# Zero-shot Sentiment Analysis in Low-Resource Languages Using a Multilingual Sentiment Lexicon

**Fajri Koto**[1] **Tilman Beck**[2] **Zeerak Talat**[1]
**Iryna Gurevych**[1] **Timothy Baldwin**[1,3]
[1]Department Natural Language Processing, MBZUAI
[2]Ubiquitous Knowledge Processing Lab, Technical University of Darmstadt
[3]The University of Melbourne
fajri.koto@mbzuai.ac.ae

## Abstract

Improving multilingual language models capabilities in low-resource languages is generally difficult due to the scarcity of large-scale data in those languages. In this paper, we relax the reliance on texts in low-resource languages by using multilingual lexicons in pretraining to enhance multilingual capabilities. Specifically, we focus on zero-shot sentiment analysis tasks across 34 languages, including 6 high/medium-resource languages, 25 low-resource languages, and 3 code-switching datasets. We demonstrate that pretraining using multilingual lexicons, without using any sentence-level sentiment data, achieves superior zero-shot performance compared to models fine-tuned on English sentiment datasets, and large language models like GPT–3.5, BLOOMZ, and XGLM. These findings are observable for unseen low-resource languages to code-mixed scenarios involving high-resource languages.[1]

## 1 Introduction

When it comes to under-represented languages, multilingual language models (Conneau et al., 2020; Xue et al., 2021; Devlin et al., 2019; Liu et al., 2020) are often considered the most viable option in the current era of pretraining and fine-tuning, primarily due to the scarcity of labeled and unlabeled training data. However, the limited language coverage of these models often results in poor cross-lingual transfer to under-represented languages (Xia et al., 2021; Wang et al., 2022).

Prior work has extended multilingual models (Conneau et al., 2020; Xue et al., 2021) to other languages by language-adaptive pretraining (i.e., continuing to pretrain on monolingual text) (e.g., Wang et al., 2020; Chau et al., 2020) and leveraging adapters (Pfeiffer et al., 2020). However, these language adaptation techniques are not compatible with low-resource languages due to the unavailability of adequate unlabeled monolingual texts.

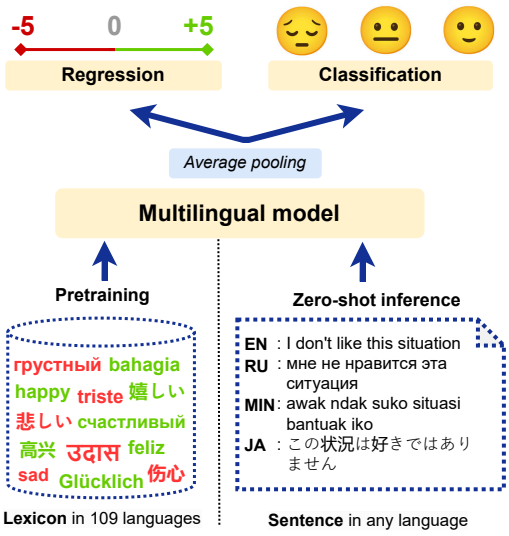


Figure 1: **Left**: pretraining with a multilingual sentiment lexicon. **Right**: zero-shot inference using sentences or documents.

Lexicons are more readily accessible and offer broader language coverage than monolingual corpora like Wikipedia and the Bible, making them a promising resource for extending multilingual models to under-represented languages. This is because when studying a new language, a lexicon is generally the first resource that field linguists develop to document its morpho-phonemics and basic vocabulary. Of the 7,000+ languages spoken worldwide, lexicons are available for approximately 70% of them, while mBERT, Wikipedia/CommonCrawl, and the Bible are available for only 1%, 4%, and 23%, respectively (Wang et al., 2022).

In prior work, Wang et al. (2022) proposed to use the Panlex translation lexicon (Baldwin et al., 2010),[2] to extend the language coverage of multilingual BERT (mBERT: Devlin et al. (2019)). They further pretrained mBERT using synthetic texts

[1]Code and dataset can be found at: https://github.com/fajri91/ZeroShotMultilingualSentiment

[2]https://panlex.org/



*Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics*
*Volume 1: Long Papers*, pages 298–320

generated through word-to-word translation, resulting in improvements in named-entity recognition tasks. Drawing inspiration from their work, our study aims to reassess the utility of sentiment lexicons in sentiment analysis tasks, which were once a prominent feature in sentiment analysis prior to the advent of pre-trained language models. Specifically, we seek to answer the following questions: (1) *To what extent do sentiment lexicons boost sentiment analysis using pretrained language models?*; and (2) *Do multilingual sentiment lexicons improve the multilingual generalizability of sentiment analysis, particularly in low-resource languages?*

We chose sentiment classification as the focus of our study for two reasons. First, there is a wealth of sentiment classification datasets across diverse languages, allowing us to conduct experiments across 34 languages, including 6 high/medium-resource languages, 25 low-resource languages, and 3 code-switching language pairs. Secondly, compared to other semantic tasks such as hate speech detection (Schmidt and Wiegand, 2017; Röttger et al., 2021) and emotion recognition (Abdul-Mageed and Ungar, 2017; Sosea and Caragea, 2020), sentiment lexicons have been studied extensively and are well-established in the field.

Cross-lingual transfer in sentiment classification is a prime case of low-resource NLP. However, existing research has predominantly focused on high/medium-resource languages (Gupta et al., 2021; Fei and Li, 2020; Lample et al., 2018), relying on sentence-level sentiment datasets in English. In this paper, we showcase how models trained on English datasets are suboptimal for low-resource languages, and introduce lexicon-based pretraining that improves multilingual sentiment modeling. Our contributions can be summarized as follows:

- Our approach is arguably cost-effective since it relies exclusively on sentiment lexicons, reducing the need for sentence-level sentiment annotation in any language, and sentence-level machine or human translation for low-resource languages, which can be challenging to access.
- We continue model pretraining using sentiment lexicons across 109 languages (see Figure 1), and demonstrate strong zero-shot performance in low-resource languages, particularly in low-resource languages that are not covered by the multilingual lexicons, and in code-mixing texts that include high-resource languages. Our approach outperforms English models fine-tuned on sentence-level sentiment datasets, as well as large language models such as XGLM (Lin et al., 2021), BLOOMZ (Muennighoff et al., 2022), and GPT–3.5 (Ouyang et al., 2022).
- We conduct comprehensive experiments in two sentiment classification scenarios: binary and 3-way classification. For each scenario, we benchmark two pretraining strategies: regression and classification. Unlike regression, the classification-based approach eases the constraint of determining the neutral class boundary before performing inference in 3-way classification in zero-shot setting.

## 2 Related Work

We briefly review three subtopics that are pertinent to this work: (1) sentiment lexicons, (2) cross-lingual adaptation for sentiment analysis, and (3) sentiment analysis in low-resource languages.

**Sentiment Lexicons** A sentiment lexicon is a curated collection of words and phrases that are classified as bearing positive or negative polarity. Such lexicons have applications in fields including NLP, cognitive science, psychology, and social science (Kiritchenko et al., 2014; Mohammad, 2018). There are two broad approaches to creating a sentiment lexicon: (1) direct annotation (Nielsen, 2011; Baccianella et al., 2010) — have annotators assign sentiment scores to individual words on a rating scale, typically ranging from $-5$ (indicating very negative) to $+5$ (indicating very positive), or based on positive/negative categorical labels (Liu et al., 2005); and (2) best–worst scaling (BWS: Kiritchenko et al. (2014, 2016); Mohammad (2018)) — have annotators select the most positive and least positive word from a collection of $n$ words, and infer a sentiment score based on the global rank of all words in the collection. BWS is considered more reliable than direct annotation as it helps mitigate annotator bias when assigning sentiment scores to individual words.

These sentiment lexicons have been constructed predominantly for English, but they also exist for languages such as Indonesian (Koto and Rahmaningtyas, 2017), Arabic (Kiritchenko et al., 2016), Persian (Dashtipour et al., 2016), Dutch (Moors et al., 2013), and Spanish (Redondo et al., 2007). In this work, we use NRC-VAD (Mohammad, 2018), which is the largest English lexicon (19,965 words)

and was built using the BWS method. Existing non-English lexicons are not just limited in size, they were also generally curated using a less reliable method (i.e., direct annotation). We instead use multilingual NRC-VAD lexicons in 108 languages, which is created by the original authors of NRC-VAD via Google Translate.[3]

**Cross-lingual Transfer in Sentiment Analysis** Most previous studies have primarily focused on cross-lingual adaptation in sentiment analysis by transferring models trained on English sentences $x_i$ and sentiment labels $y_i$ to other languages. Abdalla and Hirst (2017) developed a mapper function to convert non-English word2vec embeddings to the English embedding space (Mikolov et al., 2013). Zhou et al. (2016b,a); Wan (2009); Lambert (2015) translated English datasets to several languages, such as Chinese and Spanish, and performed joint training to improve the multilingual capabilities of the model. Fei and Li (2020) combined sentiment networks with unsupervised machine translation (Lample et al., 2018), and Meng et al. (2012); Jain and Batra (2015) have used unlabeled parallel texts in two languages to learn multilingual sentiment embeddings. In more recent work, Sun et al. (2021) used linguistic features such as language context, figurative language, and the lexification of emotional concepts to improve cross-lingual transfer, while Zhang et al. (2021) introduced a representation transformation technique from source to target languages which requires labeled English and non-English datasets.

Cross-lingual transfer in previous work relies on sentence-level labeled English datasets, and has been evaluated on high/medium-resource languages. In this work, we do not use sentence-level labeled datasets, but solely lexicons, and test our methods on low-resource languages. To the best of our knowledge, our work constitutes the first effort to perform massively multilingual sentiment pretraining using lexicons.

**Sentiment Analysis in Low-resource Languages** Most work in sentiment analysis has been applied to high/medium-resource languages, such as English (Nielsen, 2011; Baccianella et al., 2010; Koto and Adriani, 2015), Chinese (Zhou et al., 2016b,a), Japanese (Bataa and Wu, 2019), and Indonesian (Koto and Rahmaningtyas, 2017; Koto et al., 2021). There also exists a small body of work on sentiment analysis for low-resource languages. First, NusaX (Winata et al., 2023) is a parallel sentiment analysis dataset that comprises 10 local Indonesian languages, along with Indonesian and English translations. SemEval-2023 (Muhammad et al., 2022, 2023) released sentiment analysis datasets for 14 African languages. In other work, Sirajzade et al. (2020) annotated Luxembourgish sentences with sentiment labels, and Ali et al. (2021) built a sentiment lexicon for Sindhi. In this study, we include the low-resource languages of NusaX and the 14 African languages from SemEval-2023 among our test sets.

[3]The approach aligns with the utilization of bilingual lexicons such as Panlex, as demonstrated in Wang et al. (2022).

## 3 Pretraining with Sentiment Lexicons

### 3.1 Background and Problem Definition

Prior research (e.g., Zhou et al., 2016b; Zhang et al., 2021) typically assumes access to sentence-level annotated data in a source language, often English, for zero-shot cross-lingual transfer to a target language. In this work, we define zero-shot as a setting where there is no sentence-level annotated data available in the source or target languages. Instead, we use the multilingual NRC-VAD lexicon (Mohammad, 2018) which comprises words $\{w_1, w_2, .., w_n\}$ manually annotated with valence $\{v_1, v_2, .., v_n\}$, arousal $\{a_1, a_2, .., a_n\}$, and dominance $\{d_1, d_2, .., d_n\}$ scores. In this work, we train only over the valence scores $v_i$, and normalize them from a range of $[0, 1]$ to $[-5, 5]$.

Valence represents the degree of positiveness-negativeness/pleasure-displeasure and has been demonstrated to have a strong correlation with sentiment classification (Mohammad, 2018). While the valence scores are suitable for regression, we also introduce valence classes $\{s_1, s_2, .., s_n\}$ that are derived from the valence score $v_i$. For 3-way classification we set the neutral class to $[-1, 1)$, while we set 0 as the boundary between the positive and negative classes in the binary setting.

As illustrated in Figure 1, we fine-tune multilingual models (Devlin et al., 2019; Conneau et al., 2020; Liu et al., 2020; Xue et al., 2021) on the parallel NRC-VAD lexicon in 109 languages. We specifically use average pooling over all tokens prior to the regression or classification layer. During zero-shot inference, we used fine-tuned models to predict sentiment labels at the sentence level.

| | Lang | Models | | | | | | | NRC | Panlex | train/dev/test | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | mBERT | XLM-R | mBART | mT5 | BLOOMZ | XGLM | GPT–3.5 | VAD | | 3-way | Binary |
| High/Medium | en | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | 8544/1101/2210 | 6920/872/1821 |
| | ar | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | 3151/351/619 | 2162/251/428 |
| | es | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | 4802/2443/7264 | 3279/1650/5298 |
| | ru | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | 4113/726/4534 | 1205/209/1000 |
| | id | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | 3638/399/1011 | 3638/399/1011 |
| | ja | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | 3888/1112/553 | 2959/851/414 |
| Low/NusaX | ace | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✓ | 500/100/400 | 381/76/304 |
| | ban | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✓ | 500/100/400 | 381/76/304 |
| | bbc | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✓ | 500/100/400 | 381/76/304 |
| | bjn | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✓ | 500/100/400 | 381/76/304 |
| | bug | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✓ | 500/100/400 | 381/76/304 |
| | jv | ✓ | ✓ | ✗ | ✓ | ✗ | ✓ | ✓ | ✓ | ✓ | 500/100/400 | 381/76/304 |
| | mad | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | 500/100/400 | 381/76/304 |
| | min | ✓ | ✗ | ✗ | ✗ | ✗ | ✗ | ✓ | ✗ | ✓ | 500/100/400 | 381/76/304 |
| | nij | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✓ | 500/100/400 | 381/76/304 |
| | su | ✓ | ✓ | ✗ | ✓ | ✗ | ✓ | ✓ | ✓ | ✓ | 500/100/400 | 381/76/304 |
| Low/African | am | ✗ | ✓ | ✗ | ✓ | ✗ | ✓ | ✓ | ✓ | ✓ | 5984/1497/1999 | 2880/721/1775 |
| | dz | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✓ | 1651/414/958 | 1309/328/804 |
| | ha | ✗ | ✓ | ✗ | ✓ | ✗ | ✓ | ✓ | ✓ | ✓ | 14172/2677/5303 | 9260/1781/3514 |
| | ig | ✗ | ✗ | ✗ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | 10192/1841/3682 | 5684/1030/2061 |
| | kr | ✗ | ✗ | ✗ | ✗ | ✓ | ✗ | ✓ | ✓ | ✓ | 8522/2090/4515 | 2045/512/633 |
| | ma | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✓ | 5583/1215/2961 | 3422/745/1994 |
| | pcm | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✓ | 5121/1281/4154 | 5049/1260/3723 |
| | pt-MZ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | 3063/767/3662 | 1463/367/1283 |
| | sw | ✓ | ✓ | ✗ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | 1810/453/748 | 738/185/304 |
| | ts | ✗ | ✗ | ✗ | ✗ | ✓ | ✗ | ✓ | ✗ | ✓ | 804/203/254 | 668/168/211 |
| | twi | ✗ | ✗ | ✗ | ✗ | ✓ | ✗ | ✗ | ✗ | ✓ | 3481/388/949 | 2959/330/803 |
| | uo | ✓ | ✗ | ✗ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | 8522/2090/4515 | 5414/1327/2899 |
| | or | ✗ | ✓ | ✗ | ✗ | ✗ | ✓ | ✓ | ✗ | ✓ | 316/80/2096 | 218/53/1195 |
| | tg | ✗ | ✗ | ✗ | ✗ | ✗ | ✓ | ✗ | ✗ | ✓ | 318/80/2000 | 221/55/1613 |
| | aeb | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✓ | 4500/250/250 | 4284/232/235 |
| CW | en-es | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | 2449/306/307 | 1405/162/182 |
| | en-ml | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | 2856/358/335 | 2856/358/335 |
| | en-ta | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | 3233/401/398 | 3233/401/398 |

Table 1: Languages used in this paper. "✓" (green) and "✗" (red) mean that the language has and has not been seen by the models or language resources. CW indicates code-switching text. The language coverage for GPT-3.5 is derived from GPT-3 (Brown et al., 2020).

