) into negative sentiment. Texts with ambiguous emotions like anticipation and surprise were discarded. This pre-processing pipeline resulted in 79 datasets containing 6,164,762 text samples. Most of the datasets are in English (2,330,486 samples across 17 datasets), Arabic (932,075 samples across 9 datasets), and Spanish (418,712 samples across 5 datasets). The datasets represent four different domains: social media (44 datasets), reviews (24 datasets), news (5 datasets), and others (6 datasets). In addition, all datasets were processed by the cleanlab library to produce a self-confidence label-quality score for each data point.

Exhaustive testing of all configurations of the benchmark is unfeasible, but the total lack of baseline is unacceptable, either. We have decided to test the benchmark dataset by manually compiling a strong baseline. The main rationale behind this effort was the lack of coherence in annotation guidelines between considered datasets. Our baseline is built using strict annotation guidelines constructed iteratively over annotation batches, resulting in highly coherent annotations. The baseline dataset consists of 3,911 short text samples (trimmed to 350 characters) in Polish and English, annotated independently by 3 annotators fluent in these languages. Baseline texts cover multiple domains, such as social media, news sites, blogs, and Internet forums. For each instance, the majority label assigned by the annotators has been stored. The inter-rater agreement is $\kappa = 0.665$ (average Cohen’s kappa between three pairs of annotators) and $\alpha = 0.666$ (Krippendorff’s alpha).

One may think of the baseline dataset as representative of current sentiment model training, where researchers have to build domain-aligned corpora of various languages and annotate them. The comparison of results between the benchmark and the baseline datasets is a proxy of the performance trade-offs should one decide to use our benchmark dataset to train domain and language-specific sentiment classifiers.
Table 1: Summary of the corpus. Categories: N - news, R - reviews, SM - social media, O - other

<table>
<thead>
<tr>
<th>datasets</th>
<th>#samples</th>
<th>#words</th>
<th>#chars</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>17</td>
<td>304,939</td>
<td>304,939</td>
</tr>
<tr>
<td>Arabic</td>
<td>9</td>
<td>138,899</td>
<td>138,899</td>
</tr>
<tr>
<td>Spanish</td>
<td>5</td>
<td>108,733</td>
<td>108,733</td>
</tr>
<tr>
<td>Chinese</td>
<td>2</td>
<td>117,967</td>
<td>117,967</td>
</tr>
<tr>
<td>German</td>
<td>6</td>
<td>104,667</td>
<td>104,667</td>
</tr>
<tr>
<td>Polish</td>
<td>4</td>
<td>77,422</td>
<td>77,422</td>
</tr>
<tr>
<td>French</td>
<td>3</td>
<td>84,187</td>
<td>84,187</td>
</tr>
<tr>
<td>Japanese</td>
<td>1</td>
<td>83,982</td>
<td>83,982</td>
</tr>
<tr>
<td>Czech</td>
<td>4</td>
<td>39,674</td>
<td>39,674</td>
</tr>
<tr>
<td>Portuguese</td>
<td>4</td>
<td>56,827</td>
<td>56,827</td>
</tr>
<tr>
<td>Slovenian</td>
<td>2</td>
<td>33,694</td>
<td>33,694</td>
</tr>
<tr>
<td>Russian</td>
<td>2</td>
<td>31,770</td>
<td>31,770</td>
</tr>
<tr>
<td>Croatian</td>
<td>2</td>
<td>19,757</td>
<td>19,757</td>
</tr>
<tr>
<td>Serbian</td>
<td>3</td>
<td>25,089</td>
<td>25,089</td>
</tr>
<tr>
<td>Thai</td>
<td>2</td>
<td>9,326</td>
<td>9,326</td>
</tr>
<tr>
<td>Bulgarian</td>
<td>1</td>
<td>13,930</td>
<td>13,930</td>
</tr>
<tr>
<td>Hungarian</td>
<td>1</td>
<td>8,974</td>
<td>8,974</td>
</tr>
<tr>
<td>Slovak</td>
<td>1</td>
<td>14,341</td>
<td>14,341</td>
</tr>
<tr>
<td>Albanian</td>
<td>1</td>
<td>6,889</td>
<td>6,889</td>
</tr>
<tr>
<td>Swedish</td>
<td>1</td>
<td>16,266</td>
<td>16,266</td>
</tr>
<tr>
<td>Bosnian</td>
<td>1</td>
<td>11,974</td>
<td>11,974</td>
</tr>
<tr>
<td>Urdu</td>
<td>1</td>
<td>5,239</td>
<td>5,239</td>
</tr>
<tr>
<td>Hindi</td>
<td>1</td>
<td>4,992</td>
<td>4,992</td>
</tr>
<tr>
<td>Persian</td>
<td>1</td>
<td>1,602</td>
<td>1,602</td>
</tr>
<tr>
<td>Italian</td>
<td>2</td>
<td>4,043</td>
<td>4,043</td>
</tr>
<tr>
<td>Hebrew</td>
<td>1</td>
<td>2,279</td>
<td>2,279</td>
</tr>
<tr>
<td>Latvian</td>
<td>1</td>
<td>1,378</td>
<td>1,378</td>
</tr>
</tbody>
</table>

4 Multi-faceted benchmark

As we have explained in Section 1, the deployment of a multilingual sentiment classifier can be evaluated using several criteria leading to different architecture choices. Thus, we do not publish a single benchmark, but we aggregate the results along several dimensions. The benchmark available at [https://huggingface.co/spaces/Brand24/mms_benchmark](https://huggingface.co/spaces/Brand24/mms_benchmark) allows to compare models according to:

- **number of languages**: multilingual models with monolingual models,
- **training procedure**: models trained from scratch vs. fine-tuning,
- **domain language**: language to which a model is applied,
- **data modality**: news, reviews, social media,
- **knowledge transfer**: transfer between languages, transfer between domains.

