Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 258–266
Marseille, 20-25 June 2022

XLM-T: Multilingual Language Models in Twitter for Sentiment Analysis and Beyond

Francesco Barbieri♣, Luis Espinosa Anke♢, Jose Camacho-Collados♢
♣ Snap Inc., ♢ Cardiff NLP, School of Computer Science and Informatics, Cardiff University
♣ Santa Monica, California, USA ♢ Cardiff, Wales, United Kingdom
fbarbieri@snap.com, {espinosa-ankel,camachocolladosj}@cardiff.ac.uk

Abstract

Language models are ubiquitous in current NLP, and their multilingual capacity has recently attracted considerable attention. However, current analyses have almost exclusively focused on (multilingual variants of) standard benchmarks, and have relied on clean pre-training and task-specific corpora as multilingual signals. In this paper, we introduce XLM-T, a model to train and evaluate multilingual language models in Twitter. In this paper we provide: (1) a new strong multilingual baseline consisting of an XLM-R (Conneau et al., 2020) model pre-trained on millions of tweets in over thirty languages, alongside starter code to subsequently fine-tune on a target task; and (2) a set of unified sentiment analysis Twitter datasets in eight different languages and a XLM-T model fine-tuned on them.

Keywords: sentiment analysis, language models, Twitter, multilinguality

1. Introduction

Multilingual NLP is increasingly becoming popular. Despite the concerning disparity in terms of language resource availability (Joshi et al., 2020), the advent of Language Models (LMs) has indisputably enabled a myriad of multilingual architectures to fluorish, ranging from LSTMs to the arguably more popular transformer-based models (Chronopoulou et al., 2019; Pires et al., 2019). Multilingual LMs integrate streams of multilingual textual data without being tied to one single task, learning *general-purpose multilingual representations* (Hu et al., 2020). As testimony of this landscape, we find multilingual variants stemming from well-known monolingual LMs, which have now become a standard among the NLP community. For instance, mBERT from BERT (Devlin et al., 2019), mT5 (Xue et al., 2020) from T5 (Raffel et al., 2020) or XLM-R (Conneau et al., 2020) from RoBERTa (Liu et al., 2019). Social media data, however, and specifically Twitter (the platform we focus on in this paper), seem to be so far surprisingly neglected from this trend of massive multilingual pretraining. This may be due to, in addition to its well-known uncurated nature (Derczynski et al., 2013), because of discoursive and platform-specific factors such as out-of-distribution samples, misspellings, slang, vulgarisms, emoji and multimodality, among others (Barbieri et al., 2018; Camacho-Collados et al., 2020). This is an important consideration, as there is ample agreement that the quality of LM-based multilingual representations is strongly correlated with typological similarity (Hu et al., 2020), which is somewhat blurred out in the context of Twitter.

In this paper, we bridge this gap by introducing a toolkit for evaluating multilingual Twitter-specific language models. This framework, which we make available to the NLP community, is initially comprised of a large multilingual Twitter-specific LM based on XLM-R checkpoints (Section 2), from which we report an initial set of baseline results in different settings (including zero-shot). Moreover, we provide starter code for analyzing, fine-tuning and evaluating existing language models. To carry out a comprehensive multilingual evaluation, while also laying the foundations for future extensions, we devise a unified dataset in 8 languages for sentiment analysis (which we call *Unified Multilingual Sentiment Analysis Benchmark*, UMSAB henceforth), as this task is by far the most studied problem in NLP in Twitter (cf., e.g., (Salameh et al., 2015; Zhou et al., 2016; Meng et al., 2012; Chen et al., 2018; Rasooli et al., 2018; Vilares et al., 2017; Barnes et al., 2019; Patwa et al., 2020; Barriere and Balahur, 2020)). XLM-T and associated data is released at `https://github.com/cardiffnlp/xlm-t`. Finally, in order to have a solid point of comparison with respect to standard English Twitter tasks, we also report results on the TweetEval framework (Barbieri et al., 2020). Our results suggest that when fine-tuning task-specific Twitter-based multilingual LMs, a domain-specific model proves more consistent than its general-domain counterpart, and that in some cases a smart selection of training data may be preferred than large-scale fine-tuning on many languages.

2. XLM-T: Language Models in Twitter

Our framework revolves around Twitter-specific language models. In particular, we train our own multilingual language-specific language model (Section 2.1), which we then fine-tune for various monolingual and multilingual applications, and for which we provide a suitable interface (Section 2.2). Additionally, we complement these functionalities with starter code for these and other typical Twitter-related NLP tasks (Section