### 3.2 Extending the Lexicon

As shown in Table 2, the original NRC-VAD lexicon (Mohammad, 2018) comprises 19,965 English words, and has been extended to 108 languages by the original author resulting in 2.1M parallel words/phrases.[4]

In Table 1, we provide an overview of the languages and datasets used in this paper, categorized into: (1) high/medium-resource languages; (2) NusaX, covering local Indonesian languages (low resource); (3) African languages from SemEval 2023 (low resource); and (4) code-switching texts. The high/medium-resource languages and individual languages present in the code-switching texts are covered by all pretrained models and the NRC-VAD multilingual lexicon. However, for NusaX and the African languages, a considerable number of them are not covered.

Language coverage of the NRC-VAD multilingual lexicons remains limited in 109 languages. Therefore, we opt to extend the NRC-VAD lexicon using the Panlex lexicon, a "panlingual" lexicon containing translation edges between many languages. As shown in Table 1, only `mad` and `pt-MZ` are not covered by Panlex. Specifically, we focus on 15 languages that are not covered by NRC-VAD, and project the sentiment scores from English. Given an English word and its valence score pair $(w_i^{\text{en}}, v_i)$, we first obtain the translation of $w_i^{\text{en}}$ in language $L$. For each translation word $\{w_{i_1}^L, w_{i_2}^L, .., w_{i_m}^L\}$ we assign $v_i$ as the corresponding sentiment score. In total, we add 20K low-resource lexemes from 15 languages, as detailed in the Appendix (Table 7).

### 3.3 Filtering Lexemes

Although translating lexemes is relatively easier and often more accurate than sentences, the senti-

[4] https://saifmohammad.com/WebPages/nrc-vad.html

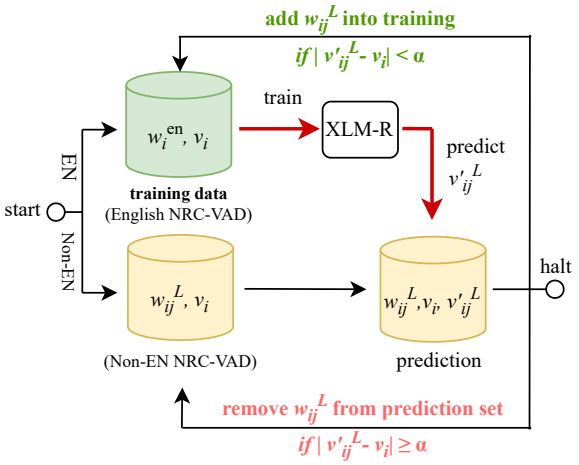


Figure 2: Lexicon filtering pipeline.

| id | Lexicon | Count |
|---|---|---|
| 1 | Original NRC-VAD | 19,965 |
| 2 | (1) + 108 translations | 2,176,185 |
| 3 | (2) + Panlex extension (15 langs) | 2,196,252 |
| 4 | (3) + Filtering | 2,071,691 |

Table 2: Statistics of the original NRC-VAD lexicon, translations, Panlex extension, and filtering.

ment score of the translated lexemes can be misleading because of word sense ambiguity. For example, the English word *cottage* refers to a small house, while the Indonesian equivalent *gubuk* "shack" from the English-Indonesian lexicon may have more negative sentiment than *cottage*.

To address the issue, we implement a filtering strategy illustrated in Figure 2. Initially, we train the English NRC-VAD $w_i^{\text{en}}$ with XLM-R (Conneau et al., 2020) using a regression approach. The training and validation data are split 80:20, with the model trained to predict the valence score $v_i$ based on the input word $w_i^{\text{en}}$. Subsequently, the model is used to predict the valence scores $v_i$ of additional lexemes (from the extended lexicons by Mohammad (2018) and Panlex). As a result, each word $w_{i_j}^{L}$ in the extended lexicon has two valence scores: the original score $v_i$ and the XLM score $\hat{v}_{i_j}^{L}$. All lexemes $w_{i_j}^{L}$ where the absolute difference $|\hat{v}_{i_j}^{L} - v_i|$ falls below a specified threshold $\alpha$ are added to the training and validation sets proportionally. This iterative process continues by training over the new extended lexicon until the number of additional words added to the training set becomes less than $\beta$.[5]

[5]We set the threshold $\alpha$ to 2.5, and $\beta$ to 1000.

# 4 Experiments

## 4.1 Data

As shown in Table 1, we use 34 languages in binary (positive, negative) and 3-way (positive, negative, neutral) classification scenarios. For binary classification, we simply remove sentences with neutral labels, resulting in a smaller dataset size. The 34 languages are grouped into 4 categories:

- **high/medium-resource languages**, including English (en: Socher et al. (2013)), Arabic (ar: Alturayeif et al. (2022)), Spanish (es: García-Vega et al. (2020)), Russian (ru: Loukachevitch et al. (2015)), Indonesian (id: Koto et al. (2020)), and Japanese (ja: Hayashibe (2020)).
- **Low-resource languages from NusaX** (Winata et al., 2023), consisting of 10 local Indonesian languages: Acehnese (ace), Balinese (ban), Batak Toba (bbc), Banjarese (bjn), Buginese (bug), Madurese (mad), Minangkabau (min), Javanese (jv), Ngaju (nij), and Sundanese (su).
- **Low-resource African languages**, based on the 14 languages of SemEval-2023 (Muhammad et al., 2022, 2023): Amharic (am), Algerian Arabic (dz), Hausa (ha), Igbo (ig), Kinyarwanda (kr), Darija (ma), Nigerian Pidgin (pcm), Mozambique Portuguese (pt-MZ),[6] Swahili (sw), Xitsonga (ts), Twi (twi), Yoruba (yo), Oromo (or), and Tigrinya (tg). We additionally include Tunisian Arabizi (aeb) from Fourati et al. (2021).
- **Code-switching texts**, involving English–Spanish (en–es: Vilares et al. (2016)), English–Malayalam (en–ml: Chakravarthi et al. (2020a)), and English–Tamil (en–ta: Chakravarthi et al. (2020b))

## 4.2 Set-Up

We perform comprehensive evaluation over six multilingual encoder and encoder–decoder pretrained language models: (1) $\text{mBERT}_{\text{Base}}$ (Devlin et al., 2019); (2) $\text{XLM-R}_{\text{Base}}$ (Conneau et al., 2020); (3) $\text{XLM-R}_{\text{Large}}$ (Conneau et al., 2020); (4) $\text{mBART}_{\text{Large}}$ (Liu et al., 2020); (5) $\text{mT5}_{\text{Base}}$ (Xue et al., 2021); and (6) $\text{mT5}_{\text{Large}}$ (Xue et al., 2021). We evaluate different scenarios, including:

[6]Mozambican Portuguese dialect differs in both pronunciation and colloquial vocabulary from standard European Portuguese.

(1) lexicon-based pretraining via regression vs. classification; and (2) binary vs. 3-way classification.

**Lexicon-based pretraining** We conduct pretraining on the six multilingual pretrained language models using three combinations of multilingual lexicons: (1) NRC-VAD; (2) NRC-VAD + Panlex; and (3) NRC-VAD + Panlex + filtering. For regression-based pretraining, we use mean square error (MSE) loss, while for binary and 3-way classification, we use cross-entropy loss. Please see the Appendix for detailed hyper-parameter settings and computational resources.

**Full Training, Few-shot and Zero-shot** Following the lexicon-based pretraining, we examine its impact on sentence-level sentiment analysis across three scenarios: (1) full training, (2) few-shot (training with limited data), and (3) zero-shot. For the first setting, we fine-tune the model with the complete training and development set of sentence-level sentiment data for each language listed in Table 1. For the second, we simulate few-shot training by randomly sampling 100 training and 50 development instances. To ensure robustness and account for variability, we repeat the experiment five times using different random seeds, and report the average performance. Please note that these first two settings are our preliminary experiments and we report the average scores of $\text{mBERT}_{\text{Base}}$ across 34 languages. Our main experiment in this work is zero-shot setting, simulating real-world scenarios for low-resource languages where no sentence-level sentiment data is available. For each of the six models, we present the average score for each language group in Section 4.1.

**Baselines** In both full and few-shot training scenarios, the baseline consists of vanilla models without lexicon-based pretraining. For zero-shot setting, we compare our approaches with (1) models trained on SST datasets (Socher et al., 2013) – a sentence-level English sentiment data; and (2) prompting via LLMs, including BLOOMZ (3B) (Muennighoff et al., 2022), XGLM (2.9B) (Lin et al., 2021), and GPT–3.5 (175B) (Ouyang et al., 2022).[7] The first baseline is zero-shot cross-lingual transfer, following prior work (Abdalla and Hirst, 2017; Zhang et al., 2021) that used English as the main training language. For robustness, we fine-tuned the models with five different seeds for the first baseline. For the LLMs, we average the results of six English prompts as detailed in the Appendix. We report weighted macro-F1 scores for all experiments.

[7]We do not include Llama-2 (Touvron et al., 2023) and Falcon (Penedo et al., 2023) as they are English-centric models.

| Models | Binary | 3-way |
|---|---|---|
| **Full training** | | |
| $\text{mBERT}_{\text{Base}}$ | 81.95 | 70.89 |
| + EN Lex. | 82.49 | 71.05 |
| + ML Lex. | 82.84 | 71.81 |
| + ML Lex. + Panlex | **83.40** | 71.82 |
| + ML Lex. + Panlex + Filtering | **83.39** | **71.98** |
| **Training with limited data** | | |
| $\text{mBERT}_{\text{Base}}$ | 68.84 | 56.76 |
| + EN Lex. | 72.47 | 60.58 |
| + ML Lex. | 75.05 | 61.42 |
| + ML Lex. + Panlex | 75.34 | 61.78 |
| + ML Lex. + Panlex + Filtering | **75.39** | **61.92** |

Table 3: Preliminary results, based on averaged macro-F1 scores across 34 languages. "EN Lex." and "ML Lex." indicate the English and multilingual NRC-VAD lexicons.

Work discussed in Section 2 is not suitable as a baseline due to the absence of word embeddings and machine translation systems in low-resource languages. For instance, Abdalla and Hirst (2017) require word2vec embeddings in the target language, while Zhou et al. (2016b,a); Wan (2009); Lambert (2015); Zhang et al. (2021) rely on sentence-level machine translation. Additionally, Meng et al. (2012); Jain and Batra (2015) require unlabeled parallel texts, which are not consistently available for low-resource languages.

### 4.3 Results

**Preliminary Results: Full and Few-shot Training** Table 3 shows the average performance of $\text{mBERT}_{\text{Base}}$ when training with full and limited training data at the sentence-level. Here we compare the vanilla multilingual model against four lexicon-based pretraining models, and its extensions (Panlex and Filtering). For each language in Table 1, we fine-tune the models and measure the macro-F1 score over the test set. For this preliminary experiment, we only use regression in lexicon-based pretraining for the binary and 3-way classification tasks. The results demonstrate that lexicon-based pretraining enhances performance, surpassing vanilla $\text{mBERT}_{\text{Base}}$ in both binary and 3-way classification settings. The proposed filtering method further slightly improves performance.