Table 2 presents the list of models included in the benchmark. For each model, we include the number of parameters and languages used in pre-training and the base model. The results presented in the benchmark reflect three possible scenarios of model deployment. A pre-trained model can be used to generate text representation only. In this scenario, denoted head-linear (HL), a model serves as a feature extractor followed by a small linear classification head. In the second scenario, the linear classification head is replaced by a BiLSTM classifier operating on features extracted by the pre-trained model. We refer to this scenario as head-bilstm (HB). Finally, each pre-trained transformer-based model (with the exception of mUSE-transformer) has been fine-tuned to the sentiment classification task. We refer to this scenario as fine-tuning (FT).
Table 2: Models included in the benchmark

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

Figure 1: Detailed comparison of models. Legend: **lang** - averaged by all languages, **ds** - averaged by dataset, **ar** - Arabic, **bg** - Bulgarian, **bs** - Bosnian, **cs** - Czech, **de** - German, **en** - English, **es** - Spanish, **fa** - Persian, **fr** - French, **he** - Hebrew, **hi** - Hindi, **hr** - Croatian, **hu** - Hungarian, **it** - Italian, **ja** - Japanese, **lv** - Latvian, **pl** - Polish, **pt** - Portuguese, **ru** - Russian, **sk** - Slovak, **sl** - Slovenian, **sq** - Albanian, **sr** - Serbian, **sv** - Swedish, **th** - Thai, **ur** - Urdu, **zh** - Chinese.

The hyperparameters of models included in the benchmark are as follows. The hidden size of the model was set to the embedding size of each model when used in HL and FT scenarios. The hidden size of the HB scenario was set to $h = 32$. The learning rate for the HL scenario was $\eta = 1 \times 10^{-3}$, fine-tuning used $\eta = 1 \times 10^{-5}$, and HB scenario used $\eta = 5 \times 10^{-3}$. The batch size in HL and HB scenarios was $b = 200$ and $b = 6$ for fine-tuning. Dropout for HB was $d = 0.5$ and $d = 0.2$ for other scenarios. Training took 5 epochs for fine-tuning and 2 epochs for HL and HB scenarios (beyond 2 epochs most models started to overfit). The models are evaluated using the traditional $F_1$ score computed on three levels: the entire dataset, averaged over all datasets, and the internal dataset.

We performed our experiments using Python 3.9 and PyTorch (1.8.1) (and Tensorflow (2.3.0) for original mUSE). Our experimental setup consists of Intel(R) Xeon(R) CPU E5-2630 v4 @ 2.20GHz and Nvidia Tesla V100 16GB.

Here we present one result available through the benchmarks. Due to space constraints, we move the presentation of other case studies to Appendix B. Figure 1 contains $F_1$ scores of all models fine-tuned on all datasets available for a given language. Interestingly, we see little variance in performance between models for high-resource languages and significant deterioration of performance for low-resource languages. An outlier is the aforementioned Portuguese, where the performance is caused by the lack of data points representing news and reviews. Figure 1 directly shows the usefulness of our benchmark. When designing sentiment classification solutions in Spanish, the choice of model is secondary (models’ performances are very similar), and other model features can be considered (such as ease of deployment, cost of inference, and memory requirements). However, if sentiment classification is to be applied to Latvian, the performance difference between models can be as high as 32 pp depending on the choice of the model.
5 Related work

5.1 Multilingual text representations

One of the foundational discoveries in multilingual presentation learning was the fact that latent vector spaces seemed to encode very similar word relationships across a wide spectrum of languages. Needless to say, first monolingual static word embeddings were quickly followed by multilingual word embeddings which provided the basis for multilingual text representations [68]. A popular solution was to use pre-trained monolingual embeddings, such as word2vec [49] and align them via linear transformations using parallel multilingual dictionaries [50]. Another approach advocated for joint learning of multilingual word embeddings on pseudo-bilingual datasets, where tokens in one language would be randomly translated to another language [92].

Static multilingual embeddings were superseded by contextual embeddings produced by language models such as BiLSTM [5] and various Transformers [30, 22, 26, 93, 95]. In the case of these more complex architectures, the multilingual capabilities of language models resulted from specific objective functions used during training. These objectives “pushed” the models toward universal multilingual representations by forcing the models to perform machine translation [5], translation language modeling (TLM) [22, 20], or translation ranking [30, 24].

The evaluation of the quality of multilingual text representation is not trivial. Cross-lingual and multilingual tasks are actively developed to foster the development of multilingual models. Examples of such cross-lingual and multilingual tasks include cross-lingual natural language inference [21], question answering [42], named entity recognition [87, 88] or parallel text extraction [97, 96]. For example, results of comparison of LASER, mBERT, and XLM in tasks of named entity recognition and part-of-speech tagging in zero-shot settings suggests that LASER outperforms the latter methods in the case of knowledge transfer [89]. Another important benchmark is XTREME [34], designed for testing the abilities of cross-lingual transfer across 40 languages and 9 tasks. Despite its massive character, XTREME lacks benchmarking task of sentiment analysis. Also, only mBERT, XLM, XLM-R, and MMTE are used as baseline models.

By far, the most popular pre-trained multilingual language model is Multilingual BERT (mBERT) [26]. It has been used in several cross-lingual studies, for instance, in zero-shot knowledge transfer between Slovenian and Croatian languages [62], exploring code-switching on Spanglish and Hinglish [61], and building hierarchical architecture for zero-shot setting [74]. The properties and limitations of mBERT have been extensively studied. One of the studies showed that mBERT is not learning a joint representation of languages. Rather, it partitions the representation space between languages [79]. The authors used the Projection Weighted Canonical Correlation Analysis (PWCCA) to analyze how translations of the same sentence are represented in mBERT layers. The correlations were stable in pre-trained and fine-tuned models, with the effect being more pronounced in deeper layers of the model. The hierarchical structure induced by correlations was similar to the structure produced by genealogical linguistics.

Another interesting finding was that mBERT encodes language-specific information within the parameter space, and this language component is not removed by fine-tuning [44]. The language-specific component of mBERT can be removed by estimating the centroid of the language (the mean of mBERT embeddings of a given language vocabulary) and subtracting this centroid from representations produced in the language. The existence of the language component has been proven in multiple tasks: language classification, language similarity, sentence retrieval, word alignment, and machine translation.

Several works tested the cross-lingual abilities of mBERT as compared to monolingual models, finding its performance on low-resource languages inferior to monolingual models [91]. One experiment used a bilingual version of mBERT and trained it in multiple configurations to test the influence of the linguistic relationship between the source and the target language, the network architecture, and the input and learning objective [37]. The authors found that structural similarity and depth of a model are the most significant factors behind mBERT’s cross-lingual performance. On the other hand, multi-head attention was not particularly important. Regarding the low-resource languages, [91] compared mBERT with baseline models on tasks of named entity recognition, universal part-of-speech tagging, and universal dependency parsing. The authors focused on the comparison between low- and high-resource languages (determined by the size of the Wikipedia dump in each language). mBERT was found to work well on high-resource languages but under-performed on low-resource
languages. Finally, Liu et al. [46] analyzed the relationship between the contextual aspect of mBERT and the training dataset size in the context of multilingual datasets. The authors compared contextual mBERT embeddings with non-contextual models of Word2Vec and GloVe. As it turns out, mBERT outperforms non-contextual models only when large datasets are available for training or fine-tuning. In the low data regime, as well as when the context window is short (limited input), multilingual mBERT is surpassed by static token embeddings.