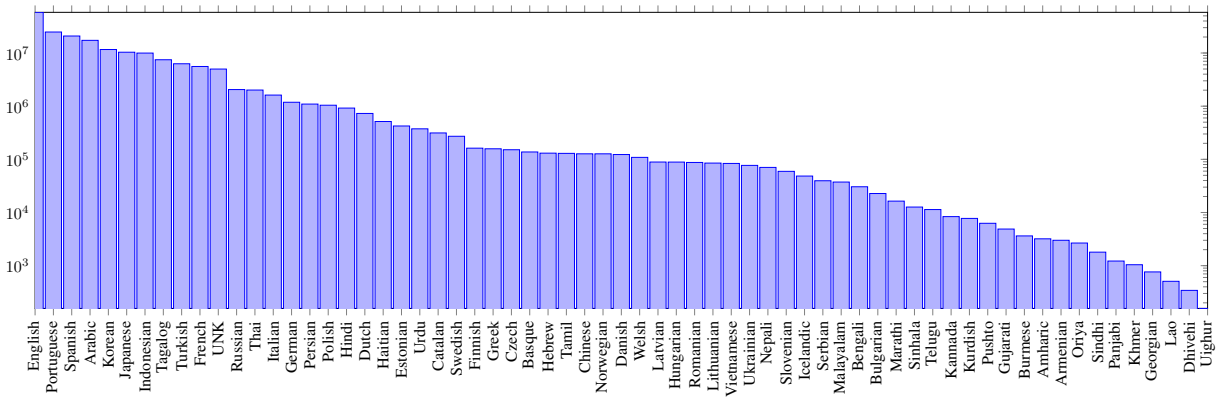

Figure 1: Distribution of languages of the 198M tweets used to finetune the Twitter-based language model (log scale). UNK corresponds to unidentified tweets according to the Twitter API.

2.3), e.g., computing tweet embeddings and multilingual sentiment analysis evaluation.

2.1. Released Language Models

We used the Twitter API to retrieve 198M tweets[1] posted between May'18 and March'20, which are our source data for LM pretraining. We only considered tweets with at least three tokens and with no URLs to avoid bot tweets and spam advertising. Additionally, we did not perform language filtering, aiming at capturing a general distribution. Figure 1 lists the 30 most represented languages by frequency, showing a prevalence of widely spoken languages such as English, Portuguese and Spanish, with the first significant drop in frequency affecting Russian at the 11th position.

In terms of opting for pretraining a LM from scratch or building upon an existing one, we follow (Gururangan et al., 2020) and (Barbieri et al., 2020) and *continue training* an XLM-R language model from publicly available checkpoints[2], which we selected due to the high results it has achieved in several multilingual NLP tasks (Hu et al., 2020). We use the same masked LM objective, and train until convergence in a validation set. The model converged after about 14 days on 8 NVIDIA V100 GPUs.[3]

While this multilingual language model (referred to as *XLM-Twitter* henceforth) is the main focus on this paper, our toolkit also integrates monolingual language models of any nature, including the English monolingual Twitter models released in Barbieri et al. (2020) and Nguyen et al. (2020).

2.2. Language Model Fine-tuning

In this section we explain the fine-tuning implementation of our framework. The main task evaluated in this paper is tweet classification, for which we provide unified datasets. One of the main differences with respect to standard fine-tuning is that we integrate the adapter technique (Houlsby et al., 2019), by means of which we freeze the LM and only fine-tune one additional classification layer. We follow the same adapter configuration proposed in Pfeiffer et al. (2020). This technique provides benefits in terms of memory and speed, which in practice facilitates the usage of multilingual language models for a wider set of NLP practitioners and researchers.

2.3. Starter code

In order to enable fast prototyping on our framework, in addition to datasets and pretrained models we also provide Python code for feature extraction from Tweets (i.e., obtaining tweet embeddings), tweet classification, model fine-tuning, and evaluation.

Feature extraction. Figure 2 shows sample code on how to extract tweet embeddings using our XLM-T language model, including its applicability for tweet similarity.

Fine-tuning. Figure 3 shows the fine-tuning procedure using a custom language model. This process can be performed with either adapters (used in our evaluation for efficiency) or the more standard language model fine-tuning. In practice, note that both options would be implemented in a very similar way, as both sit on top of the Huggingface `transformers` library.

Inference (tweet classification). We provide an easy interface to perform inference with our fine-tuned models. To this end, we rely on Hugging Face's *pipelines*. Figure 4 shows an example for a sentiment prediction using our XLM-T model fine-tuned on UMSAB. Note that, while the examples provided are for sentiment analysis, any tweet classification task such as those included in TweetEval are compatible.

Evaluation. Finally, XLM-T includes evaluation code to seamlessly evaluate any language model on senti-

[1] 1,724 million tokens (12G of uncompressed text).

[2] `https://huggingface.co/xlm-roberta-base`.

[3] The estimated cost for the language model pre-training is USD 5,000 on Google Cloud.

```
MODEL =  "cardiffnlp/twitter-xlm-roberta-base"
tokenizer = AutoTokenizer.from_pretrained(MODEL)
model = AutoModel.from_pretrained(MODEL)

def get_embedding(text):
    text = preprocess(text)
    encoded_input = tokenizer(text, return_tensors='pt')
    features = model(**encoded_input)
    features = features[0].detach().cpu().numpy()
    features_mean = np.mean(features[0], axis=0)
    return features_mean

query = "Acabo de pedir pollo frito 🐣" #spanish

tweets = ["We had a great time! ⚽", # english
          "We hebben een geweldige tijd gehad! ⛩", # dutch
          "Nous avons passé un bon moment! 🎥", # french
          "Ci siamo divertiti! 🍝"] # italian

d = defaultdict(int)
for tweet in tweets:
    sim = 1-cosine(get_embedding(query),get_embedding(tweet))
    d[tweet] = sim
```

```
print('Most similar to: ',query)
print('----------------------------------------')
for idx,x in enumerate(sorted(d.items(), key=lambda x:x[1], reverse=True)):
  print(idx+1,x[0])
```

```
Most similar to:  Acabo de pedir pollo frito 🐣
----------------------------------------
1 Ci siamo divertiti! 🍝
2 Nous avons passé un bon moment! 🎥
3 We had a great time! ⚽
4 We hebben een geweldige tijd gehad! ⛩
```