The improvements shown in Table 3 are particularly noticeable in few-shot training, with increases

| Model | Binary | | | | | 3-way | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | HM-R | NusaX | African | CS | AVERAGE | HM-R | NusaX | African | CS | AVERAGE |
| XGLM (2.9B) | 59.66 | 49.34 | 42.50 | 52.61 | 51.03 | 38.09 | 33.47 | 25.72 | 50.08 | 36.84 |
| BLOOMZ (3B) | 77.82 | 69.85 | 54.92 | 45.89 | 62.12 | 48.43 | 48.89 | 33.81 | 35.85 | 41.74 |
| GPT–3.5 (175B) | 77.50 | 63.90 | 53.82 | 73.66 | 67.22 | 67.65 | 48.50 | 38.13 | 50.41 | 51.17 |
| **mBERT$_{\text{Base}}$ (110M)** | | | | | | | | | | |
| + SST (sentence-level data) | 66.87 | 44.96 | 46.56 | 44.50 | 50.72 | 46.89 | 28.94 | 27.80 | 24.85 | 32.12 |
| + ML Lex. | 74.57 | 67.92 | 57.79 | 69.43 | 67.43 | 55.72 | 44.18 | 35.08 | 60.14 | 48.78 |
| + ML Lex. + Panlex | 74.93 | 66.71 | 58.99 | 71.58 | 68.05 | 55.42 | 43.47 | 35.13 | 61.27 | 48.82 |
| + ML Lex. + Panlex + Filtering | 74.74 | 63.95 | 58.00 | 71.88 | 67.14 | 54.21 | 39.18 | 33.42 | 58.81 | 46.40 |
| **XLM-R$_{\text{Base}}$ (270M)** | | | | | | | | | | |
| + SST (sentence-level data) | 85.51 | 59.50 | 56.59 | 49.84 | 62.86 | 68.27 | 41.34 | 35.83 | 36.30 | 45.43 |
| + ML Lex. | 82.13 | 70.94 | 61.45 | 62.44 | 69.24 | 60.52 | 33.85 | 36.31 | 38.84 | 42.38 |
| + ML Lex. + Panlex | 82.74 | 73.59 | 63.10 | 63.97 | 70.85 | 58.81 | 46.34 | 41.95 | 60.31 | 51.85 |
| + ML Lex. + Panlex + Filtering | 82.88 | 73.47 | 63.99 | 72.50 | 73.21 | 61.28 | 33.33 | 34.87 | 40.13 | 42.40 |
| **XLM-R$_{\text{Large}}$ (550M)** | | | | | | | | | | |
| + SST (sentence-level data) | **88.39** | 73.00 | 61.75 | 54.38 | 69.38 | **70.57** | 49.63 | 37.39 | 36.71 | 48.58 |
| + ML Lex. | 84.55 | 78.01 | 66.16 | 72.53 | 75.31 | 61.78 | 45.95 | 41.84 | 63.52 | 53.27 |
| + ML Lex. + Panlex | 84.89 | 70.79 | 63.85 | 76.47 | 74.00 | 64.38 | 52.84 | 42.47 | 64.73 | **56.10** |
| + ML Lex. + Panlex + Filtering | 84.20 | 66.75 | 62.28 | **78.32** | 72.89 | 64.74 | 46.67 | 43.20 | 59.59 | 53.55 |
| **mBART$_{\text{Large}}$ (600M)** | | | | | | | | | | |
| + SST (sentence-level data) | 85.41 | 65.41 | 59.62 | 62.61 | 68.26 | 66.58 | 31.80 | 27.99 | 31.52 | 39.47 |
| + ML Lex. | 83.26 | 72.26 | 62.89 | 74.10 | 73.13 | 61.25 | 41.88 | 39.18 | 54.43 | 49.19 |
| + ML Lex. + Panlex | 81.97 | 74.86 | 62.74 | 65.00 | 71.14 | 61.16 | 35.76 | 31.61 | 49.51 | 44.51 |
| + ML Lex. + Panlex + Filtering | 80.66 | 61.71 | 58.28 | 78.02 | 69.67 | 57.48 | 30.50 | 30.24 | 40.77 | 39.75 |
| **mT5$_{\text{Base}}$ (580M)** | | | | | | | | | | |
| + SST (sentence-level data) | 83.16 | 55.33 | 57.18 | 48.39 | 61.02 | 62.37 | 35.58 | 37.59 | 31.04 | 41.64 |
| + ML Lex. | 81.29 | 75.84 | 67.45 | 73.63 | 74.55 | 59.27 | 51.63 | 43.88 | 60.04 | 53.71 |
| + ML Lex. + Panlex | 79.57 | 71.37 | 66.81 | 75.60 | 73.34 | 57.14 | 50.22 | 44.62 | 58.51 | 52.62 |
| + ML Lex. + Panlex + Filtering | 82.24 | 75.52 | 67.52 | 76.33 | 75.40 | 61.72 | 45.92 | 44.45 | 59.09 | 52.79 |
| **mT5$_{\text{Large}}$** (1B) | | | | | | | | | | |
| + SST (sentence-level data) | 84.74 | 60.68 | 58.67 | 47.91 | 63.00 | 48.05 | 31.67 | 31.53 | 24.75 | 34.00 |
| + ML Lex. | 83.69 | **78.26** | 69.28 | 72.62 | 75.96 | 62.15 | 51.84 | 44.42 | 61.43 | 54.96 |
| + ML Lex. + Panlex | 82.78 | 76.70 | **70.05** | 75.32 | **76.21** | 59.59 | **53.96** | **45.12** | 62.05 | 55.18 |
| + ML Lex. + Panlex + Filtering | 81.35 | 73.37 | 68.04 | 75.43 | 74.54 | 59.52 | 46.37 | 42.75 | 60.54 | 52.29 |

Table 4: Full zero-shot results. The underlined score indicates the highest performance within the respective group, while scores in **bold** indicate the best global performance. "HM-R" = high/medium-resource languages, excluding English, "CS" = code-switched text, and "ML Lex." indicates the multilingual NRC-VAD lexicon. "SST (sentence-level data)" is cross-lingual zero-shot transfer that is trained on English sentence-level sentiment data.

of $+6.6$ and $+5.2$ for binary and 3-way classification, respectively. In the full training scenario, the increments are smaller, at only $+1.4$ and $+1.1$. These findings motivate us to further investigate zero-shot settings using all six multilingual models.

**Zero-shot Results in High/Medium-Resource Languages** Table 4 presents the averaged zero-shot performance of all models categorized by four language groups. The reported results use regression and classification in lexicon-based pretraining for binary and 3-way classification, respectively. In the case of high/medium-resource languages (HM-R), English is excluded to ensure a fair comparison with models fine-tuned on the English SST dataset. Overall, we observe that multilingual models fine-tuned on SST tend to perform the best in high/medium-resource languages, with mBERT$_{\text{Base}}$ being the exception. It is not surprising to see these multilingual models outperforming the LLMs (the first three rows), as they are specifically fine-tuned on a sentence-level dataset.

Interestingly, we also observe that most of the lexicon-based pretrained models substantially outperform the LLMs. For instance, XLM-R$_{\text{Large}}$ outperforms GPT–3.5 and XGLM by $+7$ and $+24.9$ in binary classification. In 3-way classification, GPT–3.5 tends to perform better than lexicon-based pretraining, while XGLM and BLOOMZ tend to perform poorly. It's important to note that our models are significantly smaller in size, and BLOOMZ has been fine-tuned on multilingual sentiment analysis datasets, such as Amazon reviews (Muennighoff et al., 2022).

**Zero-shot Results in Low Resource Languages** For low-resource languages, models fine-tuned on

SST (i.e., sentence-level English dataset) underperform lexicon-based pretraining by a wide margin in both binary and 3-way classification settings. Notably, despite its significantly smaller size, lexicon-based pretraining to outperform larger models like BLOOMZ and GPT–3.5. Among the models, $\text{mT5}_{\text{Large}}$ achieves the best performance in NusaX and African languages for both classification scenarios, with disparities ranging from $+8.5$ to $+15$ and $+5$ to $+7$ when compared to LLMs. The impact of incorporating Panlex and/or filtering varies across models, with notable improvements observed for $\text{XLM-R}_{\text{Base}}$ and $\text{mT5}_{\text{Large}}$.

Expanding the multilingual lexicon with Panlex tends to improve the zero-shot capability for 3-way classification. This can be attributed to the fact that NusaX and African languages have a relatively small number of new lexemes (9.5K and 8.5K, respectively). Moreover, Panlex has English as the primary source language, making it inadequate to capture the diversity of languages in our experiments.

Although adding Panlex with the filtering method showed improvements in the preliminary experiment (see the full training results in Table 3), it does not enhance the zero-shot performance in NusaX and African languages. To investigate this, we conducted a manual analysis of 100 randomly-selected samples from the 124K filtered lexemes. We compared the original sentiment scores of the corresponding English lexicon with the predicted scores generated by our filtering model. Upon back-translating the non-English words to English, we found that 63 of the original scores were either correct or better than the predicted scores, 25 predicted scores were better than the original scores, and 12 were incorrect for both. Additionally, we identified 75 unique languages among the 100 samples, indicating that our English-centric filtering might not be effective in improving low-resource languages.

**Zero-shot Results in Code-switched Text** Extending the lexicon with Panlex and the filtering method yields the best performance for code-switched text, surpassing LLMs and models fine-tuned on SST. In binary classification, our method achieves an average F1-score that is $+24.1$ higher than the models fine-tuned on SST, while in 3-way classification, our method achieves F1-scores that are $+10$ to $+20$ higher than LLMs, even though the individual languages in our code-switched texts are high-resource (i.e., English, Spanish, Tamil, and Malayalam).

| Models + ML Lex. | Binary | | 3-way | |
|---|---|---|---|---|
| | Reg. | Class. | Reg. | Class. |
| $\text{mBERT}_{\text{Base}}$ | **67.43** | 66.29 | 48.62 | **48.78** |
| $\text{XLM-R}_{\text{Base}}$ | **69.24** | 64.37 | **49.27** | 42.38 |
| $\text{XLM-R}_{\text{Large}}$ | **75.31** | 74.54 | **53.69** | 53.27 |
| $\text{mBART}_{\text{Large}}$ | **73.13** | 71.23 | 49.08 | **49.19** |
| $\text{mT5}_{\text{Base}}$ | **74.55** | 68.27 | 19.84 | **53.71** |
| $\text{mT5}_{\text{Large}}$ | **75.96** | 72.21 | 20.94 | **54.96** |

Table 5: Regression vs. classification in lexicon-based pretraining for zero-shot sentiment analysis.

| Model | Stance | Hate Speech | Emotion |
|---|---|---|---|
| *Binary classes* | | | |
| $\text{mBERT}_{\text{Base}}$ | 70.27 | 69.11 | 58.96 |
| + EN Lex. | **73.25** | **71.35** | **75.89** |
| *Original classes* | | | |
| $\text{mBERT}_{\text{Base}}$ | 52.55 | 56.36 | 18.44 |
| + EN Lex. | **53.04** | **57.59** | **23.93** |

Table 6: Lexicon-based pretraining performance (macro-F1) over stance detection, hate speech detection, and emotion classification. The results are based on the limited training data scenario.

## 5 Analysis

**Regression vs. classification in lexicon-based pretraining** In Table 5 we present the average performance across the four language groups to compare the effectiveness of lexicon-based pretraining in regression and classification tasks for both binary and 3-way classification. Our findings indicate that regression performs better for binary classification, while classification leads to better results for 3-way classification. However, regression in 3-way classification presents a challenge when determining the neutral class boundary during inference. In the zero-shot setting, we lack specific data for hyper-parameter tuning, leading us to arbitrarily set the neutral class boundaries to $-1$ and $+1$. Although this setting works reasonably well for XLM-R, it yields poor performance for mT5. A manual analysis of mT5's predictions revealed that they tend to cluster around zero.

**Performance over unseen low-resource languages** We compute the average results for languages that are completely unseen by all models, including 7 NusaX languages (`ace`, `ban`, `bbc`, `bjn`, `bug`, `mad`, `nij`) and 4 African languages (`dz`, `ma`, `pcm`, `pt-MZ`, `aeb`). We exclude lexicon-based pre-

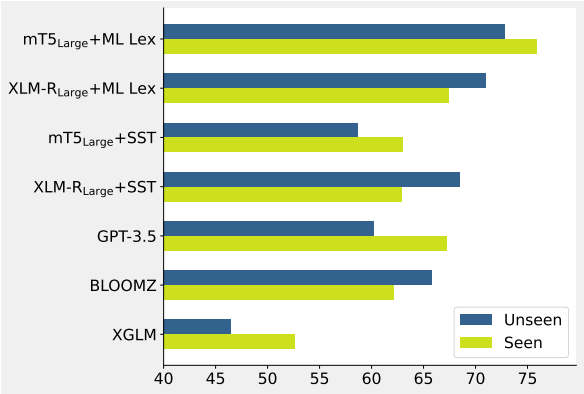


Figure 3: Average zero-shot performance of seen and unseen languages in binary classification across different models.

training with the Panlex extension since its performance is poor for low-resource languages. As a comparison, we include code-switched text as seen languages for the models. Figure 3 presents the performance of XLM-$R_{Large}$ and $mT5_{Large}$, which outperform the LLMs and models fine-tuned on SST. This suggests that the multilingual sentiment lexicon is effective at enhancing language generalization for low-resource languages.

## 6 Discussion

Given the positive results, we explore the potential applicability of our methodology to NLP tasks beyond sentiment analysis, offering valuable directions for future research. We examine if the lexicon-based pretraining yields benefits in other semantic tasks, including stance detection (Li et al., 2021), hate speech detection (Vidgen et al., 2021), and emotion classification (Demszky et al., 2020).

For each task, we take two datasets and perform experiments in few-shot training using $mBERT_{Base}$, following the setup described in Section 4.2. Instead of using the multilingual lexicon, we use the English NRC-VAD lexicon since all the data is in English. For detailed information about the datasets and results, see the Appendix. Table 6 shows the average F1 scores on each task, demonstrating that lexicon-based pretraining boosts the performance of vanilla $mBERT_{Base}$. Particularly noteworthy are the substantial improvements in the emotion classification task, with increments of $+16.9$ and $+5.5$ for the binary and original class settings, respectively. These findings highlight the potential of sentiment lexicons for various semantic tasks, particularly in the context of investigating their effectiveness in low-resource languages in future works.

## 7 Conclusion

We have demonstrated the efficacy of employing a multilingual sentiment lexicon for achieving multilingual generalization in language model pretraining. Without utilizing sentence-level datasets in any language, we provide compelling evidence of superior zero-shot performance in sentiment analysis tasks for low-resource languages, surpassing the performance of large language models. These findings open up new avenues for research in the realm of low-resource languages, not only for language understanding but also language generation tasks. Our results encourage further exploration and investigation of this exciting research direction.

## Limitations

This research focuses on general sentiment analysis, and we acknowledge that aspect-based sentiment analysis is a more fine-grained and expressive way of capturing sentiment, that warrants further exploration. Unfortunately, due to the scarcity of relevant datasets in low-resource languages, and task complexity, we were unable to explore aspect-based sentiment analysis in this work.

Regarding the proposed technique, we acknowledge four notable limitations. Firstly, due to the distinct nature of training (i.e., lexicon-level) and inference (i.e., sentence-level), our model may lack sensitivity to semantic complexity at the sentence level, encompassing nuances such as negation and sentences conveying multiple sentiments. One way to address this is to expand the NRC-VAD lexicon to include phrases, metaphors, culturally relevant words, and syntehtic sentences derived from the lexicon. Secondly, our lexicon-based pretraining is solely based on valence scores, and there is an intriguing avenue to explore the inclusion of dominance and arousal scores. Thirdly, the use of machine translation systems for translating lexicons may introduce errors in both translation and sentiment scoring. While translating lexicons is arguably less complex than translating entire sentences, a comprehensive error analysis of the translated lexicons and Panlex words could offer valuable insights into the quality of the additional lexicons. Fourthly, our filtering method (Figure 2) proves less effective in certain scenarios due to its English-centric nature. This limitation

arises because the initial filtering model is exclusively trained using English lexicons. To enhance this method, we propose that incorporating lexicons manually annotated in more diverse languages could significantly improve its efficacy.

