Fine-tuning XLM-RoBERTa on Social Media Data for Multilingual Sentiment Analysis Tasks

Feifei (Katelyn) Wang [1]
feifei@nyu.edu

Ruixi (Victoria) Zhang[1,2]
rz1424@nyu.edu

[1]Center for Data Science
New York University
60 Fifth Avenue
New York, NY 10011

[2]Courant Institute of Mathematical Sciences
New York University
251 Mercer St 801
New York, NY 10012

Abstract

This paper explores the influence of different hyperparameters and fine-tuning methods on XML-RoBERTa's performance on sentiment analysis tasks for social media texts in five languages: English, Chinese, Urdu, Marathi, and Spanish. Our data are mainly in the form of tweets, with Chinese data, from Weibo, as an exception. We selected hyperparameters like learning rates, number of batches, and epochs and run different trails to find the optimal setting. The one we chose within the time and resource limit is a learning rate of 3e$\hat{5}$ and a batch size of 64. Further, we explore XLM RoBERTa's cross-lingual transferability by fine-tuning the model on each language separately and using each model to test the performance in the same task of all five languages. As the result shows that there is relatively uniform performance in every model, our paper validated XLM RoBERTa's cross-lingual transferability.

1 Introduction

In recent years, constant efforts have been made into improving the performance of multilingual language tasks. The XLM-RoBERTa (Conneau et al., 2019) was introduced in 2019 by Facebook AI as a multilingual version of RoBERTa (Liu et al., 2019). It is a transformer-based multilingual language model (MLM) that was pre-trained on datasets in 100 languages. The XLM-RoBERTa model is famous for its state-of-the-art accuracy in different multilingual language understanding tasks.

In this paper, we seek to get more insights into how the XLM-RoBERTa model performs in sentiment analysis in different languages. Knowing that we want multilingual data, but preferably from the same domain, we decide to fine-tune XLM-RoBERTa on social media data, including tweets, Weibo posts, Instagram posts, etc. We'd like to use a multi-lingual data set for sentiment analysis to test and improve our fine-tuning process for XLM-RoBERTa.

Moreover, to further investigate XLM-RoBERTa's performance in cross-lingual transferability. Therefore, in terms of language selection, we want to take a holistic approach and look at a semantically diverse set of languages, which may also have different sentence structures. We also include data-sets for lower-resource languages.

Our code and data are available on GitHub.[1]

2 Related Works

2.1 Multilingual Language Understanding

With the development of transfer learning and mask language models(MLM), there are significant breakthroughs in improving the performing multilingual language understanding tasks in the past few years.

In 2018, Google released Bidirectional Encoder Representations from Transformers (BERT)(Devlin et al., 2018). BERT soon gains a lot of attention as the state-of-the-art model for language understanding tasks. Soon after that, Google released Multilingual Bert (M-BERT)[2], marking a breakthrough in multilingual language learning. M-BERT was pre-trained with Wikipedia Corpus in 104 languages. In the experiment by Google Research, M-BERT demonstrates the strong ability to generalize across languages(Pires et al., 2019). That is, researchers can fine-tune M-BERT on a specific task in one language and expect M-BERT to demonstrate the ability to perform the same

[1]https://github.com/Ruixi-Zhang/XLM-RoBERTa-Sentiment-Analysis

[2]https://github.com/google-research/bert/blob/master/multilingual.md

task in other languages. However, later studies revealed that such performance was partly due to linguistic and lexical similarities in the source and target language (Wang et al., 2019). The conclusion was based on an experiment in three source-target language pairs: English– Spanish, English–Russian, and English–Hindi. The result shows that the cross-lingual transferability performs the best in the English - Spanish source-target language pairs. When there are no lexical similarities between two languages, the M-BERT model learns almost no cross-lingual similarities. The study also suggested that the depth and complexity of the model also affect the model's performance in cross-language generalization. Moreover, another study by the Johns Hopkins Department of Computer Science also revealed that M-BERT performs poorly in low-resource languages and the cross-lingual transferability only covers one third of the 104 languages(Wu and Dredze, 2020).