Figure 2: Code snippet showcasing the feature extraction and tweet similarity interface. Note that using our Twitter-specific XLM-R model leads to emoji playing a crucial role in the semantics of the tweet.

```
training_args = TrainingArguments(
    output_dir='./results',                  # output directory
    num_train_epochs=EPOCHS,                 # total number of training epochs
    per_device_train_batch_size=BATCH_SIZE,  # batch size per device during training
    per_device_eval_batch_size=BATCH_SIZE,   # batch size for evaluation
    warmup_steps=100,                        # number of warmup steps for lr scheduler
    weight_decay=0.01,                       # strength of weight decay
    logging_dir='./logs',                    # directory for storing logs
    logging_steps=10,                        # when to print log
    load_best_model_at_end=True,             # load or not best model at the end
)

model = AutoModelForSequenceClassification.from_pretrained(MODEL, num_labels=num_labels)
```

```
trainer = Trainer(
    model=model,                             # the instantiated 🤗 Transformers model
    args=training_args,                      # training arguments, defined above
    train_dataset=train_dataset,             # training dataset
    eval_dataset=val_dataset                 # evaluation dataset
)

trainer.train()
trainer.save_model("./results/best_model") # save best model
```

Figure 3: Fine-tuning procedure including the declaration of dataset and parameters, training procedure and saving of the model.

ment analysis, either focusing on a subset or all of the languages included in UMSAB (cf. Section 3.2). Specifically, we provide bash scripts which handle input arguments such as gold test data, prediction files and target language(s).

3. Evaluation

We assess the reliability of our released multilingual Twitter-specific language model in three different ways: (1) we perform an evaluation on a wide range of English-specific datasets (Section 3.1); (2) we compose a large multilingual benchmark for sentiment analysis where we assess the multilingual capabilities of the language model (Section 3.2); (3) we perform a qualitative analysis based on cross-lingual tweet similarity (Section 3.3).

Experimental Setting. In each experiment we perform three runs with different seeds, and use early stop-

```
from transformers import pipeline
model_path = "cardiffnlp/twitter-xlm-roberta-base-sentiment"
sentiment_task = pipeline("sentiment-analysis", model=model_path)
sentiment_task("Huggingface es lo mejor! Awesome library 🤗😎")
```

```
[{'label': 'Positive', 'score': 0.9343640804290771}]
```

Figure 4: Sentiment analysis inference using XLM-T.

ping on the validation loss. We only tune the learning rate (0.001 and 0.0001) and, unless noted otherwise, all results we report are the average of three runs of macro-average F1 scores.In terms of models, we evaluate a standard pre-trained **XLM-R** and **XLM-Twitter**, our XLM-R model pretrained on a multilingual Twitter dataset starting from XLM-R checkpoints (see Section 2.1). For the monolingual experiments we also include a FastText (FT) baseline (Joulin et al., 2017), which relies on monolingual FT embeddings trained on Common Crawl and Wikipedia (Grave et al., 2018) as initialization for each language lookup table.

3.1. Monolingual Evaluation (TweetEval)

In order to provide an additional point of comparison for our released multilingual language model, we perform an evaluation on standard Twitter-specific tasks in English, for which we can compare its performance with existing models. In particular, we evaluate XLM-Twitter on a suite of seven heterogeneous tweet classification tasks from the TweetEval benchmark (Barbieri et al., 2020). TweetEval is composed of seven tasks: emoji prediction (Barbieri et al., 2018), emotion recognition (Mohammad et al., 2018), hate speech detection (Basile et al., 2019), irony detection (Van Hee et al., 2018), offensive language identification (Zampieri et al., 2019), sentiment analysis (Rosenthal et al., 2019) and stance detection[4] (Mohammad et al., 2016).

Table 1 shows the results of the language models and TweetEval baselines[5] As can be observed, our proposed XLM-R-Twitter improves over strong baselines such as RoBERTa-base and XLM-R that do not make use of Twitter corpora, and RoBERTa-Twitter, which is trained on Twitter corpora only. This highlights the reliability of our multilingual model in language-specific settings. However, it underperforms when compared with monolingual Twitter-specific models, such as the RoBERTa model futher pre-trained on English tweets proposed in (Barbieri et al., 2020), as well as BERTweet (Nguyen et al., 2020), which was trained on a corpus that is an order of magnitude larger.[6] This is to be expected as goes in line with previous research that shows that multilin-

[4]The stance detection dataset is in turn split into five subtopics.

[5]Please refer to the original TweetEval paper for details on the implementation of all the baselines.