## Ethical Considerations

When conducting sentiment analysis in low-resource languages, there are several important considerations that warrant reflection. First, it is crucial to ensure that the work benefits the local community rather than solely exploiting the language. In the era of large language models, the lack of computing resources often hinders the deployment of such systems in regions or countries where the language is spoken. Secondly, sentiment analysis can be subject to cultural ambiguity. Relying solely on European-centric multilingual models for sentiment prediction may introduce biases and produce inappropriate model predictions in certain cultural contexts. Therefore, cultural sensitivity and awareness are essential factors to address when conducting sentiment analysis in low-resource languages, which we leave for future work.

## References


Mohamed Abdalla and Graeme Hirst. 2017. Cross-lingual sentiment analysis without (good) translation. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 506–515, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Muhammad Abdul-Mageed and Lyle Ungar. 2017. EmoNet: Fine-grained emotion detection with gated recurrent neural networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 718–728, Vancouver, Canada. Association for Computational Linguistics.

Wazir Ali, Naveed Ali, Yong Dai, Jay Kumar, Saifullah Tumrani, and Zenglin Xu. 2021. Creating and evaluating resources for sentiment analysis in the low-resource language: Sindhi. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 188–194, Online. Association for Computational Linguistics.

Nora Saleh Alturayeif, Hamzah Abdullah Luqman, and Moataz Aly Kamaleldin Ahmed. 2022. Mawqif: A multi-label Arabic dataset for target-specific stance detection. In *Proceedings of the The Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 174–184, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Timothy Baldwin, Jonathan Pool, and Susan Colowick. 2010. PanLex and LEXTRACT: Translating all words of all languages of the world. In *Coling 2010: Demonstrations*, pages 37–40, Beijing, China. Coling 2010 Organizing Committee.

Enkhbold Bataa and Joshua Wu. 2019. An investigation of transfer learning-based sentiment analysis in Japanese. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4652–4657, Florence, Italy. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Bharathi Raja Chakravarthi, Navya Jose, Shardul Suryawanshi, Elizabeth Sherly, and John Philip McCrae. 2020a. A sentiment analysis dataset for code-mixed Malayalam-English. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 177–184, Marseille, France. European Language Resources association.

Bharathi Raja Chakravarthi, Vigneshwaran Muralidaran, Ruba Priyadharshini, and John Philip McCrae. 2020b. Corpus creation for sentiment analysis in code-mixed Tamil-English text. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 202–210, Marseille, France. European Language Resources association.

Ethan C. Chau, Lucy H. Lin, and Noah A. Smith. 2020. Parsing with multilingual BERT, a small corpus, and a small treebank. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1324–1334, Online. Association for Computational Linguistics.

Costanza Conforti, Jakob Berndt, Mohammad Taher Pilehvar, Chryssi Giannitsarou, Flavio Toxvaerd, and

Nigel Collier. 2020. Will-they-won't-they: A very large dataset for stance detection on Twitter. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1715–1724, Online. Association for Computational Linguistics.

Alexis Conneau, 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*, pages 8440–8451, Online. Association for Computational Linguistics.

Kia Dashtipour, Amir Hussain, Qiang Zhou, Alexander Gelbukh, Ahmad YA Hawalah, and Erik Cambria. 2016. PerSent: A freely available Persian sentiment lexicon. In *Advances in Brain Inspired Cognitive Systems: 8th International Conference, BICS 2016, Beijing, China, November 28-30, 2016, Proceedings 8*, pages 310–320. Springer.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. GoEmotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054, Online. Association for Computational Linguistics.

Jacob Devlin, 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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Hongliang Fei and Ping Li. 2020. Cross-lingual unsupervised sentiment classification with multi-view transfer learning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5759–5771, Online. Association for Computational Linguistics.

Antigoni Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of Twitter abusive behavior. In *Proceedings of the International AAAI Conference on Web and Social Media*.

Chayma Fourati, Hatem Haddad, Abir Messaoudi, Moez BenHajhmida, Aymen Ben Elhaj Mabrouk, and Malek Naski. 2021. Introducing a large Tunisian Arabizi dialectal dataset for sentiment analysis. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 226–230, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Manuel García-Vega, MC Díaz-Galiano, MA García-Cumbreras, FMP Del Arco, A Montejo-Ráez, SM Jiménez-Zafra, E Martínez Cámara, CA Aguilar, MAS Cabezudo, L Chiruzzo, et al. 2020. Overview of TASS 2020: Introducing emotion detection. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020) Co-Located with 36th Conference of the Spanish Society for Natural Language Processing (SEPLN 2020), Málaga, Spain*, pages 163–170.

Akshat Gupta, Sai Krishna Rallabandi, and Alan W Black. 2021. Task-specific pre-training and cross lingual transfer for sentiment analysis in Dravidian code-switched languages. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 73–79, Kyiv. Association for Computational Linguistics.

Yuta Hayashibe. 2020. Japanese realistic textual entailment corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6827–6834, Marseille, France. European Language Resources Association.

Sarthak Jain and Shashank Batra. 2015. Cross lingual sentiment analysis using modified BRAE. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 159–168, Lisbon, Portugal. Association for Computational Linguistics.

Svetlana Kiritchenko, Saif Mohammad, and Mohammad Salameh. 2016. SemEval-2016 task 7: Determining sentiment intensity of English and Arabic phrases. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 42–51, San Diego, California. Association for Computational Linguistics.

Svetlana Kiritchenko, Xiaodan Zhu, and Saif M Mohammad. 2014. Sentiment analysis of short informal texts. *Journal of Artificial Intelligence Research*, 50:723–762.

Fajri Koto and Mirna Adriani. 2015. A comparative study on Twitter sentiment analysis: Which features are good? In *Natural Language Processing and Information Systems: 20th International Conference on Applications of Natural Language to Information Systems, NLDB 2015, Passau, Germany, June 17-19, 2015, Proceedings 20*, pages 453–457. Springer.

Fajri Koto and Ikhwan Koto. 2020. Towards computational linguistics in Minangkabau language: Studies on sentiment analysis and machine translation. In *Proceedings of the 34th Pacific Asia Conference on Language, Information and Computation*, pages 138–148, Hanoi, Vietnam. Association for Computational Linguistics.

Fajri Koto, Jey Han Lau, and Timothy Baldwin. 2021. IndoBERTweet: A pretrained language model for Indonesian Twitter with effective domain-specific vocabulary initialization. In *Proceedings of the 2021*

*Conference on Empirical Methods in Natural Language Processing*, pages 10660–10668, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Fajri Koto, Afshin Rahimi, Jey Han Lau, and Timothy Baldwin. 2020. IndoLEM and IndoBERT: A benchmark dataset and pre-trained language model for Indonesian NLP. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 757–770, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Fajri Koto and Gemala Y Rahmaningtyas. 2017. Inset lexicon: Evaluation of a word list for Indonesian sentiment analysis in microblogs. In *2017 International Conference on Asian Language Processing (IALP)*, pages 391–394. IEEE.

Patrik Lambert. 2015. Aspect-level cross-lingual sentiment classification with constrained SMT. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 781–787, Beijing, China. Association for Computational Linguistics.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Unsupervised machine translation using monolingual corpora only. In *International Conference on Learning Representations*, Vancouver, Canada.

Yingjie Li, Tiberiu Sosea, Aditya Sawant, Ajith Jayaraman Nair, Diana Inkpen, and Cornelia Caragea. 2021. P-stance: A large dataset for stance detection in political domain. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2355–2365, Online. Association for Computational Linguistics.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, et al. 2021. Few-shot learning with multilingual language models. *arXiv preprint arXiv:2112.10668*.

Bing Liu, Minqing Hu, and Junsheng Cheng. 2005. Opinion observer: analyzing and comparing opinions on the web. In *Proceedings of the 14th international conference on World Wide Web*, pages 342–351.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Natalia Loukachevitch, Pavel Blinov, Evgeny Kotelnikov, Yulia Rubtsova, Vladimir Ivanov, and Elena Tutubalina. 2015. SentiRuEval: testing object-oriented sentiment analysis systems in Russian. In *Proceedings of International Conference Dialog*, volume 2, pages 3–13.

Xinfan Meng, Furu Wei, Xiaohua Liu, Ming Zhou, Ge Xu, and Houfeng Wang. 2012. Cross-lingual mixture model for sentiment classification. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 572–581, Jeju Island, Korea. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *International Conference on Learning Representations*.

Saif Mohammad. 2018. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 174–184, Melbourne, Australia. Association for Computational Linguistics.

Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. SemEval-2018 task 1: Affect in tweets. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 1–17, New Orleans, Louisiana. Association for Computational Linguistics.

Agnes Moors, Jan De Houwer, Dirk Hermans, Sabine Wanmaker, Kevin Van Schie, Anne-Laura Van Harmelen, Maarten De Schryver, Jeffrey De Winne, and Marc Brysbaert. 2013. Norms of valence, arousal, dominance, and age of acquisition for 4,300 dutch words. *Behavior research methods*, 45:169–177.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, et al. 2022. Crosslingual generalization through multitask finetuning. *arXiv preprint arXiv:2211.01786*.

Shamsuddeen Hassan Muhammad, Idris Abdulmumin, Seid Muhie Yimam, David Ifeoluwa Adelani, Ibrahim Sa'id Ahmad, Nedjma Ousidhoum, Abinew Ali Ayele, Saif M. Mohammad, Meriem Beloucif, and Sebastian Ruder. 2023. SemEval-2023 Task 12: Sentiment Analysis for African Languages (AfriSenti-SemEval). In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*. Association for Computational Linguistics.

Shamsuddeen Hassan Muhammad, David Ifeoluwa Adelani, Sebastian Ruder, Ibrahim Sa'id Ahmad, Idris Abdulmumin, Bello Shehu Bello, Monojit Choudhury, Chris Chinenye Emezue, Saheed Salahudeen Abdullahi, Anuoluwapo Aremu, Alípio Jorge, and Pavel Brazdil. 2022. NaijaSenti: A Nigerian Twitter sentiment corpus for multilingual sentiment analysis. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 590–602, Marseille, France. European Language Resources Association.

Finn Årup Nielsen. 2011. A new anew: Evaluation of a word list for sentiment analysis in microblogs. In *Workshop on'Making Sense of Microposts: Big things come in small packages*, pages 93–98.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The RefinedWeb dataset for Falcon LLM: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116*.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020. MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7654–7673, Online. Association for Computational Linguistics.

Jaime Redondo, Isabel Fraga, Isabel Padrón, and Montserrat Comesaña. 2007. The spanish adaptation of anew (affective norms for english words). *Behavior research methods*, 39(3):600–605.

Paul Röttger, Bertie Vidgen, Dong Nguyen, Zeerak Waseem, Helen Margetts, and Janet Pierrehumbert. 2021. HateCheck: Functional tests for hate speech detection models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 41–58, Online. Association for Computational Linguistics.

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10, Valencia, Spain. Association for Computational Linguistics.

Joshgun Sirajzade, Daniela Gierschek, and Christoph Schommer. 2020. An annotation framework for Luxembourgish sentiment analysis. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 172–176, Marseille, France. European Language Resources association.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Tiberiu Sosea and Cornelia Caragea. 2020. CancerEmo: A dataset for fine-grained emotion detection. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8892–8904, Online. Association for Computational Linguistics.

Jimin Sun, Hwijeen Ahn, Chan Young Park, Yulia Tsvetkov, and David R. Mortensen. 2021. Cross-cultural similarity features for cross-lingual transfer learning of pragmatically motivated tasks. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2403–2414, Online. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Bertie Vidgen, Tristan Thrush, Zeerak Waseem, and Douwe Kiela. 2021. Learning from the worst: Dynamically generated datasets to improve online hate detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1667–1682, Online. Association for Computational Linguistics.

David Vilares, Miguel A. Alonso, and Carlos Gómez-Rodríguez. 2016. EN-ES-CS: An English-Spanish code-switching Twitter corpus for multilingual sentiment analysis. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4149–4153, Portorož, Slovenia. European Language Resources Association (ELRA).

Xiaojun Wan. 2009. Co-training for cross-lingual sentiment classification. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 235–243, Suntec, Singapore. Association for Computational Linguistics.

Xinyi Wang, Sebastian Ruder, and Graham Neubig. 2022. Expanding pretrained models to thousands more languages via lexicon-based adaptation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 863–877, Dublin, Ireland. Association for Computational Linguistics.

Zihan Wang, Karthikeyan K, Stephen Mayhew, and Dan Roth. 2020. Extending multilingual BERT to low-resource languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages

2649–2656, Online. Association for Computational Linguistics.

Genta Indra Winata, Alham Fikri Aji, Samuel Cahyawijaya, Rahmad Mahendra, Fajri Koto, Ade Romadhony, Kemal Kurniawan, David Moeljadi, Radityo Eko Prasojo, Pascale Fung, Timothy Baldwin, Jey Han Lau, Rico Sennrich, and Sebastian Ruder. 2023. NusaX: Multilingual parallel sentiment dataset for 10 Indonesian local languages. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 815–834, Dubrovnik, Croatia. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Mengzhou Xia, Guoqing Zheng, Subhabrata Mukherjee, Milad Shokouhi, Graham Neubig, and Ahmed Hassan Awadallah. 2021. MetaXL: Meta representation transformation for low-resource cross-lingual learning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 499–511, Online. Association for Computational Linguistics.

Linting Xue, 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*, pages 483–498, Online. Association for Computational Linguistics.

Wenxuan Zhang, Ruidan He, Haiyun Peng, Lidong Bing, and Wai Lam. 2021. Cross-lingual aspect-based sentiment analysis with aspect term code-switching. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9220–9230, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xinjie Zhou, Xiaojun Wan, and Jianguo Xiao. 2016a. Attention-based LSTM network for cross-lingual sentiment classification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 247–256, Austin, Texas. Association for Computational Linguistics.

Xinjie Zhou, Xiaojun Wan, and Jianguo Xiao. 2016b. Cross-lingual sentiment classification with bilingual document representation learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1403–1412, Berlin, Germany. Association for Computational Linguistics.

## A Additional lexemes from Panlex

| Language | #words |
|---|---|
| ace | 548 |
| aeb | 257 |
| arq | 91 |
| ary | 1702 |
| ban | 1435 |
| bbc | 857 |
| bjn | 377 |
| bug | 1001 |
| gaz | 1253 |
| min | 5755 |
| nij | 1326 |
| pcm | 58 |
| tir | 5367 |
| tso | 27 |
| twi | 13 |
| **Total** | 20067 |

Table 7: Total lexemes added from Panlex. For min we additionally extend the lexicon with a bilingual min id lexicon from Koto and Koto (2020).