5.2 Multilingual sentiment classification

Several surveys present an overview of traditional sentiment classification methods (lexicon-based approaches and shallow models with lexical features engineering) [25, 70]. Recently, deep-learning approaches became more prevalent. Attia et al. [6] use convolutional neural networks on word-level embeddings of texts in English, German and Arabic. The approach is computationally expensive as it requires separate embedding dictionaries for each language. An alternative approach is to use character-level embeddings. Wehrmann et al. [90] trained such a model for binary sentiment classification of English, German, Portuguese, and Spanish tweets. A similar model was used in [4], but the authors applied multi-stage pre-processing, including dictionary checks, Soundex algorithm, and lemmatization. Other examples of deep neural models for multilingual sentiment classification include recurrent neural networks supplied with machine translation. Can et al. [16] trained a model on English reviews and evaluated it on machine-translated reviews in Russian, Spanish, Turkish, and Dutch using the Google Translation API and pre-trained GloVe embeddings for English. A similar approach is presented in [38], where the authors used LASER sentence embeddings to train a sentiment classifier on Polish reviews and used this classifier to score reviews translated into other languages.

5.3 Multilingual sentiment datasets

Multilingual sentiment datasets are crucial resources for training sentiment analysis models that can operate across various languages. Unfortunately, very few multilingual datasets contain sentiment polarity annotations. The Multilingual Amazon Reviews Corpus [57] is a noteworthy example of such a dataset. It provides customer review data in English, Japanese, German, French, Spanish, and Chinese, annotated for sentiment. Another significant resource is the Multilingual Sentiment Lexicons developed by Chen and Skiena [18]), which provides sentiment lexicons for 136 languages. Being a lexicon, it cannot be used to train sentiment models directly, but it can be used for weak supervision. The NTCIR corpus contains information on sentiment polarity for news related to sport and politics for articles in English, Chinese, and Japanese [76]. The JMTC dataset [52] (Jigsaw Multilingual Toxic Comment Classification) dataset contains comments from Wikipedia’s talk page edits and is available in multiple languages. It is annotated for toxic vs. non-toxic sentiment, which cannot be aligned with our 3-class annotation schema. An interesting resource is Task 4 from SemEval providing English, Spanish, Dutch, Arabic, Russian, and Turkish tweets and SMS messages annotated for 3-class sentiment [54, 67]. The most similar dataset to ours is XED compiled by Öhman et al. [59]. However, this dataset contains original annotations only for English and Finnish, and the remaining 30 languages are "projected" (i.e., annotated samples in English and Finnish are translated into multiple languages).

6 Discussion

The main focus of the paper is the introduction of a massively multilingual large collection of sentiment datasets and the extensive benchmark of model training and validation scenarios. During the construction of the benchmark, we have gathered valuable experiences that we want to share with the community. Our most important observation is that a single multilingual sentiment classifier can perform approximately equally well for all languages. A small group of pre-trained models (XML-R, LaBSE, MPNet) under fine-tuning scenario produces the best results relative to other models. Obviously, the absolute $F_1$ score across languages differs significantly (due to data scarcity, data quality, or difficulty of samples), but these selected models always produce top classifiers. Secondly, all models perform better under the fine-tuning scenario, but the performance gain varies from model to model. When evaluated on the test dataset, models gained between 4 pp. (mUSE-dist) and 9 pp. (mBERT, DistilmBERT). When tested on the internal datasets, $F_1$ gains varied between 0 pp (mUSE-
dist) and 20 pp. (DistilmBERT), with mT5 and XML-R improving by 17 pp. and 15 pp., respectively. In general, the largest gains can be observed for models trained with the masked language modeling technique (XML-R, mBERT). We also observe that models which produce lower-quality sentence embeddings gain more from fine-tuning.

We also find that bigger models tend to achieve better performance in all data modalities and training scenarios. Of course, there are counterexamples, e.g., mUSE-dist is smaller than mBERT but achieves better performance in the HL scenario for all datasets. This indicates that the size of the model is an important factor for determining its performance, but other factors, like the domain and the type of pretraining task, may also affect the results. Moreover, we observe that the correlation between model size and model performance is weaker after fine-tuning. This means that one may often find a competitive model with similar performance to the state-of-the-art model but significantly smaller and faster for the production environment.

Our final remark is a warning: in our experiments, we have encountered a significant variability in performance conditioned on a particular data split. In one of the cross-validation fine-tuning experiments, we observed a data fold that produced consistently worse results (up to 4 pp.) for all models. This fold was produced by the same random seed and (most probably) consisted of samples with increased difficulty. Since our experiments involved multiple models, multiple training scenarios, multiple datasets, and multiple languages, we had to rely on data subsampling. Conducting experiments on a fixed seed for sample selection could lead to a similar data fold and, consequently, to biased experimental results.

7 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.

An important limitation of our dataset collection is a significant variance in sample quality across all datasets and all languages. Figure 2 presents the distribution of self-confidence label-quality score for each data point computed by the \texttt{cleanlab [58]}. 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.
Acknowledgments and Disclosure of Funding

The work presents results achieved in projects funded by (1) European Regional Development Fund (ERDF) in RPO WD 2014-2020 (project no. RPDS.01.02.01-02-0065/20); (2) the Polish Ministry of Education and Science, CLARIN-PL; (3) the ERDF in the 2014-2020 Smart Growth Operational Programme, CLARIN – Common Language Resources and Technology Infrastructure; (4) project CLARIN-Q (agreement no. 2022/WK/09); (5) the Department of Artificial Intelligence at Wroclaw University of Science and Technology. Mikofaj Morzy is supported by (6) EEA Financial Mechanism 2014-2021 Project 2019/35/J/HS6/03498.

References


A Datasets

We present detailed lists of datasets included in our research in Tables[3] and [4]. They include language, category, dataset size, class balance, and basic dataset characteristics.