[6]While XLM-R-Twitter was fine-tuned on the same amount of English tweets (60M) than RoBERTa-Tw, BERTweet was trained on 850M English tweets.

	Emoji	Emotion	Hate	Irony	Offensive	Sentiment	Stance	ALL
SVM	29.3	64.7	36.7	61.7	52.3	62.9	67.3	53.5
FastText	25.8	65.2	50.6	63.1	73.4	62.9	65.4	58.1
BLSTM	24.7	66.0	52.6	62.8	71.7	58.3	59.4	56.5
RoB-Bs	30.9±0.2 (30.8)	76.1±0.5 (76.6)	46.6±2.5 (44.9)	59.7±5.0 (55.2)	79.5±0.7 (78.7)	71.3±1.1 (72.0)	68±0.8 (70.9)	61.3
RoB-RT	31.4±0.4 (**31.6**)	78.5±1.2 (**79.8**)	52.3±0.2 (**55.5**)	61.7±0.6 (62.5)	80.5±1.4 (**81.6**)	72.6±0.4 (**72.9**)	69.3±1.1 (**72.6**)	**65.2**
RoB-Tw	29.3±0.4 (29.5)	72.0±0.9 (71.7)	46.9±2.9 (45.1)	65.4±3.1 (65.1)	77.1±1.3 (78.6)	69.1±1.2 (69.3)	66.7±1.0 (67.9)	61.0
XLM-R	28.6±0.7 (27.7)	72.3±3.6 (68.5)	44.4±0.7 (43.9)	57.4±4.7 (54.2)	75.7±1.9 (73.6)	68.6±1.2 (69.6)	65.4±0.8 (66.0)	57.6
XLM-Tw	30.9±0.5 (30.8)	77.0±1.5 (78.3)	50.8±0.6 (51.5)	69.9±1.0 (**70.0**)	79.9±0.8 (79.3)	72.3±0.2 (72.3)	67.1±1.4 (68.7)	64.4
SotA	*33.4*	*79.3*	*56.4*	*82.1*	*79.5*	*73.4*	*71.2*	*67.9*
Metric	M-F1	M-F1	M-F1	$F^{(i)}$	M-F1	M-Rec	AVG ($F^{(a)}$,$F^{(f)}$)	TE

Table 1: TweetEval test results. For neural models we report both the average result from three runs and its standard deviation, and the best result according to the validation set (parentheses). *SotA* results correspond to the best TweetEval reported system, i.e., BERTweet.

gual models tend to underperform monolingual models in language-specific tasks (Rust et al., 2020).[7] In the following section we evaluate XLM-Twitter on multilingual settings, including evaluation in monolingual and cross-lingual scenarios.

3.2. Multilingual Evaluation (Sentiment Analysis)

We focus our evaluation on multilingual Sentiment Analysis (SA). We first flesh out the process followed to compile and unify our cross-lingual SA benchmark (Section 3.2.1). Our experiments[8] can then be grouped into two types: when no training in the target language is available, i.e., zero-shot (Section 3.2.2), and when the evaluated models have access to target language training data, either alone or as part of a larger fully multilingual training set (Section 3.2.3).

3.2.1. Unified Multilingual Sentiment Analysis Benchmark (UMSAB)

We aim at constructing a balanced multilingual SA dataset, i.e., where all languages are equally distributed in terms of frequency, and with representation of typologically distant languages. To this end, we compiled monolingual SA datasets for eight diverse languages. We list the languages and relevant statistics in Table 3, as well as their spanning timeframes. Given that retaining the original distribution would skew the unified dataset towards the most frequent languages, we established a maximum number of tweets corresponding to the size of the smallest dataset, specifically the 3,033 for the Hindi portion, and prune all data splits for all languages with this threshold. This leaves 1,839 training tweets (with 15% of them allocated to a fixed validation set), and 870 for testing. The total size of the dataset is thus 24,262 tweets. Let us highlight two additional important design decisions: first, we enforced a balanced distribution across the three labels (positive, negative and neutral), and second, we kept the original training/test splits in each dataset. After this preprocessing, we obtain 8 datasets of 3,033 instances, respectively. Note that some languages in this dataset agglutinate or refer to specific variations. In particular, we use Hindi to refer to the grouping of Hindi, Bengalu and Tamil, Portuguese for Brazilian Portuguese, and Spanish for Iberian, Peruvian and Costa Rican variations.

3.2.2. Zero-shot Cross-lingual Transfer

Table 2 shows zero-shot results of XLM-R and XLM-Twitter in our multilingual sentiment analysis benchmark. The performance of both models is competitive, especially considering the diversity of domains[9] and that the source language was not seen during training. An interesting observation concerns those cases in which zero-shot models outperform their monolingual counterparts (e.g., English→Arabic or Italian→Hindi). Additionally, XLM-Twitter proves more robust, achieving the best overall results in six of the eight languages, with consistent improvements in general, and with remarkable improvements in e.g., Hindi, outperforming XLM-R by 7.9 absolute points. Finally, let us provide some insights on the results obtained in an all-minus-one (the **All-1** columns in Table 2) setting. Here, notable cases are, first, Hindi, in which XLM-R and XLM-Twitter models benefit substantially by having access to more training data, with this improvement being more pronounced in XLM-Twitter. Second, the results for the English dataset suggest that compiling a larger training set helps, although this may be also attributed to identical tokens shared between English and the other languages, such as named entities, hashtags or colloquialisms and slang.