## B Languages in Multilingual Sentiment Lexicon

The NRC-VAD lexicon was initially developed in English and later translated into 108 languages by the original author using the Google Translate API (Mohammad, 2018). The lexicon covers a total of 109 languages, including English, as follows:

Afrikaans (af), Albanian (sq), Amharic (am), Arabic (ar), Armenian (hy), Azerbaijani (az), Basque (eu), Belarusian (be), Bengali (bn), Bosnian (bs), Bulgarian (bg), Catalan (ca), Cebuano (ceb), Chichewa (ny), Chinese-Simplified (zh), Chinese-Traditional (zh), Corsican (co), Croatian (hr), Czech (cs), Danish (da), Dutch (nl), English (en), Esperanto (eo), Estonian (et), Filipino (fil), Finnish (fi), French (fr), Frisian (fy), Galician (gl), Georgian (ka), German (de), Greek (el), Gujarati (gu), Haitian-Creole (ht), Hausa (ha), Hawaiian (haw), Hebrew (he), Hindi (hi), Hmong (hmn), Hungarian (hu), Icelandic (is), Igbo (ig), Indonesian (id), Irish (ga), Italian (it), Japanese (ja), Javanese (jv), Kannada (kn), Kazakh (kk), Khmer (km), Kinyarwanda (rw), Korean (ko), Kurdish-Kurmanji (ku), Kyrgyz (ky), Lao (lo), Latin (la), Latvian (lv), Lithuanian (lt), Luxembourgish (lb), Macedonian (mk), Malagasy (mg), Malay (ms), Malayalam (ml), Maltese (mt), Maori (mi), Marathi (mr), Mongolian (mn), Myanmar-Burmese (my), Nepali (ne), Norwegian (no), Odia (or), Pashto (ps), Persian (fa), Polish (pl), Portuguese (pt), Punjabi (pa), Romanian (ro), Russian (ru), Samoan (sm), Scots-Gaelic (sco), Serbian (sr), Sesotho (st), Shona (sn), Sindhi (sd), Sinhala (si), Slovak (sk), Slovenian (sl), Somali (so), Spanish (es), Sundanese (su), Swahili (sw), Swedish (sv), Tajik (tg), Tamil (ta), Tatar (tt), Telugu (te), Thai (th), Turkish (tr), Turkmen (tk), Ukrainian (uk), Urdu (ur), Uyghur (ug), Uzbek (uz), Vietnamese (vi), Welsh (cy), Xhosa (xh), Yiddish (yi), Yoruba (yo), Zulu (zu)

## C Model Artifacts

| Models (#parameters) | Source |
|---|---|
| $\text{mBERT}_{\text{Base}}$ (110M) | bert-base-multilingual-cased |
| $\text{XLM-R}_{\text{Base}}$ (270M) | xlm-roberta-base |
| $\text{XLM-R}_{\text{Large}}$ (550M) | xlm-roberta-large |
| $\text{mBART}_{\text{Large}}$ (600M) | mbart-large-50 |
| $\text{mT5}_{\text{Base}}$ (580M) | google/mt5-base |
| $\text{mT5}_{\text{Large}}$ (1B) | google/mt5-large |
| XGLM (2.9B) | facebook/xglm-2.9B |
| BLOOMZ (3B) | bigscience/bloomz-3b |
| GPT–3.5 (175B) | text-davinci-003 |

Table 8: With the exception of GPT–3.5 (Ouyang et al., 2022), all models used in this study were sourced from Huggingface (Wolf et al., 2020).

## D Prompts

We use six different prompts in evaluating large language models:

- `[INPUT]`
  `What would be the sentiment of the text above? [LABELS].`
- `What is the sentiment of this text?`
  `Text: [INPUT]`
  `Sentiment: [LABELS].`
- `Text: [INPUT]`
  `Please classify the sentiment of above text: [LABELS].`
- `[INPUT]`
  `What would be the sentiment of the text above? [OPTIONS]? [LABELS].`
- `What is the sentiment of this text?`
  `Text: [INPUT]`
  `Answer with [OPTIONS]: [LABELS].`
- `Text: [INPUT]`
  `Please classify the sentiment of above text. Answer with [OPTIONS]: [LABELS].`

where [INPUT] is the input text, [OPTIONS] list all sentiment labels, and [LABELS] represent a sentiment class. For instance, given the text *I love you* in binary classification, for the first prompt, we compare two normalized log-likelihood of:

- `I love you`
  `What would be the sentiment of the text above? positive`
- `I love you`
  `What would be the sentiment of the text above? negative`

## E Detailed Experimental Results

See Table 9, Table 10, Table 11 for full results of binary classification, and Table 12, Table 13, Table 14 for full results of 3-way classification.

## F Hyperparameters and Training Configurations

For lexicon-based pretraining, we utilize a single 32GB A100 GPU. We set the initial learning rate to 2e-5 and the maximum number of epochs to 100. A patience value of 5 is used for early stopping, and a dropout rate of 0.2 is applied. Additionally, we set the maximum token length to 10. Different batch sizes are employed for each model: $\text{mBERT}_{\text{Base}}$=4000, $\text{XLM-R}_{\text{Base}}$=4000, $\text{XLM-R}_{\text{Large}}$=1000, $\text{mBART}_{\text{Large}}$=1000, $\text{mT5}_{\text{Base}}$=500, and $\text{mT5}_{\text{Large}}$=500.

For fine-tuning the mBERT model in both the full and limited training data scenarios, we also configure the initial learning rate to 2e-5 and set the maximum number of epochs to 20. A patience value of 5 is employed for early stopping, while a dropout rate of 0.2 is utilized. The maximum token length is set to 512, and a batch size of 32 is used. We also use these settings when training the English SST baseline model for five different seeds.

## G Additional Experiments

We present details of the datasets used in the additional experiments in Table 15, and present the detailed results in Table 16.

For binary experiment settings, we conduct the transformations:

- `WT-WT` (Conforti et al., 2020):: We consider the label *support* as the positive class, *refute* as the negative class, and discard *comment* and *unrelated* categories.
- `P-Stance` (Li et al., 2021): We assign the label *favor* as the positive class and *against* as the negative class.
- `HS1` (Founta et al., 2018): We map the *normal* class to the positive class and *abusive* and *hateful* classes to the negative class. We exclude the *spam* class as it is unrelated to sentiment analysis.
- `HS2` (Vidgen et al., 2021): We map the *none* label to the positive class and the remaining labels to the negative class. We acknowledge that this projection introduces noise as there is no absolute positive class available in this dataset.
- `GoEmotions` (Demszky et al., 2020): For the positive class, we include emotions such as *admiration*, *amusement*, *approval*, *caring*, *curiosity*, *desire*, *excitement*, *gratitude*, *joy*, *love*, *optimism*, *pride*, *realization*, *relief*, and *surprise*. For the negative class, we include emotions such as *annoyance*, *confusion*, *disappointment*, *disapproval*, *disgust*, *embarrassment*, *fear*, *grief*, *nervousness*, *remorse*, and *sadness*. Since `GoEmotions` is a multi-label dataset, we tally the positive and negative counts for each sentence and discard sentences with an equal count of positive and negative labels.
- `SemEval2018` (Mohammad et al., 2018): For the positive class, we include emotions such as *anticipation*, *joy*, *love*, *optimism*, *pessimism*, *surprise*, and *trust*. For the negative class, we include emotions such as *anger*, *disgust*, *fear*, and *sadness*. Similar to `GoEmotions`, since `SemEval2018` is a multi-label dataset, we employ the same strategy to determine the final class.

| Method | High/Medium-Resource | | | | | | Code-switching | | |
|---|---|---|---|---|---|---|---|---|---|
| | en | ar | es | ru | id | ja | en-es | en-ml | en-ta |
| **Full training** | | | | | | | | | |
| $\text{mBERT}_{\text{Base}}$ | 88.40 | 80.94 | 82.66 | 81.73 | 85.91 | 93.34 | 84.25 | 88.69 | 79.83 |
| + ML Lex. | 88.46 | 81.89 | 83.21 | 82.45 | 85.73 | 93.72 | 85.15 | 88.98 | 80.86 |
| + ML Lex. + Panlex | 88.69 | 82.15 | 83.59 | 84.87 | 86.06 | 94.58 | 87.77 | 89.58 | 79.73 |
| + ML Lex. + Panlex + Filtering | 88.53 | 81.46 | 83.24 | 85.95 | 85.92 | 94.07 | 86.78 | 89.28 | 81.11 |
| **Training with limited data** | | | | | | | | | |
| $\text{mBERT}_{\text{Base}}$ | 63.83 | 71.49 | 67.55 | 70.04 | 71.15 | 80.07 | 65.69 | 72.05 | 77.95 |
| + ML Lex. | 76.66 | 73.41 | 78.55 | 75.44 | 76.45 | 84.74 | 80.55 | 79.97 | 77.78 |
| + ML Lex. + Panlex | 75.97 | 75.69 | 79.69 | 75.11 | 78.21 | 83.07 | 79.92 | 77.83 | 78.03 |
| + ML Lex. + Panlex + Filtering | 76.27 | 75.19 | 79.95 | 76.91 | 77.15 | 84.39 | 78.19 | 79.34 | 77.98 |
| **Zero-shot (LLMs)** | | | | | | | | | |
| XGLM (2.9B) | 51.18 | 54.37 | 53.09 | 55.08 | 56.32 | 79.45 | 55.85 | 50.50 | 51.47 |
| BLOOMZ (3B) | 95.05 | 81.08 | 81.99 | 71.58 | 78.45 | 75.99 | 71.83 | 27.89 | 37.95 |
| GPT–3.5 (175B) | 84.52 | 72.14 | 72.53 | 80.29 | 72.68 | 89.85 | 79.83 | 68.50 | 72.66 |
| **Zero-shot (SST and Lexicon-based pretraining with regression)** | | | | | | | | | |
| $\text{mBERT}_{\text{Base}}$ (110M) | | | | | | | | | |
| + fine-tuned on SST | 88.42 | 69.92 | 74.39 | 65.64 | 67.24 | 57.17 | 61.38 | 29.87 | 42.26 |
| + ML Lex. | 71.42 | 73.06 | 76.89 | 81.84 | 72.73 | 68.35 | 69.61 | 69.58 | 69.09 |
| + ML Lex. + Panlex | 70.39 | 75.31 | 77.08 | 77.27 | 68.40 | 60.41 | 72.44 | 61.18 | 50.47 |
| + ML Lex. + Panlex + Filtering | 71.44 | 75.20 | 78.86 | 75.75 | 70.87 | 73.00 | 70.43 | 72.29 | 72.93 |
| $\text{XLM-R}_{\text{Base}}$ (270M) | | | | | | | | | |
| + fine-tuned on SST | 91.48 | 81.53 | 83.98 | 88.40 | 81.70 | 91.94 | 67.79 | 35.75 | 45.99 |
| + ML Lex. | 73.89 | 76.16 | 78.84 | 84.63 | 80.22 | 90.82 | 68.58 | 54.54 | 64.19 |
| + ML Lex. + Panlex | 74.09 | 77.80 | 79.45 | 84.23 | 81.80 | 90.43 | 69.29 | 57.17 | 65.46 |
| + ML Lex. + Panlex + Filtering | 75.85 | 77.80 | 80.43 | 83.76 | 80.92 | 91.48 | 74.31 | 71.44 | 71.76 |
| $\text{XLM-R}_{\text{Large}}$ (550M) | | | | | | | | | |
| + fine-tuned on SST | 94.00 | 85.37 | 85.95 | 92.64 | 82.88 | 95.11 | 75.36 | 38.21 | 49.57 |
| + ML Lex. | 78.14 | 79.85 | 82.12 | 88.46 | 82.47 | 89.87 | 75.41 | 70.70 | 71.47 |
| + ML Lex. + Panlex | 76.00 | 80.73 | 81.76 | 87.48 | 82.88 | 91.61 | 77.10 | 75.23 | 77.06 |
| + ML Lex. + Panlex + Filtering | 77.82 | 78.99 | 81.55 | 87.95 | 80.42 | 92.06 | 78.66 | 79.80 | 76.49 |
| $\text{mBART}_{\text{Large}}$ (600M) | | | | | | | | | |
| + fine-tuned on SST | 94.00 | 85.37 | 85.95 | 92.64 | 82.88 | 95.11 | 75.36 | 38.21 | 49.57 |
| + ML Lex. | 75.73 | 82.31 | 76.31 | 87.18 | 80.38 | 90.15 | 78.45 | 71.20 | 72.67 |
| + ML Lex. + Panlex | 78.96 | 81.18 | 79.52 | 87.07 | 81.49 | 80.59 | 74.38 | 58.79 | 61.83 |
| + ML Lex. + Panlex + Filtering | 74.38 | 58.79 | 61.83 | 74.38 | 58.79 | 61.83 | 76.40 | 73.67 | 76.23 |
| $\text{mT5}_{\text{Base}}$ (580M) | | | | | | | | | |
| + fine-tuned on SST | 90.21 | 79.99 | 82.58 | 84.00 | 78.87 | 90.37 | 70.47 | 30.49 | 44.21 |
| + ML Lex. | 72.55 | 76.48 | 79.18 | 83.58 | 75.11 | 92.11 | 77.53 | 69.86 | 73.49 |
| + ML Lex. + Panlex | 71.67 | 76.29 | 76.53 | 79.60 | 74.34 | 91.08 | 79.65 | 71.53 | 75.63 |
| + ML Lex. + Panlex + Filtering | 75.62 | 77.52 | 80.93 | 84.55 | 75.34 | 92.83 | 77.62 | 74.30 | 77.06 |
| $\text{mT5}_{\text{Large}}$ (1B) | | | | | | | | | |
| + fine-tuned on SST | 91.13 | 81.92 | 83.14 | 87.48 | 79.82 | 91.34 | 70.18 | 29.47 | 44.07 |
| + ML Lex. | 71.99 | 82.82 | 79.06 | 88.42 | 76.07 | 92.06 | 77.62 | 68.57 | 71.67 |
| + ML Lex. + Panlex | 72.37 | 80.44 | 77.34 | 86.11 | 77.97 | 92.02 | 78.66 | 74.73 | 72.56 |
| + ML Lex. + Panlex + Filtering | 70.52 | 80.26 | 75.80 | 81.44 | 78.01 | 91.22 | 76.40 | 73.67 | 76.23 |

Table 9: All **binary classification** results for high/medium-resource languages and code-switched texts.