Table 3: List of all monolingual datasets used in experiments. Category (Cat.): R - Reviews, SM - Social Media, C - Chats, N - News, P - Poems, M - Mixed. HL - human labeled, #Words and #Chars are mean values

<table>
<thead>
<tr>
<th>Paper</th>
<th>Lang</th>
<th>Cat.</th>
<th>HL</th>
<th>Samples</th>
<th>NEG/NEU/POS</th>
<th>#Words</th>
<th>#Chars</th>
</tr>
</thead>
<tbody>
<tr>
<td>Al Omari et al. [11]</td>
<td>ar</td>
<td>R</td>
<td>No</td>
<td>3096</td>
<td>13.0/10.2/76.8</td>
<td>9</td>
<td>51</td>
</tr>
<tr>
<td>Elnagar et al. [29]</td>
<td>ar</td>
<td>R</td>
<td>No</td>
<td>400101</td>
<td>13.0/19.9/67.1</td>
<td>22</td>
<td>127</td>
</tr>
<tr>
<td>Aly and Atiya [12]</td>
<td>ar</td>
<td>R</td>
<td>No</td>
<td>6250</td>
<td>11.6/17.9/70.5</td>
<td>65</td>
<td>343</td>
</tr>
<tr>
<td>Elnagar and Einea [28]</td>
<td>ar</td>
<td>No</td>
<td>504007</td>
<td>15.8/21.0/63.6</td>
<td>77</td>
<td>424</td>
<td></td>
</tr>
<tr>
<td>Baly et al. [7]</td>
<td>ar</td>
<td>SM</td>
<td>Yes</td>
<td>2809</td>
<td>47.2/23.9/29.0</td>
<td>22</td>
<td>130</td>
</tr>
<tr>
<td>Nabil et al. [53]</td>
<td>ar</td>
<td>SM</td>
<td>Yes</td>
<td>3224</td>
<td>59.9/25.0/24.1</td>
<td>16</td>
<td>94</td>
</tr>
<tr>
<td>Salameh et al. [71]</td>
<td>ar</td>
<td>SM</td>
<td>Yes</td>
<td>1199</td>
<td>48.0/10.5/41.5</td>
<td>11</td>
<td>51</td>
</tr>
<tr>
<td>Salameh et al. [71]</td>
<td>ar</td>
<td>SM</td>
<td>Yes</td>
<td>1998</td>
<td>67.5/10.1/22.4</td>
<td>20</td>
<td>107</td>
</tr>
<tr>
<td>Habernal et al. [52]</td>
<td>cs</td>
<td>R</td>
<td>No</td>
<td>91140</td>
<td>32.4/33.7/33.9</td>
<td>50</td>
<td>311</td>
</tr>
<tr>
<td>Habernal et al. [52]</td>
<td>cs</td>
<td>R</td>
<td>No</td>
<td>92758</td>
<td>7.9/23.4/68.7</td>
<td>20</td>
<td>131</td>
</tr>
<tr>
<td>Habernal et al. [52]</td>
<td>cs</td>
<td>SM</td>
<td>Yes</td>
<td>9752</td>
<td>20.4/53.1/26.5</td>
<td>10</td>
<td>59</td>
</tr>
<tr>
<td>Habernal et al. [52]</td>
<td>cs</td>
<td>SM</td>
<td>Yes</td>
<td>2637</td>
<td>30.8/60.6/8.6</td>
<td>33</td>
<td>170</td>
</tr>
<tr>
<td>Cieliebak et al. [19]</td>
<td>de</td>
<td>SM</td>
<td>Yes</td>
<td>9948</td>
<td>16.3/59.2/24.6</td>
<td>11</td>
<td>86</td>
</tr>
<tr>
<td>Schabus and Skowron [75]</td>
<td>de</td>
<td>SM</td>
<td>Yes</td>
<td>3598</td>
<td>47.3/51.5/1.2</td>
<td>33</td>
<td>237</td>
</tr>
<tr>
<td>Chapuis et al. [17]</td>
<td>en</td>
<td>C</td>
<td>Yes</td>
<td>12138</td>
<td>31.8/46.5/21.7</td>
<td>12</td>
<td>48</td>
</tr>
<tr>
<td>Chapuis et al. [17]</td>
<td>en</td>
<td>C</td>
<td>Yes</td>
<td>4643</td>
<td>22.3/48.9/28.8</td>
<td>15</td>
<td>71</td>
</tr>
<tr>
<td>Malo et al. [48]</td>
<td>en</td>
<td>N</td>
<td>Yes</td>
<td>3448</td>
<td>12.2/62.1/25.7</td>
<td>22</td>
<td>124</td>
</tr>
<tr>
<td>Bastan et al. [10]</td>
<td>en</td>
<td>N</td>
<td>Yes</td>
<td>5333</td>
<td>11.6/37.3/51.8</td>
<td>388</td>
<td>2129</td>
</tr>
<tr>
<td>Hutto and Gilbert [55]</td>
<td>en</td>
<td>N</td>
<td>No</td>
<td>5190</td>
<td>29.3/52.9/17.8</td>
<td>17</td>
<td>104</td>
</tr>
<tr>
<td>Sheng and Uthaisuk [78]</td>
<td>en</td>
<td>P</td>
<td>Yes</td>
<td>1052</td>
<td>18.3/15.8/65.9</td>
<td>7</td>
<td>37</td>
</tr>
<tr>
<td>Hutto and Gilbert [55]</td>
<td>en</td>
<td>R</td>
<td>No</td>
<td>3708</td>
<td>34.2/19.5/46.3</td>
<td>16</td>
<td>87</td>
</tr>
<tr>
<td>Hutto and Gilbert [55]</td>
<td>en</td>
<td>R</td>
<td>No</td>
<td>10605</td>
<td>49.6/1.5/48.9</td>
<td>19</td>
<td>111</td>
</tr>
<tr>
<td>Ni et al. [56]</td>
<td>en</td>
<td>R</td>
<td>No</td>
<td>1883238</td>
<td>8.3/80.8/37.3</td>
<td>70</td>
<td>382</td>
</tr>
<tr>
<td>Sanders [72]</td>
<td>en</td>
<td>SM</td>
<td>Yes</td>
<td>3424</td>
<td>16.7/68.1/15.2</td>
<td>14</td>
<td>97</td>
</tr>
<tr>
<td>Thelwall et al. [84]</td>
<td>en</td>
<td>SM</td>
<td>Yes</td>
<td>11759</td>
<td>28.0/30.4/38.0</td>
<td>26</td>
<td>147</td>
</tr>
<tr>
<td>Inc. [36]</td>
<td>en</td>
<td>SM</td>
<td>Yes</td>
<td>14427</td>
<td>63.2/21.2/15.8</td>
<td>17</td>
<td>104</td>
</tr>
<tr>
<td>Hutto and Gilbert [55]</td>
<td>en</td>
<td>SM</td>
<td>No</td>
<td>4200</td>
<td>26.9/17.0/56.1</td>
<td>13</td>
<td>79</td>
</tr>
<tr>
<td>Keith Norambuena et al. [59]</td>
<td>es</td>
<td>R</td>
<td>Yes</td>
<td>382</td>
<td>45.0/27.2/27.8</td>
<td>165</td>
<td>1033</td>
</tr>
<tr>
<td>Cruz et al. [23]</td>
<td>es</td>
<td>R</td>
<td>No</td>
<td>3871</td>
<td>32.8/32.3/34.9</td>
<td>511</td>
<td>3000</td>
</tr>
<tr>
<td>Hosseini et al. [53]</td>
<td>fa</td>
<td>R</td>
<td>Yes</td>
<td>13525</td>
<td>12.0/37.5/50.5</td>
<td>21</td>
<td>104</td>
</tr>
<tr>
<td>Amram et al. [6]</td>
<td>he</td>
<td>SM</td>
<td>Yes</td>
<td>8619</td>
<td>26.5/2.8/71.8</td>
<td>22</td>
<td>110</td>
</tr>
<tr>
<td>Pelicon et al. [53]</td>
<td>hr</td>
<td>N</td>
<td>Yes</td>
<td>2025</td>
<td>22.5/61/46.0</td>
<td>161</td>
<td>1021</td>
</tr>
<tr>
<td>Barbieri et al. [68]</td>
<td>it</td>
<td>SM</td>
<td>Yes</td>
<td>8926</td>
<td>36.7/41.7/21.6</td>
<td>14</td>
<td>101</td>
</tr>
<tr>
<td>Sprugnoli [82]</td>
<td>it</td>
<td>SM</td>
<td>Yes</td>
<td>3139</td>
<td>24.4/14.9/60.6</td>
<td>17</td>
<td>106</td>
</tr>
<tr>
<td>Sprogis and Rikters [81]</td>
<td>lv</td>
<td>SM</td>
<td>Yes</td>
<td>5790</td>
<td>23.8/45.2/31.0</td>
<td>20</td>
<td>138</td>
</tr>
<tr>
<td>Rybak et al. [69]</td>
<td>pl</td>
<td>R</td>
<td>No</td>
<td>10074</td>
<td>30.8/13.2/56.0</td>
<td>80</td>
<td>494</td>
</tr>
<tr>
<td>Kocoen et al. [41]</td>
<td>pl</td>
<td>R</td>
<td>Yes</td>
<td>57098</td>
<td>42.4/26.8/30.8</td>
<td>30</td>
<td>175</td>
</tr>
<tr>
<td>Sobkowicz and Sobkowicz [80]</td>
<td>pl</td>
<td>SM</td>
<td>Yes</td>
<td>645</td>
<td>50.7/74.3/2.0</td>
<td>33</td>
<td>230</td>
</tr>
<tr>
<td>Rogers et al. [65]</td>
<td>ru</td>
<td>SM</td>
<td>Yes</td>
<td>23226</td>
<td>16.8/54.6/28.6</td>
<td>12</td>
<td>79</td>
</tr>
<tr>
<td>Bučar et al. [14]</td>
<td>sl</td>
<td>N</td>
<td>Yes</td>
<td>10417</td>
<td>32.0/52.0/16.0</td>
<td>309</td>
<td>2017</td>
</tr>
<tr>
<td>Batanović et al. [11]</td>
<td>sr</td>
<td>R</td>
<td>No</td>
<td>4724</td>
<td>17.8/43.7/38.5</td>
<td>498</td>
<td>3097</td>
</tr>
<tr>
<td>Batanović et al. [12]</td>
<td>sr</td>
<td>R</td>
<td>Yes</td>
<td>3948</td>
<td>30.3/18.1/51.5</td>
<td>18</td>
<td>105</td>
</tr>
<tr>
<td>Thongthanomkul et al. [86]</td>
<td>th</td>
<td>No</td>
<td>46193</td>
<td>5.3/30.5/61.1</td>
<td>25</td>
<td>544</td>
<td></td>
</tr>
<tr>
<td>Suriyawongkul et al. [83]</td>
<td>th</td>
<td>SM</td>
<td>Yes</td>
<td>26126</td>
<td>27.1/55.6/18.3</td>
<td>6</td>
<td>90</td>
</tr>
<tr>
<td>Sharf and Rahman [72]</td>
<td>ur</td>
<td>M</td>
<td>Yes</td>
<td>19660</td>
<td>26.7/43.6/29.7</td>
<td>13</td>
<td>69</td>
</tr>
<tr>
<td>Lin et al. [45]</td>
<td>zh</td>
<td>R</td>
<td>No</td>
<td>125725</td>
<td>28.6/21.9/49.5</td>
<td>51</td>
<td>128</td>
</tr>
</tbody>
</table>
Table 4: List of all multilingual datasets used in experiments. Category (Cat.): R - Reviews, SM - Social Media, C - Chats, N - News, P - Poems, M - Mixed. HL - human labeled.