[7] It has been shown that this performance difference could be further decreased by using language-specific tokenizers (Rust et al., 2020), but this was out of scope for this paper.

[8] Standard deviation and best run results are provided, for completeness, in the appendix.

[9] For instance, for Arabic we find trending topics such as iPhone or vegetarianism, where the Portuguese dataset is dominated by comments on TV shows.

	XLM-R									XLM-Twitter								
	Ar	**En**	**Fr**	**De**	**Hi**	**It**	**Pt**	**Es**	***All-1***	**Ar**	**En**	**Fr**	**De**	**Hi**	**It**	**Pt**	**Es**	***All-1***
Ar	63.6	**64.1**	54.4	53.9	22.9	57.4	62.4	62.2	*59.2*	67.7	**66.6**	62.1	59.3	46.3	63.0	60.1	65.3	*64.3*
En	64.2	68.2	61.6	63.5	23.7	**68.1**	65.9	67.8	*68.2*	64.0	66.9	60.6	67.8	35.2	67.7	61.6	**68.7**	*70.3*
Fr	45.4	52.1	72.0	36.5	16.7	43.3	40.8	**56.7**	*53.6*	47.7	**59.2**	68.2	38.7	20.9	45.1	38.6	52.5	*50.0*
De	43.5	**64.4**	55.2	73.6	21.5	60.8	60.1	62.0	*63.6*	46.5	65.0	56.4	76.1	36.9	**66.3**	65.1	65.8	*65.9*
Hi	48.2	52.7	43.6	47.6	36.6	**54.4**	51.6	51.7	*49.9*	50.0	55.5	51.5	44.4	40.3	**56.1**	51.2	49.5	*57.8*
It	48.8	65.7	63.9	**66.9**	22.1	71.5	63.1	58.9	*65.7*	41.9	59.6	60.8	64.5	24.6	70.9	**64.7**	55.1	*65.2*
Pt	41.5	63.2	57.9	59.7	26.5	59.6	67.1	**65.0**	*65.0*	56.4	**67.7**	62.8	64.4	26.0	67.1	76.0	64.0	*71.4*
Es	47.1	63.1	56.8	57.2	26.2	57.6	**63.1**	65.9	*63.0*	52.9	66.0	64.5	58.7	30.7	62.4	**67.9**	68.5	*66.2*

Table 2: Zero-shot cross-lingual sentiment analysis results (F1). We use the best model in the language on the column and evaluate on the test set of the language of each row. For example, when we forward the best XLM-R trained on English text on the Arabic test set we obtain 64.1. In the columns *All minus one (**All-1**)* we train on all the languages excluding the one of each row. For example, we obtain a F1 of 59.2 on the Arabic test set when we train an XLM-R using all the languages excluding Arabic. On the diagonals, in gray, models are trained and evaluated on the same language.

Lang.	Dataset	Time-Train	Time-Test
Arabic	SemEval-17 (Rosenthal et al., 2017)	09/16-11/16	12/16-1/17
English	SemEval-17 (Rosenthal et al., 2017)	01/12-12/15	12/16-1/17
French	Deft-17 (Benamara et al., 2017)	2014-2016	Same
German	SB-10K (Cieliebak et al., 2017)	8/13-10/13	Same
Hindi	SAIL 2015 (Patra et al., 2015)	NA,3-month	Same
Italian	Sentipolc-16 (Barbieri et al., 2016)	2013-2016	2016
Portug.	SentiBR (Brum and Nunes, 2017)	1/17-7/17	Same
Spanish	Intertass (Díaz-Galiano et al., 2018)	7/16-01/17	Same

Table 3: Sentiment analysis datasets for the eight languages used in our experiments.

	Monolingual			Bilingual		Multilingual	
	FT	**XLM-R**	**XLM-T**	**XLM-R**	**XLM-T**	**XLM-R**	**XLM-T**
Ar	45.98	63.56	**67.67**	63.63 (En)	67.65 (En)	64.31	66.89
En	50.85	68.18	66.89	65.07 (It)	67.47 (Es)	68.52	**70.63**
Fr	54.82	71.98	68.19	**73.55** (Sp)	68.24 (En)	70.52	71.18
De	59.56	73.61	76.13	72.48 (En)	75.49 (It)	72.84	**77.35**
Hi	37.08	36.60	40.29	33.57 (It)	55.35 (It)	53.39	**56.39**
It	54.65	71.47	70.91	70.43 (Ge)	**73.50** (Pt)	68.62	69.06
Pt	55.05	67.11	75.98	71.87 (Sp)	**76.08** (En)	69.79	75.42
Sp	50.06	65.87	68.52	67.68 (Po)	**68.68** (Pt)	66.03	67.91
All	51.01	64.80	66.82	64.78	69.06	66.75	**69.35**

Table 4: Cross-lingual sentiment analysis F1 results on target languages using target language training data (Monolingual) only, combined with training data from another language (Bilingual) and with all languages at once (Multilingual). "All" is computed as the average of all individual results.