| Method | NusaX | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | **ace** | **ban** | **bbc** | **bjn** | **bug** | **jav** | **mad** | **min** | **nij** | **sun** |
| **Full training** | | | | | | | | | | |
| $\text{mBERT}_{\text{Base}}$ | 83.84 | 85.44 | 84.42 | 84.99 | 83.85 | 87.82 | 84.31 | 87.87 | 83.93 | 85.90 |
| + ML Lex. | 84.96 | 85.97 | 84.16 | 89.44 | 84.19 | 92.04 | 86.04 | 88.28 | 84.84 | 89.14 |
| + ML Lex. + Panlex | 85.32 | 87.26 | 85.58 | 88.65 | 83.64 | 92.56 | 85.51 | 88.88 | 86.82 | 89.79 |
| + ML Lex. + Panlex + Filtering | 84.33 | 87.22 | 84.52 | 89.83 | 83.73 | 93.35 | 85.17 | 89.33 | 87.66 | 90.06 |
| **Training with limited data** | | | | | | | | | | |
| $\text{mBERT}_{\text{Base}}$ | 70.64 | 71.89 | 69.77 | 77.49 | 68.61 | 81.58 | 72.01 | 79.57 | 74.03 | 79.04 |
| + ML Lex. | 79.72 | 76.15 | 76.51 | 79.20 | 76.14 | 89.65 | 78.13 | 85.73 | 81.67 | 87.95 |
| + ML Lex. + Panlex | 80.70 | 78.22 | 76.84 | 83.21 | 75.60 | 91.36 | 80.86 | 85.31 | 81.00 | 86.55 |
| + ML Lex. + Panlex + Filtering | 78.85 | 77.13 | 76.57 | 83.70 | 75.57 | 90.37 | 79.22 | 87.27 | 82.88 | 85.84 |
| **Zero-shot (LLMs)** | | | | | | | | | | |
| XGLM (2.9B) | 48.19 | 53.01 | 41.04 | 53.20 | 39.45 | 55.61 | 51.38 | 53.45 | 46.49 | 51.55 |
| BLOOMZ (3B) | 74.10 | 74.05 | 55.56 | 83.20 | 49.92 | 81.35 | 67.08 | 78.66 | 68.97 | 65.56 |
| GPT–3.5 (175B) | 63.52 | 63.53 | 55.37 | 72.84 | 46.19 | 75.43 | 61.01 | 72.49 | 60.28 | 68.32 |
| **Zero-shot (SST and Lexicon-based pretraining with regression)** | | | | | | | | | | |
| $\text{mBERT}_{\text{Base}}$ (110M) | | | | | | | | | | |
| + fine-tuned on SST | 38.05 | 47.89 | 39.04 | 46.86 | 34.94 | 54.90 | 40.48 | 51.26 | 44.80 | 51.37 |
| + ML Lex. | 66.00 | 67.58 | 62.38 | 69.22 | 58.68 | 80.23 | 65.66 | 70.38 | 61.23 | 77.88 |
| + ML Lex. + Panlex | 64.06 | 62.30 | 59.23 | 66.60 | 53.98 | 80.58 | 68.63 | 72.50 | 62.48 | 76.73 |
| + ML Lex. + Panlex + Filtering | 60.07 | 58.00 | 58.64 | 63.41 | 48.48 | 80.22 | 62.71 | 72.00 | 61.62 | 74.31 |
| $\text{XLM-R}_{\text{Base}}$ (270M) | | | | | | | | | | |
| + fine-tuned on SST | 56.63 | 62.59 | 38.17 | 74.12 | 35.34 | 80.86 | 59.55 | 65.79 | 58.28 | 63.71 |
| + ML Lex. | 68.56 | 80.23 | 49.21 | 83.21 | 44.25 | 90.79 | 66.42 | 81.20 | 64.39 | 81.17 |
| + ML Lex. + Panlex | 66.57 | 80.77 | 57.25 | 83.87 | 50.90 | 89.80 | 69.23 | 84.21 | 71.13 | 82.17 |
| + ML Lex. + Panlex + Filtering | 62.47 | 77.40 | 64.54 | 79.22 | 59.19 | 86.18 | 72.37 | 83.19 | 70.57 | 79.59 |
| $\text{XLM-R}_{\text{Large}}$ (550M) | | | | | | | | | | |
| + fine-tuned on SST | 68.43 | 78.26 | 49.79 | 86.87 | 44.73 | 91.82 | 75.23 | 80.73 | 69.62 | 84.48 |
| + ML Lex. | 73.19 | 77.61 | 66.00 | 82.35 | 63.41 | 92.43 | 74.31 | 84.81 | 74.88 | 91.11 |
| + ML Lex. + Panlex | 60.81 | 64.60 | 59.76 | 77.01 | 55.83 | 89.40 | 69.28 | 81.87 | 64.28 | 85.01 |
| + ML Lex. + Panlex + Filtering | 52.42 | 65.90 | 53.43 | 72.08 | 40.84 | 87.71 | 65.75 | 78.81 | 65.27 | 85.33 |
| $\text{mBART}_{\text{Large}}$ (600M) | | | | | | | | | | |
| + fine-tuned on SST | 62.93 | 70.97 | 46.23 | 75.61 | 46.88 | 81.68 | 71.92 | 72.80 | 63.00 | 62.09 |
| + ML Lex. | 67.10 | 66.14 | 65.31 | 80.49 | 56.64 | 82.35 | 75.17 | 81.75 | 71.23 | 76.36 |
| + ML Lex. + Panlex | 70.26 | 76.38 | 67.40 | 81.15 | 62.77 | 82.89 | 70.94 | 84.54 | 71.36 | 80.90 |
| + ML Lex. + Panlex + Filtering | 50.12 | 61.09 | 48.79 | 70.58 | 45.53 | 73.10 | 60.91 | 75.94 | 64.45 | 66.59 |
| $\text{mT5}_{\text{Base}}$ (580M) | | | | | | | | | | |
| + fine-tuned on SST | 47.86 | 61.23 | 36.09 | 65.18 | 35.36 | 82.62 | 51.24 | 53.58 | 52.54 | 67.63 |
| + ML Lex. | 74.66 | 74.59 | 64.85 | 77.96 | 64.62 | 87.06 | 73.96 | 80.90 | 75.30 | 84.46 |
| + ML Lex. + Panlex | 70.89 | 72.58 | 56.52 | 76.29 | 57.11 | 84.75 | 66.78 | 76.97 | 73.00 | 78.78 |
| + ML Lex. + Panlex + Filtering | 71.26 | 73.04 | 63.75 | 80.03 | 63.09 | 89.10 | 75.30 | 80.26 | 73.88 | 85.50 |
| $\text{mT5}_{\text{Large}}$ (1B) | | | | | | | | | | |
| + fine-tuned on SST | 57.80 | 69.19 | 39.04 | 68.32 | 36.78 | 88.56 | 54.81 | 61.52 | 55.26 | 75.54 |
| + ML Lex. | 75.37 | 76.10 | 66.95 | 82.67 | 66.88 | 90.41 | 75.75 | 85.49 | 74.22 | 88.75 |
| + ML Lex. + Panlex | 74.90 | 76.62 | 61.84 | 81.91 | 63.84 | 89.73 | 73.88 | 80.38 | 75.54 | 88.39 |
| + ML Lex. + Panlex + Filtering | 68.17 | 75.17 | 62.05 | 79.53 | 63.47 | 83.49 | 67.75 | 78.90 | 71.26 | 83.86 |

Table 10: All **binary classification** results for NusaX.

| Method | African | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **am** | **dz** | **ha** | **ig** | **kr** | **ma** | **pcm** | **pt** | **sw** | **ts** | **twi** | **yo** | **or** | **tg** | **aeb** |
| **Full training** | | | | | | | | | | | | | | | |
| $\text{mBERT}_{\text{Base}}$ | 67.44 | 70.15 | 88.09 | 91.23 | 76.25 | 67.91 | 75.10 | 77.52 | 73.39 | 63.60 | 78.48 | 83.17 | 54.96 | 49.18 | 75.37 |
| + ML Lex. | 65.50 | 71.95 | 88.89 | 91.74 | 79.96 | 69.86 | 75.83 | 78.70 | 77.37 | 64.71 | 77.67 | 84.11 | 51.76 | 49.12 | 75.93 |
| + ML Lex. + Panlex | 70.13 | 72.07 | 88.91 | 91.94 | 78.03 | 69.20 | 75.89 | 79.51 | 76.09 | 64.63 | 77.64 | 84.29 | 52.01 | 50.01 | 77.34 |
| + ML Lex. + Panlex + Filtering | 68.37 | 72.63 | 89.14 | 91.96 | 77.65 | 67.95 | 76.60 | 79.19 | 75.65 | 64.30 | 79.04 | 83.67 | 55.33 | 49.36 | 75.88 |
| **Training with limited data** | | | | | | | | | | | | | | | |
| $\text{mBERT}_{\text{Base}}$ | 45.84 | 62.02 | 71.98 | 61.94 | 62.18 | 52.72 | 56.23 | 66.55 | 66.28 | 56.83 | 53.55 | 65.36 | 42.98 | 45.98 | 63.91 |
| + ML Lex. | 48.62 | 65.39 | 75.71 | 71.50 | 71.70 | 58.40 | 61.62 | 75.06 | 70.79 | 55.63 | 54.97 | 67.81 | 47.21 | 45.45 | 62.11 |
| + ML Lex. + Panlex | 50.92 | 64.31 | 74.09 | 70.68 | 70.96 | 58.19 | 62.81 | 75.67 | 73.27 | 54.86 | 56.71 | 67.91 | 52.80 | 45.67 | 64.08 |
| + ML Lex. + Panlex + Filtering | 53.07 | 65.11 | 75.05 | 71.41 | 69.94 | 57.91 | 63.26 | 76.17 | 70.74 | 55.26 | 56.39 | 68.23 | 53.60 | 45.24 | 63.73 |
| **Zero-shot (LLMs)** | | | | | | | | | | | | | | | |
| XGLM (2.9B) | 25.28 | 30.34 | 45.48 | 41.24 | 45.14 | 52.24 | 50.75 | 47.25 | 56.46 | 43.94 | 43.08 | 39.23 | 38.70 | 33.57 | 44.78 |
| BLOOMZ (3B) | 62.98 | 67.26 | 51.12 | 52.65 | 64.79 | 57.67 | 62.69 | 80.72 | 45.90 | 49.09 | 44.90 | 38.46 | 49.45 | 48.34 | 47.76 |
| GPT–3.5 (174B) | 41.45 | 54.90 | 57.76 | 51.12 | 50.35 | 59.39 | 63.08 | 70.34 | 70.89 | 54.10 | 51.46 | 50.69 | 40.06 | 39.76 | 52.02 |
| **Zero-shot (SST and Lexicon-based pretraining with regression)** | | | | | | | | | | | | | | | |
| $\text{mBERT}_{\text{Base}}$ (110M) | | | | | | | | | | | | | | | |
| + fine-tuned on SST | 31.75 | 59.08 | 53.21 | 47.76 | 46.52 | 55.87 | 61.74 | 65.13 | 32.51 | 37.70 | 40.13 | 31.84 | 49.33 | 37.10 | 48.71 |
| + ML Lex. | 51.10 | 61.20 | 61.06 | 67.50 | 62.64 | 56.61 | 65.10 | 69.06 | 72.03 | 52.79 | 45.46 | 58.49 | 48.53 | 47.10 | 48.24 |
| + ML Lex. + Panlex | 52.56 | 61.67 | 60.16 | 69.96 | 65.34 | 60.17 | 66.72 | 72.27 | 68.01 | 46.52 | 50.21 | 60.31 | 49.92 | 48.75 | 52.36 |
| + ML Lex. + Panlex + Filtering | 50.66 | 61.24 | 53.87 | 68.89 | 58.36 | 58.04 | 67.41 | 72.45 | 68.46 | 53.33 | 51.20 | 62.08 | 47.30 | 48.33 | 48.45 |
| $\text{XLM-R}_{\text{Base}}$ (270M) | | | | | | | | | | | | | | | |
| + fine-tuned on SST | 80.76 | 72.71 | 65.89 | 45.38 | 48.62 | 60.71 | 65.70 | 83.67 | 59.49 | 39.69 | 38.23 | 27.30 | 48.04 | 60.13 | 52.57 |
| + ML Lex. | 81.62 | 68.67 | 60.54 | 64.43 | 59.22 | 59.36 | 64.97 | 80.31 | 70.30 | 56.90 | 51.12 | 34.80 | 53.18 | 62.36 | 54.03 |
| + ML Lex. + Panlex | 81.17 | 70.19 | 60.81 | 65.92 | 63.80 | 61.00 | 66.05 | 79.58 | 72.92 | 60.03 | 52.09 | 40.16 | 54.50 | 65.09 | 53.13 |
| + ML Lex. + Panlex + Filtering | 81.14 | 70.54 | 64.42 | 65.06 | 60.92 | 61.08 | 67.56 | 80.31 | 75.35 | 58.16 | 51.96 | 57.85 | 54.33 | 62.04 | 49.19 |
| $\text{XLM-R}_{\text{Large}}$ (550M) | | | | | | | | | | | | | | | |
| + fine-tuned on SST | 82.92 | 75.06 | 70.90 | 52.31 | 56.65 | 64.66 | 66.81 | 85.71 | 69.60 | 46.03 | 44.20 | 36.17 | 55.45 | 62.47 | 57.34 |
| + ML Lex. | 81.52 | 69.43 | 71.28 | 70.63 | 62.06 | 64.88 | 67.72 | 80.68 | 77.75 | 61.23 | 53.34 | 53.72 | 53.95 | 67.05 | 57.19 |
| + ML Lex. + Panlex | 78.26 | 69.28 | 68.08 | 67.55 | 52.02 | 60.45 | 68.38 | 83.45 | 80.70 | 63.39 | 56.32 | 60.13 | 46.16 | 46.27 | 57.38 |
| + ML Lex. + Panlex + Filtering | 78.84 | 67.74 | 67.81 | 63.83 | 46.45 | 61.22 | 69.81 | 84.49 | 83.26 | 55.39 | 51.16 | 59.35 | 47.43 | 47.74 | 49.62 |
| $\text{mBART}_{\text{Large}}$ (600M) | | | | | | | | | | | | | | | |
| + fine-tuned on SST | 65.74 | 71.64 | 66.99 | 52.74 | 54.24 | 62.94 | 67.80 | 82.91 | 61.32 | 54.90 | 50.94 | 37.19 | 53.62 | 48.43 | 62.95 |
| + ML Lex. | 63.40 | 69.53 | 64.93 | 65.09 | 62.53 | 61.31 | 70.38 | 83.21 | 82.08 | 54.91 | 56.44 | 56.14 | 48.10 | 47.62 | 57.61 |
| + ML Lex. + Panlex | 63.23 | 69.43 | 67.95 | 70.72 | 65.20 | 60.86 | 68.68 | 83.48 | 77.31 | 52.93 | 50.99 | 53.53 | 52.49 | 47.25 | 57.14 |
| + ML Lex. + Panlex + Filtering | 58.19 | 62.84 | 50.86 | 59.26 | 49.83 | 58.42 | 65.55 | 82.66 | 83.68 | 58.04 | 53.25 | 59.62 | 38.75 | 46.71 | 46.58 |
| $\text{mT5}_{\text{Base}}$ (580M) | | | | | | | | | | | | | | | |
| + fine-tuned on SST | 79.97 | 67.21 | 68.15 | 54.55 | 66.64 | 59.14 | 65.70 | 79.85 | 56.31 | 36.35 | 34.85 | 30.46 | 48.40 | 59.71 | 50.38 |
| + ML Lex. | 78.20 | 68.65 | 78.16 | 67.04 | 75.52 | 67.93 | 69.81 | 80.59 | 80.32 | 54.40 | 51.52 | 49.12 | 56.30 | 72.56 | 61.63 |
| + ML Lex. + Panlex | 78.62 | 69.08 | 76.56 | 69.00 | 74.35 | 68.08 | 68.53 | 79.27 | 77.16 | 59.46 | 54.96 | 42.08 | 56.20 | 72.07 | 56.71 |
| + ML Lex. + Panlex + Filtering | 77.31 | 68.49 | 75.92 | 67.16 | 75.97 | 67.01 | 70.45 | 82.65 | 81.84 | 54.67 | 54.64 | 52.40 | 53.69 | 70.58 | 60.08 |
| $\text{mT5}_{\text{Large}}$ (1B) | | | | | | | | | | | | | | | |
| + fine-tuned on SST | 81.45 | 69.14 | 66.54 | 54.94 | 73.30 | 59.49 | 65.43 | 81.83 | 59.14 | 39.07 | 36.09 | 32.45 | 49.32 | 64.23 | 47.66 |
| + ML Lex. | 83.09 | 73.54 | 77.57 | 70.64 | 79.14 | 67.58 | 69.18 | 82.78 | 80.85 | 58.42 | 55.06 | 48.31 | 56.27 | 74.45 | 62.26 |
| + ML Lex. + Panlex | 81.20 | 70.21 | 79.67 | 72.68 | 75.91 | 67.07 | 70.03 | 82.14 | 82.74 | 59.93 | 59.97 | 51.87 | 57.53 | 78.40 | 61.33 |
| + ML Lex. + Panlex + Filtering | 79.09 | 70.35 | 75.20 | 69.27 | 71.35 | 66.87 | 68.95 | 80.86 | 78.85 | 59.81 | 56.89 | 58.52 | 56.20 | 71.62 | 56.74 |