<table>
<thead>
<tr>
<th>Paper</th>
<th>Cat.</th>
<th>Lang</th>
<th>HL</th>
<th>Samples</th>
<th>(NEG/NEU/POS)</th>
<th>#Words</th>
<th>#Char.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Narr et al. [55]</td>
<td>SM</td>
<td>de</td>
<td>Yes</td>
<td>953</td>
<td>10.0/75.1/14.9</td>
<td>12</td>
<td>80</td>
</tr>
<tr>
<td></td>
<td></td>
<td>en</td>
<td>Yes</td>
<td>1781</td>
<td>16.9/63.3/19.8</td>
<td>13</td>
<td>81</td>
</tr>
<tr>
<td></td>
<td></td>
<td>fr</td>
<td>Yes</td>
<td>7073</td>
<td>17.4/60.0/22.6</td>
<td>14</td>
<td>78</td>
</tr>
<tr>
<td>Keung et al. [40]</td>
<td>R</td>
<td>de</td>
<td>No</td>
<td>209073</td>
<td>40.1/20.0/39.9</td>
<td>33</td>
<td>208</td>
</tr>
<tr>
<td></td>
<td></td>
<td>en</td>
<td>No</td>
<td>209393</td>
<td>40.0/20.0/40.0</td>
<td>34</td>
<td>179</td>
</tr>
<tr>
<td></td>
<td></td>
<td>es</td>
<td>No</td>
<td>208127</td>
<td>40.2/20.0/39.8</td>
<td>27</td>
<td>152</td>
</tr>
<tr>
<td></td>
<td></td>
<td>fr</td>
<td>No</td>
<td>208160</td>
<td>40.2/20.1/39.7</td>
<td>28</td>
<td>160</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ja</td>
<td>No</td>
<td>209780</td>
<td>40.0/20.0/40.0</td>
<td>2</td>
<td>101</td>
</tr>
<tr>
<td></td>
<td></td>
<td>zh</td>
<td>No</td>
<td>205977</td>
<td>39.8/20.1/40.1</td>
<td>1</td>
<td>50</td>
</tr>
<tr>
<td>Rosenthal et al. [66]</td>
<td>M</td>
<td>ar</td>
<td>Yes</td>
<td>9391</td>
<td>35.5/40.6/23.9</td>
<td>14</td>
<td>105</td>
</tr>
<tr>
<td>Patwa et al. [60]</td>
<td>SM</td>
<td>es</td>
<td>Yes</td>
<td>14920</td>
<td>16.8/33.1/50.0</td>
<td>16</td>
<td>86</td>
</tr>
<tr>
<td>Mozetič et al. [51]</td>
<td>SM</td>
<td>bg</td>
<td>Yes</td>
<td>62150</td>
<td>22.6/45.9/31.5</td>
<td>12</td>
<td>85</td>
</tr>
<tr>
<td></td>
<td></td>
<td>bs</td>
<td>Yes</td>
<td>36183</td>
<td>33.4/30.5/36.1</td>
<td>12</td>
<td>75</td>
</tr>
<tr>
<td></td>
<td></td>
<td>de</td>
<td>Yes</td>
<td>90534</td>
<td>19.7/52.8/27.4</td>
<td>12</td>
<td>94</td>
</tr>
<tr>
<td></td>
<td></td>
<td>en</td>
<td>Yes</td>
<td>85784</td>
<td>26.8/44.1/29.1</td>
<td>12</td>
<td>77</td>
</tr>
<tr>
<td></td>
<td></td>
<td>es</td>
<td>Yes</td>
<td>191412</td>
<td>11.8/37.9/50.3</td>
<td>14</td>
<td>92</td>
</tr>
<tr>
<td></td>
<td></td>
<td>hr</td>
<td>Yes</td>
<td>75569</td>
<td>25.7/23.9/50.4</td>
<td>12</td>
<td>91</td>
</tr>
<tr>
<td></td>
<td></td>
<td>hu</td>
<td>Yes</td>
<td>56682</td>
<td>15.9/31.0/53.1</td>
<td>11</td>
<td>83</td>
</tr>
<tr>
<td></td>
<td></td>
<td>pl</td>
<td>Yes</td>
<td>168931</td>
<td>30.0/26.1/43.9</td>
<td>11</td>
<td>82</td>
</tr>
<tr>
<td></td>
<td></td>
<td>pt</td>
<td>Yes</td>
<td>145197</td>
<td>37.2/35.0/27.8</td>
<td>10</td>
<td>61</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ru</td>
<td>Yes</td>
<td>87704</td>
<td>32.0/40.1/27.8</td>
<td>10</td>
<td>67</td>
</tr>
<tr>
<td></td>
<td></td>
<td>sk</td>
<td>Yes</td>
<td>56623</td>
<td>25.6/22.5/51.9</td>
<td>13</td>
<td>97</td>
</tr>
<tr>
<td></td>
<td></td>
<td>sl</td>
<td>Yes</td>
<td>103126</td>
<td>29.9/43.3/26.8</td>
<td>13</td>
<td>91</td>
</tr>
<tr>
<td></td>
<td></td>
<td>sq</td>
<td>Yes</td>
<td>44284</td>
<td>15.7/33.1/51.1</td>
<td>13</td>
<td>90</td>
</tr>
<tr>
<td></td>
<td></td>
<td>sr</td>
<td>Yes</td>
<td>67696</td>
<td>34.8/42.8/22.4</td>
<td>13</td>
<td>81</td>
</tr>
<tr>
<td></td>
<td></td>
<td>sv</td>
<td>Yes</td>
<td>41346</td>
<td>40.3/31.2/28.5</td>
<td>14</td>
<td>94</td>
</tr>
</tbody>
</table>