3.2.3. Cross-lingual Transfer with Target Language Training Data

Table 4 shows macro-F1 results for the following three settings: (1) **monolingual**, where we train and test in one single language; (2) **bilingual**, where we use the best-performing cross-lingual zero-shot model, and continue fine-tuning on training data from the target language; and (3) an entirely **multilingual** setting where we train with data from all languages. One of the most notable conclusions in the light of these figures is that increasing the training data even in different languages is a useful strategy, and is particularly rewarding in the case of XLM-Twitter and in challenging datasets and languages (e.g., the Hindi results significantly increase from 40.29 to 56.39). Interestingly, a smart selection of languages based on validation accuracy achieves better results than if trained on all languages in half of the cases. This may be due to the (dis)similarity of the datasets (in terms of topic or typological proximity), although overall the main conclusion we can draw is that there is an obvious trade-off, as a single multilingual model is often more practical and versatile.

3.3. Qualitative Analysis

As an additional qualitative analysis, we plot in Figure 5 a sample of similarity scores (by cosine distance) between XLM-Twitter-based embeddings obtained from the English *training set* and the sentiment analysis test sets for the other 7 languages (see Section 3.2.1). In addition to the clearly low resemblance with Hindi, we find that the most similar languages in the embedding space are English and French, suggesting that not only typology, but also topic overlap, may play an important role in the quality of these multilingual representations. This becomes even more apparent in Arabic, which differs from English in typology and script, but has similar representations. The Arabic and English datasets were obtained using the same keywords.

4. Conclusions

We have presented a comprehensive framework for Twitter-based multilingual LMs, including the release of a new multilingual LM trained on almost 200M tweets. As main test bed for our multilingual experiments, we focused on sentiment analysis, for which we collected datasets in eight languages. After a unification and standardization of the evaluation benchmark, we compared the Twitter-based multilingual language model with a standard multilingual language model trained on general-domain corpora. This multilingual language model along with starting and evaluation code are re-

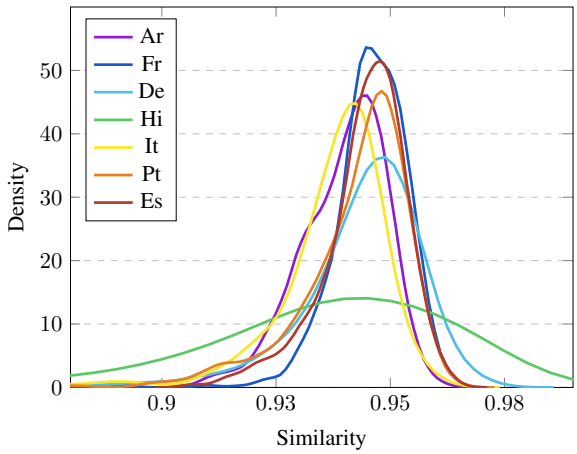

Figure 5: Cross-lingual similarity (by cosine distance) between the English training set and the test sets in the other 7 languages. The embeddings are obtained by averaging all the XLM-Twitter contextualized embeddings for each tweet.

leased to facilitate research in Twitter at a multilingual scale (over thirty languages used for training data). The results highlight the potential of the domain-specific language model, as more suited to handle social media and specifically multilingual SA. Finally, our analysis reveals trends and potential for this Twitter-based multilingual language model in zero-shot cross-lingual settings when language-specific training data is not available. For future work we are planning to extend this analysis to more languages and tasks, but also to deepen the cross-lingual zero and few shot analysis, particularly focusing on typologically similar languages. Finally, and due to the seasonal nature of Twitter, it would also be interesting to explore correlations between topic distribution and trends and performance in downstream applications.

Acknowledgments

We would like to thank Eugenio Martínez Cámara for his involvement in the first stages of this project. Jose Camacho-Collados is supported with a UKRI Future Leaders Fellowship.

5. Bibliographical References

Barbieri, F., Basile, V., Croce, D., Nissim, M., Novielli, N., and Patti, V. (2016). Overview of the evalita 2016 sentiment polarity classification task. In *Proceedings of third Italian conference on computational linguistics (CLiC-it 2016) & fifth evaluation campaign of natural language processing and speech tools for Italian. Final Workshop (EVALITA 2016)*.

Barbieri, F., Camacho-Collados, J., Ronzano, F., Anke, L. E., Ballesteros, M., Basile, V., Patti, V., and Saggion, H. (2018). Semeval 2018 task 2: Multilingual emoji prediction. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 24–33.

Barbieri, F., Camacho-Collados, J., Espinosa Anke, L., and Neves, L. (2020). TweetEval: Unified benchmark and comparative evaluation for tweet classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1644–1650, Online, November. Association for Computational Linguistics.

Barnes, J., Øvrelid, L., and Velldal, E. (2019). Sentiment analysis is not solved! assessing and probing sentiment classification. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 12–23, Florence, Italy, August. Association for Computational Linguistics.