Table 11: All **binary classification** results for the 14 African languages from SemEval 2023.

| Method | High/Medium-Resource | | | | | | Code-switching | | |
|---|---|---|---|---|---|---|---|---|---|
| | en | ar | es | ru | id | ja | en-es | en-ml | en-ta |
| **Full training** | | | | | | | | | |
| $\text{mBERT}_{\text{Base}}$ | 72.65 | 64.15 | 62.21 | 80.05 | 85.91 | 83.04 | 63.58 | 88.69 | 79.83 |
| + ML Lex. | 72.78 | 64.68 | 64.11 | 81.10 | 85.73 | 83.12 | 65.42 | 88.98 | 80.86 |
| + ML Lex. + Panlex | 72.44 | 65.85 | 64.14 | 81.97 | 86.06 | 82.54 | 64.34 | 89.58 | 79.73 |
| + ML Lex. + Panlex + Filtering | 73.10 | 64.78 | 63.57 | 81.70 | 85.92 | 82.67 | 64.91 | 89.28 | 81.11 |
| **Training with limited data** | | | | | | | | | |
| $\text{mBERT}_{\text{Base}}$ | 49.49 | 49.54 | 43.58 | 73.32 | 71.15 | 60.96 | 43.48 | 72.05 | 77.95 |
| + ML Lex. | 59.55 | 50.05 | 53.79 | 77.52 | 76.45 | 66.20 | 56.05 | 79.97 | 77.78 |
| + ML Lex. + Panlex | 59.32 | 51.96 | 55.65 | 78.20 | 78.21 | 66.51 | 54.40 | 77.83 | 78.03 |
| + ML Lex. + Panlex + Filtering | 60.03 | 52.15 | 57.32 | 78.35 | 77.15 | 68.46 | 51.96 | 79.34 | 77.98 |
| **Zero-shot (LLMs)** | | | | | | | | | |
| XGLM (2.9B) | 39.88 | 29.44 | 37.37 | 16.52 | 54.86 | 52.25 | 31.89 | 57.58 | 60.78 |
| BLOOMZ (3B) | 71.12 | 48.66 | 52.04 | 6.93 | 80.57 | 53.95 | 34.20 | 31.12 | 42.23 |
| GPT–3.5 (175B) | 67.61 | 55.56 | 60.56 | 83.35 | 66.58 | 72.22 | 58.97 | 42.85 | 49.42 |
| **Zero-shot (SST and Lexicon-based pretraining with classification)** | | | | | | | | | |
| $\text{mBERT}_{\text{Base}}$ (110M) | | | | | | | | | |
| + fine-tuned on SST | 70.13 | 43.44 | 52.00 | 42.54 | 57.21 | 39.26 | 40.82 | 12.98 | 20.73 |
| + ML Lex. | 53.06 | 49.53 | 50.80 | 69.44 | 54.21 | 54.62 | 52.22 | 65.38 | 62.81 |
| + ML Lex. + Panlex | 53.76 | 49.00 | 48.64 | 67.20 | 55.72 | 56.54 | 53.06 | 67.10 | 63.65 |
| + ML Lex. + Panlex + Filtering | 49.90 | 49.48 | 50.83 | 64.17 | 50.87 | 55.68 | 55.21 | 60.96 | 60.27 |
| $\text{XLM-R}_{\text{Base}}$ (270M) | | | | | | | | | |
| + fine-tuned on SST | 73.54 | 56.51 | 62.46 | 77.20 | 69.01 | 76.18 | 50.32 | 25.64 | 32.94 |
| + ML Lex. | 50.79 | 52.38 | 49.41 | 72.99 | 58.75 | 69.05 | 47.23 | 31.06 | 38.23 |
| + ML Lex. + Panlex | 48.79 | 52.82 | 48.63 | 72.38 | 58.96 | 68.30 | 47.89 | 33.24 | 39.26 |
| + ML Lex. + Panlex + Filtering | 46.10 | 54.11 | 50.20 | 73.62 | 58.87 | 69.62 | 51.26 | 32.34 | 36.79 |
| $\text{XLM-R}_{\text{Large}}$ (550M) | | | | | | | | | |
| + fine-tuned on SST | 76.07 | 61.67 | 63.52 | 83.19 | 64.17 | 80.30 | 54.75 | 23.79 | 31.59 |
| + ML Lex. | 52.82 | 56.18 | 53.57 | 69.47 | 64.13 | 65.55 | 54.93 | 66.14 | 69.50 |
| + ML Lex. + Panlex | 53.07 | 57.94 | 56.12 | 73.80 | 70.06 | 63.99 | 56.75 | 69.11 | 68.34 |
| + ML Lex. + Panlex + Filtering | 57.25 | 58.00 | 57.57 | 69.45 | 74.67 | 64.01 | 59.01 | 58.74 | 61.03 |
| $\text{mBART}_{\text{Large}}$ (600M) | | | | | | | | | |
| + fine-tuned on SST | 74.18 | 56.87 | 61.34 | 82.02 | 56.44 | 76.21 | 46.81 | 20.15 | 27.61 |
| + ML Lex. | 51.98 | 57.05 | 51.20 | 67.36 | 59.81 | 70.84 | 53.70 | 48.22 | 61.36 |
| + ML Lex. + Panlex | 49.09 | 56.82 | 54.20 | 69.91 | 54.94 | 69.95 | 51.58 | 44.43 | 52.51 |
| + ML Lex. + Panlex + Filtering | 47.30 | 53.74 | 49.83 | 71.56 | 45.16 | 67.12 | 48.30 | 31.99 | 42.00 |
| $\text{mT5}_{\text{Base}}$ (580M) | | | | | | | | | |
| + fine-tuned on SST | 71.14 | 52.12 | 58.48 | 61.45 | 68.82 | 70.98 | 46.00 | 18.54 | 28.58 |
| + ML Lex. | 47.54 | 51.64 | 48.11 | 60.88 | 69.66 | 66.07 | 49.57 | 62.13 | 68.41 |
| + ML Lex. + Panlex | 47.50 | 49.59 | 48.96 | 58.89 | 62.64 | 65.62 | 48.18 | 60.45 | 66.90 |
| + ML Lex. + Panlex + Filtering | 51.85 | 50.53 | 52.23 | 70.95 | 68.98 | 65.93 | 52.56 | 59.94 | 64.76 |
| $\text{mT5}_{\text{Large}}$ (1B) | | | | | | | | | |
| + fine-tuned on SST | 65.71 | 37.38 | 43.89 | 55.95 | 51.91 | 51.10 | 36.49 | 14.22 | 23.55 |
| + ML Lex. | 51.61 | 54.48 | 53.08 | 70.37 | 65.08 | 67.75 | 51.95 | 65.78 | 66.55 |
| + ML Lex. + Panlex | 50.24 | 52.71 | 49.18 | 68.20 | 62.20 | 65.68 | 47.55 | 67.86 | 70.75 |
| + ML Lex. + Panlex + Filtering | 48.71 | 54.31 | 49.56 | 70.24 | 56.60 | 66.88 | 54.53 | 58.84 | 68.24 |

Table 12: All **3-way classification** results for high/medium-resource languages and code-switched texts.

| Method | NusaX | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | **ace** | **ban** | **bbc** | **bjn** | **bug** | **jav** | **mad** | **min** | **nij** | **sun** |
| **Full training** | | | | | | | | | | |
| $\text{mBERT}_{\text{Base}}$ | 75.67 | 76.09 | 71.71 | 75.94 | 74.64 | 78.37 | 72.88 | 76.60 | 73.40 | 76.61 |
| + ML Lex. | 76.74 | 75.69 | 75.08 | 78.54 | 75.41 | 81.07 | 72.39 | 81.21 | 74.05 | 79.40 |
| + ML Lex. + Panlex | 77.31 | 76.55 | 75.01 | 78.12 | 76.08 | 82.09 | 74.90 | 79.35 | 74.84 | 79.35 |
| + ML Lex. + Panlex + Filtering | 76.91 | 76.41 | 74.50 | 80.17 | 74.79 | 82.34 | 75.01 | 78.43 | 75.63 | 79.91 |
| **Training with limited data** | | | | | | | | | | |
| $\text{mBERT}_{\text{Base}}$ | 60.66 | 62.71 | 60.40 | 66.41 | 58.91 | 68.77 | 60.87 | 63.22 | 62.09 | 69.10 |
| + ML Lex. | 63.21 | 66.87 | 63.02 | 64.98 | 64.08 | 76.34 | 61.16 | 69.56 | 63.56 | 73.84 |
| + ML Lex. + Panlex | 64.86 | 67.43 | 65.86 | 69.42 | 65.02 | 77.92 | 64.16 | 70.73 | 65.30 | 72.20 |
| + ML Lex. + Panlex + Filtering | 65.51 | 66.93 | 65.17 | 68.82 | 64.26 | 78.03 | 63.11 | 70.27 | 66.74 | 72.67 |
| **Zero-shot (LLMs)** | | | | | | | | | | |
| XGLM (2.9B) | 32.90 | 35.42 | 28.20 | 35.14 | 27.02 | 37.60 | 34.54 | 36.28 | 31.79 | 35.77 |
| BLOOMZ (3B) | 51.91 | 51.68 | 39.36 | 57.28 | 34.49 | 56.56 | 47.45 | 54.08 | 48.91 | 47.15 |
| GPT–3.5 (175B) | 49.19 | 48.96 | 33.09 | 59.82 | 26.38 | 63.74 | 45.65 | 59.10 | 44.42 | 54.63 |
| **Zero-shot (SST and Lexicon-based pretraining with classification)** | | | | | | | | | | |
| $\text{mBERT}_{\text{Base}}$ (110M) | | | | | | | | | | |
| + fine-tuned on SST | 24.89 | 30.64 | 23.33 | 30.83 | 23.80 | 34.03 | 27.41 | 33.43 | 28.66 | 32.44 |
| + ML Lex. | 35.04 | 43.33 | 36.73 | 43.62 | 36.73 | 60.91 | 45.16 | 45.25 | 42.75 | 52.25 |
| + ML Lex. + Panlex | 36.91 | 41.95 | 39.15 | 42.40 | 37.21 | 57.96 | 42.18 | 44.58 | 41.04 | 51.29 |
| + ML Lex. + Panlex + Filtering | 35.95 | 38.95 | 30.83 | 40.70 | 28.02 | 54.72 | 37.23 | 42.34 | 37.52 | 45.51 |
| $\text{XLM-R}_{\text{Base}}$ (270M) | | | | | | | | | | |
| + fine-tuned on SST | 34.36 | 40.39 | 27.88 | 50.40 | 25.19 | 62.25 | 37.43 | 48.13 | 41.43 | 45.96 |
| + ML Lex. | 29.57 | 35.27 | 13.99 | 43.52 | 13.58 | 52.47 | 28.47 | 47.01 | 25.54 | 49.11 |
| + ML Lex. + Panlex | 29.78 | 35.42 | 15.25 | 46.90 | 14.94 | 50.94 | 28.83 | 45.31 | 28.23 | 47.82 |
| + ML Lex. + Panlex + Filtering | 32.11 | 33.78 | 13.26 | 44.18 | 10.57 | 52.70 | 32.69 | 45.18 | 22.25 | 46.56 |
| $\text{XLM-R}_{\text{Large}}$ (550M) | | | | | | | | | | |
| + fine-tuned on SST | 42.83 | 50.93 | 25.56 | 64.92 | 22.00 | 76.62 | 44.20 | 61.42 | 42.41 | 65.42 |
| + ML Lex. | 43.37 | 43.84 | 30.07 | 51.68 | 32.12 | 59.79 | 44.57 | 55.74 | 42.20 | 56.11 |
| + ML Lex. + Panlex | 48.76 | 51.25 | 43.75 | 56.95 | 37.66 | 62.81 | 48.42 | 60.28 | 52.68 | 65.80 |
| + ML Lex. + Panlex + Filtering | 42.39 | 47.69 | 31.67 | 49.93 | 26.87 | 61.18 | 47.59 | 56.83 | 43.68 | 58.89 |
| $\text{mBART}_{\text{Large}}$ (600M) | | | | | | | | | | |
| + fine-tuned on SST | 25.81 | 33.64 | 17.25 | 46.62 | 13.31 | 46.02 | 30.52 | 44.44 | 32.60 | 27.76 |
| + ML Lex. | 36.09 | 41.12 | 31.95 | 52.44 | 27.32 | 50.21 | 43.97 | 48.85 | 40.15 | 46.70 |
| + ML Lex. + Panlex | 31.24 | 38.97 | 26.54 | 44.26 | 22.03 | 43.78 | 34.47 | 42.80 | 33.77 | 39.76 |
| + ML Lex. + Panlex + Filtering | 28.73 | 34.82 | 17.64 | 38.36 | 15.81 | 40.94 | 27.59 | 38.22 | 27.60 | 35.31 |
| $\text{mT5}_{\text{Base}}$ (580M) | | | | | | | | | | |
| + fine-tuned on SST | 29.77 | 37.00 | 25.70 | 35.83 | 27.29 | 55.31 | 31.09 | 33.12 | 34.25 | 46.41 |
| + ML Lex. | 48.47 | 48.56 | 44.55 | 55.73 | 45.58 | 61.24 | 48.94 | 55.49 | 52.53 | 55.26 |
| + ML Lex. + Panlex | 49.56 | 49.90 | 38.50 | 56.42 | 42.40 | 62.17 | 44.03 | 53.39 | 49.45 | 56.43 |
| + ML Lex. + Panlex + Filtering | 45.54 | 41.50 | 31.16 | 53.93 | 29.57 | 62.00 | 41.52 | 53.02 | 45.69 | 55.26 |
| $\text{mT5}_{\text{Large}}$ (1B) | | | | | | | | | | |
| + fine-tuned on SST | 29.32 | 32.70 | 28.89 | 32.90 | 28.10 | 37.39 | 29.24 | 32.32 | 30.95 | 34.90 |
| + ML Lex. | 48.28 | 48.30 | 41.65 | 58.03 | 40.26 | 62.78 | 53.14 | 55.63 | 52.47 | 57.88 |
| + ML Lex. + Panlex | 54.32 | 52.35 | 43.84 | 59.30 | 47.76 | 61.94 | 49.85 | 58.63 | 54.95 | 56.65 |
| + ML Lex. + Panlex + Filtering | 41.64 | 48.12 | 33.72 | 51.50 | 34.45 | 59.36 | 42.19 | 54.99 | 44.15 | 53.62 |