B Experiments

B.1 Data quality

In this section, we expand upon the results of experiments included in the benchmark. We begin with the presentation of the distribution of data point quality across all datasets. The data point quality has been computed using the cleanlab library self-confidence label-quality score [58]. Figure 2 presents the cumulative distribution of data quality score. As we can see, the data quality still is a pressing issue, despite our efforts to publish a well-curated collection. Over 40% of all data points have quality 0.6 or worse. As we have mentioned, low-quality data points are not distributed uniformly across all languages and datasets, but the impact of low-quality data is most pronounced for low-resource languages.

B.2 Pairwise ranked comparison of models

Of course, everyone is most interested in finding the answer to the following question: "Which model should I use?" The answer, as usual, is: "It depends". Figure 3 presents the results of the Nemenyi pairwise rank comparison test. In neither of the considered training scenarios (HL, HB, TF), a single model is statistically better than others. However, these results allow us to draw partial conclusions. For instance, the MPNet performs best in the HL scenario, and the same model is not significantly worse than XML-M, which is the best model in HB and TF scenarios. We can also notice that mBERT-based models (mBERT and DistillmBERT) proved to be the worst language models for our
tasks. Three models stand out as the most promising choices: XLM-R, LaBSE, and MPNet. They achieve comparable performance in all scenarios and test cases. Furthermore, they are better than other models in almost all test cases.
Table 5: Aggregated $F_1$ results of models. The best results for each test set are highlighted. W - whole test, A - avg. by dataset, I - internal