Barriere, V. and Balahur, A. (2020). Improving sentiment analysis over non-English tweets using multilingual transformers and automatic translation for data-augmentation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 266–271, Barcelona, Spain (Online), December. International Committee on Computational Linguistics.

Basile, V., Bosco, C., Fersini, E., Nozza, D., Patti, V., Rangel Pardo, F. M., Rosso, P., and Sanguinetti, M. (2019). SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63, Minneapolis, Minnesota, USA, June. Association for Computational Linguistics.

Benamara, F., Grouin, C., Karoui, J., Moriceau, V., and Robba, I. (2017). Analyse d'opinion et langage figuratif dans des tweets: présentation et résultats du défi fouille de textes deft2017. In *Défi Fouille de Textes DEFT2017. Atelier TALN 2017*. Association pour le Traitement Automatique des Langues (ATALA).

Brum, H. B. and Nunes, M. d. G. V. (2017). Building a sentiment corpus of tweets in brazilian portuguese. *arXiv preprint arXiv:1712.08917*.

Camacho-Collados, J., Doval, Y., Martínez-Cámara, E., Espinosa-Anke, L., Barbieri, F., and Schockaert, S. (2020). Learning cross-lingual word embeddings from twitter via distant supervision. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 72–82.

Chen, X., Sun, Y., Athiwaratkun, B., Cardie, C., and Weinberger, K. (2018). Adversarial deep averaging networks for cross-lingual sentiment classification. *Transactions of the Association for Computational Linguistics*, 6:557–570.

Chronopoulou, A., Baziotis, C., and Potamianos, A. (2019). An embarrassingly simple approach for transfer learning from pretrained language models. 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 2089–2095.

Cieliebak, M., Deriu, J. M., Egger, D., and Uzdilli, F.

(2017). A twitter corpus and benchmark resources for german sentiment analysis. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 45–51.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (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, July. Association for Computational Linguistics.

Derczynski, L., Ritter, A., Clark, S., and Bontcheva, K. (2013). Twitter part-of-speech tagging for all: Overcoming sparse and noisy data. In *Proceedings of the international conference recent advances in natural language processing ranlp 2013*, pages 198–206.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (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, June. Association for Computational Linguistics.

Díaz-Galiano, M. C., Martínez-Cámara, E., Ángel García Cumbreras, M., Vega, M. G., and Román, J. V. (2018). The democratization of deep learning in tass 2017. *Procesamiento del Lenguaje Natural*, 60(0):37–44.

Grave, E., Bojanowski, P., Gupta, P., Joulin, A., and Mikolov, T. (2018). Learning word vectors for 157 languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., and Smith, N. A. (2020). Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online, July. Association for Computational Linguistics.

Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., Attariyan, M., and Gelly, S. (2019). Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pages 2790–2799. PMLR.

Hu, J., Ruder, S., Siddhant, A., Neubig, G., Firat, O., and Johnson, M. (2020). Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *International Conference on Machine Learning*, pages 4411–4421. PMLR.

Joshi, P., Santy, S., Budhiraja, A., Bali, K., and Choudhury, M. (2020). The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online, July. Association for Computational Linguistics.

Joulin, A., Grave, E., Bojanowski, P., and Mikolov, T. (2017). Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain, April. Association for Computational Linguistics.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Meng, X., Wei, F., Liu, X., Zhou, M., Xu, G., and Wang, H. (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. Association for Computational Linguistics.

Mohammad, S., Kiritchenko, S., Sobhani, P., Zhu, X., and Cherry, C. (2016). Semeval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 31–41.

Mohammad, S., Bravo-Marquez, F., Salameh, M., and Kiritchenko, S. (2018). Semeval-2018 task 1: Affect in tweets. In *Proceedings of the 12th international workshop on semantic evaluation*, pages 1–17.

Nguyen, D. Q., Vu, T., and Tuan Nguyen, A. (2020). BERTweet: A pre-trained language model for English tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 9–14, Online, October. Association for Computational Linguistics.

Patra, B. G., Das, D., Das, A., and Prasath, R. (2015). Shared task on sentiment analysis in indian languages (sail) tweets-an overview. In *International Conference on Mining Intelligence and Knowledge Exploration*, pages 650–655. Springer.

Patwa, P., Aguilar, G., Kar, S., Pandey, S., PYKL, S., Gambäck, B., Chakraborty, T., Solorio, T., and Das, A. (2020). SemEval-2020 task 9: Overview of sentiment analysis of code-mixed tweets. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 774–790, Barcelona (online), December. International Committee for Computational Linguistics.

Pfeiffer, J., Kamath, A., Rücklé, A., Cho, K., and Gurevych, I. (2020). Adapterfusion: Non-destructive task composition for transfer learning. *arXiv preprint arXiv:2005.00247*.

Pires, T., Schlinger, E., and Garrette, D. (2019). How multilingual is multilingual bert? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2020). Exploring the limits of transfer learning with a unified

text-to-text transformer. *Journal of Machine Learning Research*, 21:1–67.

Rasooli, M. S., Farra, N., Radeva, A., Yu, T., and McKeown, K. (2018). Cross-lingual sentiment transfer with limited resources. *Machine Translation*, 32(1):143–165, Jun.