Table 13: All **3-way classification** results for NusaX.

| Method | African | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **am** | **dz** | **ha** | **ig** | **kr** | **ma** | **pcm** | **pt** | **sw** | **ts** | **twi** | **yo** | **or** | **tg** | **aeb** |
| **Full training** | | | | | | | | | | | | | | | |
| $\text{mBERT}_{\text{Base}}$ | 17.51 | 58.38 | 75.27 | 77.73 | 56.41 | 48.21 | 62.93 | 63.52 | 55.04 | 49.61 | 65.01 | 71.30 | 35.38 | 36.89 | 71.59 |
| + ML Lex. | 10.10 | 59.99 | 75.98 | 78.93 | 59.23 | 48.37 | 63.66 | 65.08 | 57.18 | 51.66 | 64.59 | 71.77 | 32.82 | 38.22 | 71.32 |
| + ML Lex. + Panlex | 11.48 | 59.75 | 75.75 | 78.72 | 59.29 | 47.93 | 64.36 | 65.13 | 56.15 | 49.19 | 64.16 | 72.49 | 33.58 | 38.25 | 71.80 |
| + ML Lex. + Panlex + Filtering | 12.14 | 59.04 | 75.70 | 79.29 | 58.20 | 47.27 | 64.65 | 64.86 | 56.36 | 49.89 | 65.44 | 71.82 | 34.67 | 39.27 | 73.16 |
| **Training with limited data** | | | | | | | | | | | | | | | |
| $\text{mBERT}_{\text{Base}}$ | 5.42 | 49.46 | 53.66 | 39.71 | 40.48 | 31.72 | 50.61 | 53.60 | 47.82 | 39.86 | 42.32 | 42.56 | 29.36 | 32.06 | 59.45 |
| + ML Lex. | 5.99 | 52.00 | 54.24 | 53.02 | 40.34 | 38.82 | 55.19 | 59.76 | 47.38 | 40.34 | 41.82 | 42.25 | 30.48 | 35.70 | 60.20 |
| + ML Lex. + Panlex | 6.20 | 51.63 | 50.19 | 53.87 | 39.54 | 36.49 | 55.30 | 60.46 | 51.81 | 41.17 | 44.31 | 43.53 | 31.05 | 33.47 | 57.72 |
| + ML Lex. + Panlex + Filtering | 5.27 | 52.00 | 50.53 | 53.84 | 41.15 | 38.41 | 56.11 | 61.02 | 50.06 | 42.80 | 43.12 | 42.97 | 30.86 | 35.55 | 59.43 |
| **Zero-shot (LLMs)** | | | | | | | | | | | | | | | |
| XGLM (2.9B) | 16.23 | 22.74 | 26.25 | 18.00 | 24.12 | 28.74 | 43.38 | 17.63 | 15.25 | 35.90 | 35.80 | 22.19 | 17.03 | 20.51 | 41.98 |
| BLOOMZ (3B) | 52.47 | 53.82 | 27.95 | 23.05 | 32.61 | 34.12 | 54.69 | 16.07 | 13.67 | 40.01 | 36.76 | 21.13 | 20.70 | 34.58 | 45.51 |
| GPT–3.5 (175B) | 22.03 | 40.45 | 45.05 | 40.29 | 37.67 | 41.62 | 50.57 | 59.19 | 51.48 | 33.49 | 29.09 | 33.52 | 32.82 | 22.29 | 32.42 |
| **Zero-shot (SST and Lexicon-based pretraining with classification)** | | | | | | | | | | | | | | | |
| $\text{mBERT}_{\text{Base}}$ (110M) | | | | | | | | | | | | | | | |
| + fine-tuned on SST | 16.16 | 32.77 | 25.12 | 31.83 | 28.94 | 33.93 | 38.70 | 44.91 | 39.26 | 15.71 | 11.76 | 22.95 | 32.94 | 17.67 | 24.35 |
| + ML Lex. | 9.39 | 45.13 | 36.55 | 38.75 | 38.72 | 40.17 | 46.68 | 45.93 | 34.88 | 31.97 | 35.11 | 40.21 | 31.49 | 15.24 | 35.97 |
| + ML Lex. + Panlex | 7.58 | 40.85 | 36.55 | 37.76 | 40.61 | 39.75 | 45.65 | 47.51 | 35.26 | 36.60 | 34.69 | 42.31 | 32.71 | 12.94 | 36.14 |
| + ML Lex. + Panlex + Filtering | 5.64 | 39.13 | 39.13 | 44.50 | 34.07 | 38.94 | 42.71 | 52.01 | 40.50 | 24.53 | 32.75 | 37.72 | 29.35 | 9.81 | 30.54 |
| $\text{XLM-R}_{\text{Base}}$ (270M) | | | | | | | | | | | | | | | |
| + fine-tuned on SST | 61.62 | 43.28 | 33.82 | 33.92 | 33.74 | 36.17 | 47.65 | 51.17 | 44.29 | 17.91 | 20.39 | 23.65 | 31.96 | 34.30 | 23.51 |
| + ML Lex. | 48.15 | 37.74 | 43.92 | 41.78 | 35.17 | 36.31 | 31.30 | 60.23 | 55.02 | 33.12 | 19.57 | 24.09 | 36.74 | 20.48 | 21.03 |
| + ML Lex. + Panlex | 48.60 | 40.92 | 43.16 | 43.11 | 36.16 | 37.12 | 31.05 | 59.20 | 53.36 | 33.74 | 18.96 | 24.41 | 36.35 | 24.39 | 18.28 |
| + ML Lex. + Panlex + Filtering | 41.84 | 35.22 | 39.44 | 42.71 | 29.06 | 35.57 | 33.69 | 62.43 | 53.13 | 24.66 | 22.09 | 24.27 | 35.09 | 19.21 | 24.68 |
| $\text{XLM-R}_{\text{Large}}$ (550M) | | | | | | | | | | | | | | | |
| + fine-tuned on SST | 60.78 | 49.27 | 41.82 | 33.61 | 32.37 | 42.48 | 45.67 | 55.20 | 50.55 | 20.05 | 18.49 | 23.05 | 35.20 | 27.51 | 24.76 |
| + ML Lex. | 52.71 | 42.98 | 50.26 | 46.10 | 38.54 | 41.44 | 49.93 | 43.45 | 46.11 | 40.24 | 40.36 | 38.47 | 28.53 | 32.91 | 35.60 |
| + ML Lex. + Panlex | 58.72 | 49.15 | 45.26 | 32.79 | 38.65 | 40.92 | 52.81 | 46.26 | 45.54 | 47.24 | 44.14 | 32.83 | 28.07 | 33.10 | 41.52 |
| + ML Lex. + Panlex + Filtering | 61.22 | 49.35 | 50.09 | 49.17 | 39.86 | 41.90 | 48.18 | 44.12 | 50.91 | 42.49 | 37.08 | 33.39 | 35.47 | 33.81 | 30.98 |
| $\text{mBART}_{\text{Large}}$ (600M) | | | | | | | | | | | | | | | |
| + fine-tuned on SST | 6.19 | 42.10 | 23.12 | 32.42 | 29.14 | 35.25 | 39.74 | 57.52 | 49.02 | 12.53 | 14.57 | 23.87 | 28.16 | 8.97 | 17.25 |
| + ML Lex. | 57.28 | 42.14 | 36.71 | 43.00 | 35.83 | 38.61 | 39.75 | 55.71 | 51.55 | 29.21 | 23.80 | 29.45 | 39.28 | 38.26 | 27.17 |
| + ML Lex. + Panlex | 25.49 | 33.58 | 29.64 | 40.12 | 27.91 | 35.70 | 35.43 | 60.61 | 54.16 | 21.83 | 18.87 | 26.18 | 29.34 | 18.25 | 16.99 |
| + ML Lex. + Panlex + Filtering | 27.30 | 21.83 | 24.94 | 39.75 | 29.31 | 33.02 | 29.58 | 62.92 | 54.19 | 21.33 | 17.03 | 30.69 | 28.52 | 19.56 | 13.68 |
| $\text{mT5}_{\text{Base}}$ (580M) | | | | | | | | | | | | | | | |
| + fine-tuned on SST | 57.69 | 44.85 | 40.66 | 34.80 | 36.91 | 40.86 | 50.48 | 44.64 | 38.68 | 23.48 | 23.66 | 22.65 | 31.24 | 38.38 | 34.94 |
| + ML Lex. | 60.93 | 50.01 | 48.75 | 41.54 | 45.58 | 44.56 | 46.89 | 44.97 | 36.91 | 38.82 | 39.60 | 27.89 | 28.33 | 51.10 | 52.30 |
| + ML Lex. + Panlex | 62.41 | 51.19 | 47.82 | 43.07 | 44.17 | 44.95 | 53.56 | 48.19 | 42.02 | 33.94 | 40.23 | 26.34 | 29.17 | 52.87 | 49.41 |
| + ML Lex. + Panlex + Filtering | 63.52 | 50.95 | 43.89 | 45.86 | 44.18 | 44.97 | 51.92 | 48.81 | 48.98 | 34.52 | 34.56 | 34.42 | 31.24 | 49.92 | 38.98 |
| $\text{mT5}_{\text{Large}}$ (1B) | | | | | | | | | | | | | | | |
| + fine-tuned on SST | 43.87 | 35.09 | 33.22 | 29.86 | 33.75 | 32.12 | 46.47 | 44.47 | 26.31 | 20.38 | 21.80 | 21.63 | 27.80 | 27.45 | 28.68 |
| + ML Lex. | 55.66 | 50.79 | 51.59 | 43.68 | 45.17 | 43.39 | 47.26 | 46.83 | 40.84 | 37.76 | 38.81 | 36.61 | 25.23 | 52.16 | 50.53 |
| + ML Lex. + Panlex | 63.45 | 54.80 | 53.36 | 43.53 | 41.60 | 44.54 | 53.44 | 33.77 | 44.00 | 39.55 | 43.49 | 32.35 | 31.39 | 53.37 | 44.17 |
| + ML Lex. + Panlex + Filtering | 56.09 | 48.07 | 46.59 | 44.85 | 44.35 | 44.63 | 46.60 | 48.82 | 46.01 | 36.35 | 34.41 | 37.06 | 25.89 | 46.93 | 34.57 |

Table 14: All **3-way classification** results for the 14 African languages from SemEval 2023..

| Task | Data | Label | Multilabel | Original (train/dev/test) | Binary (train/dev/test) |
|---|---|---|---|---|---|
| Stance | WT-WT (Conforti et al., 2020) | support, refute, comment, unrelated | No | 41027/5128/5129 | 8663/1070/1151 |
| | P-Stance Li et al. (2021) | favor, against | No | 17224/2193/2157 | 17224/2193/2157 |
| Hate speech | HS1 (Founta et al., 2018) | abusive, normal, hateful, spam | No | 79996/10000/10000 | 68803/8560/8603 |
| | HS2 Vidgen et al. (2021) | none, derogation, animosity, dehumanization, threatening, support | No | 27256/3422/3356 | 27256/3422/3356 |
| Emotion | GoEmotions (Demszky et al., 2020) | 27 emotions | Yes | 43410/5426/5427 | 28264/3566/3551 |
| | SemEval2018 (Mohammad et al., 2018) | 11 emotions | Yes | 6838/886/3259 | 5902/795/2846 |

Table 15: Datasets used in our additional experiments.

| Model | Stance | | Hate Speech | | Emotion | |
|---|---|---|---|---|---|---|
| | **WT-WT** | **P-Stance** | **HS1** | **HS2** | **GoEmotions** | **SemEval2018** |
| *Binary classes* | | | | | | |
| $\text{mBERT}_{\text{Base}}$ | 79.88 | 60.66 | 83.72 | 54.49 | 57.45 | 60.46 |
| + EN Lex. | **85.39** | **61.16** | **87.95** | **54.74** | **72.17** | **79.62** |
| *Original classes* | | | | | | |
| $\text{mBERT}_{\text{Base}}$ | 44.43 | 60.66 | 69.45 | 43.26 | 14.20 | 17.57 |
| + EN Lex. | **44.97** | **61.12** | **71.28** | **43.89** | **14.91** | **32.93** |

Table 16: Lexicon-based pretraining performance (macro-F1) in six other English semantic tasks. The results are based on the limited training data scenario.