<table>
<thead>
<tr>
<th></th>
<th>XLM-R</th>
<th>LaBSE</th>
<th>MPNet</th>
<th>XLM-R-dist</th>
<th>mT5</th>
<th>mBERT</th>
<th>DistlmBERT</th>
<th>mUSE-dist</th>
<th>LASER</th>
<th>mUSE-trans</th>
<th>mUSE-cnn</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>HL scenario</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>W</td>
<td>62</td>
<td>62</td>
<td>63</td>
<td>60</td>
<td>59</td>
<td>56</td>
<td>55</td>
<td>59</td>
<td>55</td>
<td>55</td>
<td>54</td>
</tr>
<tr>
<td>A</td>
<td>51</td>
<td>54</td>
<td>55</td>
<td>51</td>
<td>49</td>
<td>45</td>
<td>43</td>
<td>50</td>
<td>47</td>
<td>47</td>
<td>45</td>
</tr>
<tr>
<td>I</td>
<td>55</td>
<td>61</td>
<td>61</td>
<td>56</td>
<td>50</td>
<td>43</td>
<td>38</td>
<td>60</td>
<td>50</td>
<td>49</td>
<td>50</td>
</tr>
<tr>
<td><strong>HB scenario</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>W</td>
<td>66</td>
<td>62</td>
<td>63</td>
<td>62</td>
<td>65</td>
<td>60</td>
<td>59</td>
<td>62</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>A</td>
<td>57</td>
<td>55</td>
<td>56</td>
<td>54</td>
<td>56</td>
<td>49</td>
<td>48</td>
<td>54</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>I</td>
<td>64</td>
<td>63</td>
<td>64</td>
<td>63</td>
<td>54</td>
<td>48</td>
<td>64</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td><strong>FT scenario</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>W</td>
<td>68</td>
<td>68</td>
<td>67</td>
<td>67</td>
<td>66</td>
<td>65</td>
<td>64</td>
<td>63</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>A</td>
<td>61</td>
<td>62</td>
<td>62</td>
<td>62</td>
<td>60</td>
<td>56</td>
<td>56</td>
<td>56</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>I</td>
<td>70</td>
<td>69</td>
<td>65</td>
<td>67</td>
<td>57</td>
<td>58</td>
<td>60</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

B.3 Detailed results for each model

Table 5 shows the $F_1$ scores of all models aggregated by datasets. The models improve with fine-tuning (up to 0.7 $F_1$) as compared to linear (0.61) or BiLSTM (0.64) layers operating on embeddings produced by models. The performance gains are more pronounced for models trained with masked language modeling objectives (mBERT, XLM-R) than for models trained with sentence classification or sentence similarity objectives (LaBSE). Fine-tuning reduces inequalities in the results between models (0.55 vs. 0.43 for best and worst models in the HL scenario). The additional BiLSTM layer on top of transformer-based token embeddings has the capacity to improve the results compared to the linear classification layer model. The differences are most clear in the case of the results for our internal dataset, where the result improved even by 13 pp. for the mT5 model.
B.4 Detailed results for each language

We assessed the performance of each model in each experimental scenario per language. The texts were sub-sampled with stratification by language and class label so that language distribution in the test dataset follows the distribution in the whole dataset. We also include the total macro $F_1$ score value in column “all”. Results are presented in Figure 4. Those results confirm our previous findings regarding the advantage of XLM-R, LaBSE, and MPNet models. They outperform other models in most languages, with no clear indication of superiority among them.

![Figure 4: Detailed results of models’ comparison.](image)

B.5 Transfer between data modalities

The last example compares the effectiveness of transfer learning between data modalities. The results are presented in Figure 5. As expected, when models are tested against datasets from the same domain (news, social media, reviews), the average performance is much higher than on out-of-domain datasets. What is striking is the visible deterioration of the performance of models trained in the news domain. When knowledge transfer happens between social media and review domains, the average performance of models stays relatively the same.
The ease of loading and using more than 6M sentiment annotated texts is as simple as the listings above. All data will be downloaded automatically; you must only appropriately filter it to your use case.

```python
import datasets
mms_dataset = datasets.load_dataset("Brand24/mms")
slavic = mms_dataset.filter(lambda row: row["Genus"] == "Slavic" and row["Polar questions"] == "interrogative word order")
```

Listing 1: Loading and filtering sentiment examples for Slavic genus with polar questions formed using interrogative word order.

```python
import datasets
mms_dataset = datasets.load_dataset("Brand24/mms")
afro_asiatic = mms_dataset.filter(lambda row: row["Family"] == "Afro-Asiatic" and row["Number of cases"] == "no morphological case-making")
```

Listing 2: Loading and filtering sentiment examples for Afro-Asiatic Language Family with no morphological case-making.

```python
import datasets
mms_dataset = datasets.load_dataset("Brand24/mms")
pl_only_social_media = mms_dataset.filter(lambda row: row["language"] == "pl" and row["domain"] == "social_media")
```

Listing 3: Loading and filtering sentiment examples for specific language and domain.

```python
import datasets
CLEANLAB_THRESHOLD = 0.7
mms_dataset = datasets.load_dataset("Brand24/mms")
afro_asian = mms_dataset.filter(lambda row: row["cleanlab_self_confidence"] >= CLEANLAB_THRESHOLD)
```

Listing 4: Loading data based on cleanlab data quality score.
D Error analysis

We present the error analysis of sentiment model results. The Tables 6 and 7 could be a little bit small, so we advise checking the editable version\(^1\) of results with the option to check ours or create your pivot tables of the results.

A couple of patterns could be spotted in Table 6:

- Japanese neutral class is consistently low performing across all models, not exceeding even 50% f1 score. Interestingly, even well-resourced languages like English still have problems with the neutral class. only one English model FT_XLM-R achieved 0.7 F1 score.
- The positive class is the easiest class to learn for the models, but the performance of neutrals and negatives varies very much, and it always needs to be verified, and potentially there could be a need to annotate more data for these classes.

Table 6: F1 scores across all languages and sentiment classes. Legend: ar - Arabic, bg - Bulgarian, bs - Bosnian, cs - Czech, de - German, en - English, es - Spanish, fa - Persian, fr - French, he - Hebrew, hi - Hindi, hr - Croatian, hu - Hungarian, it - Italian, ja - Japanese, lv - Latvian, pl - Polish, pt - Portuguese, ru - Russian, sk - Slovak, sl - Slovenian, sq - Albanian, sr - Serbian, sv - Swedish, th - Thai, ur - Urdu, zh - Chinese.