In 2019, Facebook AI introduced XLM-RoBERTa(Conneau et al., 2019), which outperformed M-BERT in many multiple multilingual tasks in terms of accuracy and F1 scores. On top of that, Facebook claims that XLM-RoBERTa performs 23% better than M-BERT for low-resource languages in cross-lingual classification and many other tasks(Conneau et al., 2019). There is an essential trade-off in the construction of multilingual language models. Adding more languages will ensure better performance on lower-resource languages while causing the overall performance of the language model to decrease. With XLM-RoBERTa, Facebook AI demonstrated that such trade-off can be alleviated with a greater model capacity. Unlike how M-BERT was trained with Wikipedia texts, XLM-RoBERTa was trained with data from the Common Crawl corpus. Common Crawl corpus includes web data that was collected from 2008 until now, including web page data, extracted metadata, and text extractions[3]. The Common Crawl corpus had become an essential resource for NLP and studies related to language understanding with its variety of languages and high-quality content (Wenzek et al., 2019). The diversity and abundance of web data provide XLM-RoBERTa with a stronger foundation for model capacity, especially in lower-resource languages. Lower-resource languages were also included in the ablation studies process, ensuring that the measurement of the model performance also took lower-resource languages into account.

Therefore, we believe that XLM-RoBERTa is indeed the state-of-the-art model for multilingual language tasks and it would be a suitable choice for conducting sentiment analysis.

2.2 Social Media Posts Sentiment Analysis

Sentiment analysis, sometimes called opinion mining(Alsaeedi and Khan, 2019), is a process that extracts text's emotions or opinions through natural language understanding/processing models. The sentiment analysis of social media posts has been extensively studied by researchers from both academia and corporate, who put it into real-world applications such as market emotion indicators and customer feedback. Out of all platforms, the researchers use English tweets as training data very frequently. A small portion of them has also stepped foot into analyzing the emotion of social media posts in other languages.

As industries use language understanding models more pervasively, sentiment analysis tasks are carried out using entirely different groups of models. Previously, the sentiment analysis researcher got benefited from the development of the word vectors model in non-English monolingual sentiment analysis, including word2vec(Xue et al., 2014). The word2vec model by Google(Mikolov et al., 2013), specifically, creates word embeddings that reflect the words' semantics and correlations, which can be used for sentiment analysis. In the more recent years, transformer-based pre-trained models like BERT, RoBERTa, and XLM-RoBERTa emerged as the amount of training data and computing speed increased. XLM-Twitter, built by (Barbieri et al., 2021) is one of the recent examples of twitter-based sentiment analysis fine-tuned models.

Though versatile, the BERT model family requires further fine-tuning to be fitted into specific tasks. For multi-lingual sentimental analysis, (Barriere and Balahur, 2020) proposes to use data augmentation–auto translating English tweets to other languages–to create more training data for languages with insufficient training data (and data in resource-limiting languages). Though the performance of RoBERTa in resource-limited-language texts has been improved, data augmenta-

[3] https://commoncrawl.org/the-data/get-started

tion is still very effective during the pre-training process. Another paper(Pota et al., 2021) discusses the pre-processing techniques for BERT sentiment analysis based on social media posts. A lot of the papers had done modular tests to analyze the performance of different pre-processing methods, including tackling specific characteristics of social media posts (e.g. emoticons, emojis, mentions, hashtags, and URL), and other classic NLP pre-processing steps (e.g. removing stopwords, numbers, and punctuations). This paper seeks to get a better understanding of improving the fine-tuning and pre-processing effectiveness when doing sentiment analysis with RoBERTa, which is relatively new.

3 Method

3.1 Data Source

We collected pre-labeled social media sentiment analysis datasets in five different languages: Chinese, Spanish, English, Marathi, and Urdu. For Chinese, we use the weibo_senti_100k dataset [4] which contains prelabeled Weibo comments. The data was labeled in two categories, positive and negative. The English corpus is the largest, which is named Sentiment140 [5] and contains 1.6 million tweets extracted with Twitter API. The Spanish dataset [6] contains 7,867 airline tweets in 2018 and was used for the Kaggle competition in the past. We also use the Urdu Sentiment Corpus(Khan and Nizami, 2020), which contains 17,185 tweets in Urdu labeled as positive, negative, or neutral. For Marathi, we use the L3-Cube-MahaSent corpus(Kulkarni et al., 2021), containing 16,000 tweets labeled in a similar manner.