Rosenthal, S., Farra, N., and Nakov, P. (2017). Semeval-2017 task 4: Sentiment analysis in twitter. In *Proceedings of the 11th international workshop on semantic evaluation (SemEval-2017)*, pages 502–518.

Rosenthal, S., Farra, N., and Nakov, P. (2019). Semeval-2017 task 4: Sentiment analysis in twitter. *arXiv preprint arXiv:1912.00741*.

Rust, P., Pfeiffer, J., Vulić, I., Ruder, S., and Gurevych, I. (2020). How good is your tokenizer? on the monolingual performance of multilingual language models. *arXiv preprint arXiv:2012.15613*.

Salameh, M., Mohammad, S., and Kiritchenko, S. (2015). Sentiment after translation: A case-study on arabic social media posts. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 767–777, Denver, Colorado, May–June. Association for Computational Linguistics.

Van Hee, C., Lefever, E., and Hoste, V. (2018). Semeval-2018 task 3: Irony detection in english tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 39–50.

Vilares, D., Alonso, M. A., and Gómez-Rodríguez, C. (2017). Supervised sentiment analysis in multilingual environments. *Information Processing & Management*, 53(3):595 – 607.

Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., and Raffel, C. (2020). mt5: A massively multilingual pre-trained text-to-text transformer.

Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., and Kumar, R. (2019). SemEval-2019 task 6: Identifying and categorizing offensive language in social media (OffensEval). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 75–86, Minneapolis, Minnesota, USA, June. Association for Computational Linguistics.

Zhou, X., Wan, X., and Xiao, J. (2016). Cross-lingual sentiment classification with bilingual document representation learning. In *Proceedings of ACL*, pages 1403–1412, August.

A. Full Experimental Results

This appendix includes the full experimental results, including standard deviation after three runs and the best runs according to the validation set. Table 5 includes the monolingual results; Table 6, the cross-lingual results; and Table 7, the multilingual experiments.

	XLM		XLM-Twitter	
	F1 macro	**F1 Best**	**F1 macro**	**F1 Best**
Ar	63.56 ±1.29	64.89	**67.67 ±1.25**	69.03
En	**68.18 ±2.57**	69.64	66.89 ±1.19	67.82
Fr	**71.98 ±1.46**	72.86	68.19 ±1.55	69.20
De	73.61 ±0.22	73.75	**76.13 ±0.53**	76.58
Hi	36.6 ±4.36	41.46	**40.29 ±7.37**	48.79
It	**71.47 ±1.35**	73.02	70.91 ±0.87	71.41
Pt	67.11 ±1.1	67.89	**75.98 ±0.03**	76.01
Es	65.87 ±1.67	67.75	**68.52 ±0.69**	69.01

Table 5: Monolingual experiments. XLM and XLM-Twitter are finetuned for each language. F1 macro is the average of three runs and F1 best is the best one of them.

	XLM			XLM-Twitter		
Tar.	**Pre.**	**F1 Macro**	**F1 Best**	**Pre.**	**F1 Macro**	**F1 Best**
Ar	En	63.63 ±2.71	65.25	En	**67.65 ±0.1**	67.76
En	It	65.07 ±1.8	66.93	Sp	**67.47 ±0.46**	67.85
Fr	Sp	**73.55 ±0.92**	74.21	En	68.24 ±5.2	71.66
De	En	72.48 ±0.44	72.97	It	**75.49 ±0.67**	76.18
Hi	It	33.57 ±9.34	39.41	It	**55.35 ±0.38**	55.68
It	Ge	70.43 ±1.51	71.4	Po	**73.5 ±0.58**	74.12
Pt	Sp	71.87 ±0.24	72.14	En	**76.08 ±1.08**	76.78
Es	Po	67.68 ±0.87	68.66	Po	**68.68 ±0.2**	68.85

Table 6: Bilingual experiments. We finetune XLM and XLM-Twitter models for S/A in the target language (Tar.) but instead of starting with random initialization of the adapter, we start with the adapter pretrained (Pre.) in the language that best performed in the zero shot classification for the Target language (using validation).

	XLM		XLM-Twitter	
	F1 Avg	**F1 Best**	**F1 Avg**	**F1 Best**
Ar	64.31 ±1.92	66.52	**66.89 ±1.18**	67.68
En	68.52 ±1.42	69.85	**70.63 ±1.04**	71.76
Fr	70.52 ±1.76	72.24	**71.18 ±1.06**	72.32
De	72.84 ±0.28	73.15	**77.35 ±0.27**	77.62
Hi	53.39 ±2.00	54.97	**56.39 ±1.60**	57.32
It	68.62 ±2.23	70.97	**69.06 ±1.07**	70.12
Pt	69.79 ±0.57	70.37	**75.42 ±0.49**	75.86
Es	66.03 ±1.31	66.94	**67.91 ±1.43**	69.03
All	66.93 ±0.16	67.07	**69.45 ±0.63**	70.11

Table 7: Multilingual experiments. XLM-R and XLM-Twitter are finetuned using one single multilingual dataset. We evaluate the two multilingual models with the test set of each language and with the composition of all the test sets (All). F1 macro is the average of three runs and F1 best is the best one of them.