<table>
<thead>
<tr>
<th>Language</th>
<th>f1_pos</th>
<th>f1_neg</th>
<th>f1_neu</th>
</tr>
</thead>
<tbody>
<tr>
<td>en</td>
<td>0.70</td>
<td>0.80</td>
<td>0.78</td>
</tr>
<tr>
<td>de</td>
<td>0.67</td>
<td>0.70</td>
<td>0.78</td>
</tr>
<tr>
<td>thunder</td>
<td>0.87</td>
<td>0.90</td>
<td>0.78</td>
</tr>
<tr>
<td>janet</td>
<td>0.41</td>
<td>0.50</td>
<td>0.36</td>
</tr>
<tr>
<td>emma</td>
<td>0.47</td>
<td>0.79</td>
<td>0.65</td>
</tr>
<tr>
<td>sarah</td>
<td>0.84</td>
<td>0.68</td>
<td>0.74</td>
</tr>
<tr>
<td>sophia</td>
<td>0.56</td>
<td>0.77</td>
<td>0.70</td>
</tr>
<tr>
<td>pauline</td>
<td>0.73</td>
<td>0.71</td>
<td>0.79</td>
</tr>
<tr>
<td>kate</td>
<td>0.07</td>
<td>0.44</td>
<td>0.09</td>
</tr>
<tr>
<td>lisa</td>
<td>0.67</td>
<td>0.77</td>
<td>0.81</td>
</tr>
<tr>
<td>sarah</td>
<td>0.96</td>
<td>0.97</td>
<td>0.96</td>
</tr>
</tbody>
</table>

A couple of patterns could be spotted in Table 7:

- The first interesting pattern is a hard-to-learn neutral class for ‘poems’. It could be related to a tiny dataset, only with approximately 1000 English examples (693 positive, 193 negative, and 166 neutral examples).
- The review texts are easy to train for positive and negative; however, they are tough to work decently for neutral class. The reason for it could be mixing positive and negative sentiment in many neutral annotated reviews, mentioning both advantages and disadvantages of products or services.
- News domain shows interestingly high performance for the neutral class; however, it could be related to skewed distribution in this domain for the neutral class. We have neutral as many positive and negative examples summed together.

There are many other angles on how we can use our benchmarking results. We advise you to check the editable results and create your pivot tables for your own purposes and needs.

\(^{1}\)https://drive.google.com/drive/folders/1JKBslTXXiTVu9cHhQL7Fo4GdhUmoeac?usp=sharing

<table>
<thead>
<tr>
<th>Model</th>
<th>chats</th>
<th>mixed</th>
<th>news</th>
<th>poems</th>
<th>reviews</th>
<th>social_media</th>
</tr>
</thead>
<tbody>
<tr>
<td>FT_DistilmBERT</td>
<td>0.70</td>
<td>0.75</td>
<td>0.48</td>
<td>0.75</td>
<td>0.72</td>
<td>0.65</td>
</tr>
<tr>
<td>FT_LaBSE</td>
<td>0.66</td>
<td>0.59</td>
<td>0.65</td>
<td>0.80</td>
<td>0.73</td>
<td>0.73</td>
</tr>
<tr>
<td>FT_mBERT</td>
<td>0.66</td>
<td>0.80</td>
<td>0.55</td>
<td>0.67</td>
<td>0.80</td>
<td>0.69</td>
</tr>
<tr>
<td>FT_MultiNet</td>
<td>0.72</td>
<td>0.62</td>
<td>0.57</td>
<td>0.78</td>
<td>0.74</td>
<td>0.74</td>
</tr>
<tr>
<td>FT_MixSE-dot</td>
<td>0.70</td>
<td>0.64</td>
<td>0.55</td>
<td>0.67</td>
<td>0.73</td>
<td>0.73</td>
</tr>
<tr>
<td>FT_XLM-R</td>
<td>0.61</td>
<td>0.71</td>
<td>0.63</td>
<td>0.78</td>
<td>0.77</td>
<td>0.68</td>
</tr>
<tr>
<td>FT_XLM-R-dist</td>
<td>0.70</td>
<td>0.77</td>
<td>0.70</td>
<td>0.77</td>
<td>0.77</td>
<td>0.71</td>
</tr>
<tr>
<td>HB_DistilmBERT</td>
<td>0.80</td>
<td>0.42</td>
<td>0.34</td>
<td>0.74</td>
<td>0.68</td>
<td>0.42</td>
</tr>
<tr>
<td>HB_LaBSE</td>
<td>0.50</td>
<td>0.87</td>
<td>0.67</td>
<td>0.61</td>
<td>0.72</td>
<td>0.59</td>
</tr>
<tr>
<td>HB_MultiNet</td>
<td>0.66</td>
<td>0.67</td>
<td>0.57</td>
<td>0.76</td>
<td>0.50</td>
<td>0.67</td>
</tr>
<tr>
<td>HB_MixSE-dot</td>
<td>0.76</td>
<td>0.46</td>
<td>0.37</td>
<td>0.76</td>
<td>0.72</td>
<td>0.58</td>
</tr>
<tr>
<td>HB_XLM-R</td>
<td>0.76</td>
<td>0.65</td>
<td>0.45</td>
<td>0.74</td>
<td>0.74</td>
<td>0.56</td>
</tr>
<tr>
<td>HB_XLM-R-dist</td>
<td>0.70</td>
<td>0.77</td>
<td>0.70</td>
<td>0.77</td>
<td>0.77</td>
<td>0.71</td>
</tr>
<tr>
<td>HL_DistilmBERT</td>
<td>0.84</td>
<td>0.34</td>
<td>0.52</td>
<td>0.74</td>
<td>0.57</td>
<td>0.43</td>
</tr>
<tr>
<td>HL_LaBSE</td>
<td>0.70</td>
<td>0.87</td>
<td>0.67</td>
<td>0.51</td>
<td>0.72</td>
<td>0.59</td>
</tr>
<tr>
<td>HL_MultiNet</td>
<td>0.66</td>
<td>0.87</td>
<td>0.57</td>
<td>0.76</td>
<td>0.50</td>
<td>0.67</td>
</tr>
<tr>
<td>HL_MixSE-dot</td>
<td>0.76</td>
<td>0.46</td>
<td>0.37</td>
<td>0.76</td>
<td>0.72</td>
<td>0.58</td>
</tr>
<tr>
<td>HL_XLM-R</td>
<td>0.76</td>
<td>0.65</td>
<td>0.45</td>
<td>0.74</td>
<td>0.74</td>
<td>0.56</td>
</tr>
<tr>
<td>HL_XLM-R-dist</td>
<td>0.70</td>
<td>0.77</td>
<td>0.70</td>
<td>0.77</td>
<td>0.77</td>
<td>0.71</td>
</tr>
</tbody>
</table>

Note: The table contains F1 scores for various models across different domains, with columns for chats, mixed, news, poems, reviews, and social_media, and rows for different language models.