3.2 Data Preprocessing

To begin with, we discovered that each dataset is labeled differently. To ensure consistency, we decide to unify the label by setting 1 as positive sentiment and 0 as negative sentiment. Also, since the two of our largest datasets - Chinese and English - do not contain neutral sentiment, we decide to remove rows with neutral sentiment in the Spanish, Urdu, and Marathi datasets. Due to the limited capacity of our model, we decided to cut the size of the Chinese and English datasets by sampling 9,000 rows from each of them. Then, we split our data into three sets: test set(15%), validation set(15%), and training set (70%). After the previous preprocessing, we drew the following summaries on our dataset.

Test Set						
	C	E	M	S	R	T
P	697	699	585	241	74	2296
N	653	651	626	547	72	2549
T	1350	1350	1211	788	146	4845
Validation Set						
	C	E	M	S	U	T
P	228	659	675	605	72	2239
N	560	691	675	606	74	2606
T	788	1350	1350	1211	146	4845
Training Set						
	C	E	M	S	U	T
P	3162	3174	1020	334	2848	10538
N	3138	3126	2662	353	2806	12085
T	6300	6300	3682	687	5654	22623

Table 1

3.3 Baseline Result

For the baseline result, we input our test set directly into the XLM RoBERTa for Sequence Classification Model. This model is pre-train to perform text classification tasks. However, it hasn't been fine-tuned. We ran a separate test for each of the five languages. Table 2 shows the result of our baseline test.

Lang	C	E	U	M	S
Acc	0.51	0.50	0.51	0.50	0.71

Table 2

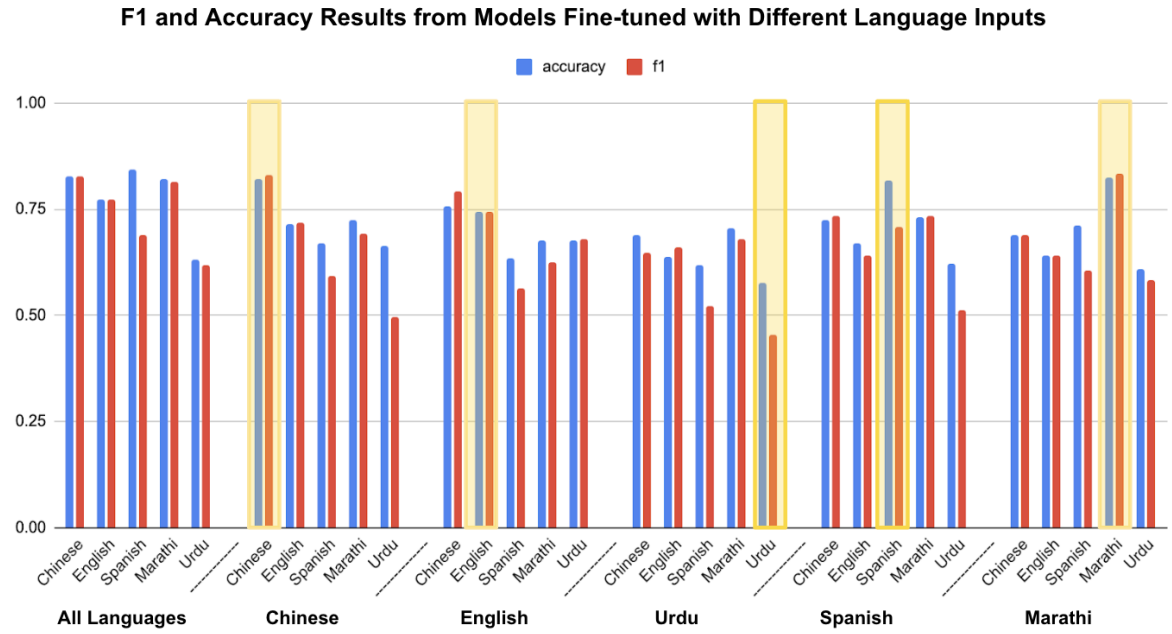
3.4 Optimizing Hyperparameters

We proposed the following options of hyperparameters: batch size [32, 64]; learning rate [1e^5, 2e^5, 3e^5]. To find the most optimal hyperparameters for our fine-tuning process, we ran 6 trails with each hyperparameter combination. We recorded the validation loss, validation accuracy and validation F1 score for each epoch. We observed these results and the improvement of each

[4] https://github.com/SophonPlus/ChineseNlpCorpus/blob/master/datasets/weibo_senti_100k/intro.ipynb

[5] https://www.kaggle.com/datasets/kazanova/sentiment140

[6] https://www.kaggle.com/competitions/spanish-arilines-tweets-sentiment-analysis/data



metric over each epoch. Table 3 shows that the trail with batch size 64 and learning rate 3e-5 produced the most optimal output. Therefore, we decide to use this hyperparameter combination for our fine-tuning process.

Note: We followed a tutorial from Kaggle[7] for our model setup and fine-tuning process.

4 Result

We fine-tuned our model with batch size 64 and learning rate 3e-5 in two major directions. First, we use the test and validation set for the combined language. Then, we created five different models, each trained with the validation set of one single language. With the six models, we seek to test the test set in every language and evaluate each language's performance using accuracy, F1, precision, and recall. In this way, we can gain insights into the cross-lingual transferability of the XLM-RoBERTa model. The result of all six models can be found at the graph above and table 4.

5 Analysis and Discussion

To test the cross-lingual transferability of the pre-trained models, we compare the resulting f1 scores from the fine-tuned models. Assuming the cross-lingual transferability is weak, one should see less accuracy when the input test data are in a different language from the fine-tuned layer. At the same time, models fine-tuned in English and Spanish may result in higher scores in the model fine-tuned on each other's, since English and Spanish are similar regarding grammar, letters, and language family. However, we found that the outputs have a negative relationship. This might be due to the fact that the Spanish dataset has much more negative data input than the positive ones (around 70% of data are classified as having negative sentiment). This relationship proves that the improvement in performance may not be based on the similarity in the languages themselves.

The five models that were fine-tuned with one single language demonstrated a surprisingly good performance in performing sentiment analysis tasks on the languages that the model wasn't trained on. The overall performance is relatively uniform in every model. The only exception is that F1 and accuracy of Urdu were relatively low compared to other languages. However, this pattern can be found in all six models, including the one that was solely trained in Urdu. Therefore, we can suspect that the poor performance may be due to the small size of the Urdu dataset, as this is a common factor for all models. Thus, we can infer that the improvements in f1 scores could be transferable across the five languages we selected.

[7] https://www.kaggle.com/code/vbookshelf/basics-of-bert-and-xlm-roberta-pytorch/notebook#Section-3

6 Collaboration Statement

Overall, we plan to divide work responsibilities into 50% and 50%, each of us taking part in every step of the research process. For the "Related Works" section, Zhang was responsible for learning and understanding the milestones and important models for multilingual language understanding and writing the literature review for section 2.1. Wang was responsible for researching past studies on social media sentiment analysis and conducting the literature review for section 2.2. Each of us found 2-3 datasets for this research project. Wang found the Spanish, English, and Chinese datasets, while Zhang discovers two sentiment analysis corpora in lower-resource languages - Marathi and Urdu. Each of us work on the pre-processing of 2-3 data set, and collaborate during the hyperparameter optimization and fine-tuning procedure. Regarding the analysis and discussion section, the language-specific analysis is written by Wang, and the performance across 5 models were run and evaluated by Zhang.

References

Abdullah Alsaeedi and Mohammad Zubair Khan. 2019. A study on sentiment analysis techniques of twitter data. *International Journal of Advanced Computer Science and Applications* 10(2):361–374.

Francesco Barbieri, Luis Espinosa Anke, and Jose Camacho-Collados. 2021. Xlm-t: A multilingual language model toolkit for twitter. *arXiv preprint arXiv:2104.12250* .

Valentin Barriere and Alexandra Balahur. 2020. Improving sentiment analysis over non-english tweets using multilingual transformers and automatic translation for data-augmentation. *arXiv preprint arXiv:2010.03486* .

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116* .

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* .

Muhammad Yaseen Khan and Muhammad Suffian Nizami. 2020. Urdu sentiment corpus (v1.0): Linguistic exploration and visualization of labeled datasetfor urdu sentiment analysis. In *2020 IEEE 2nd International Conference On Information Science Communication Technology (ICISCT)*. IEEE.

Atharva Kulkarni, Meet Mandhane, Manali Likhitkar, Gayatri Kshirsagar, and Raviraj Joshi. 2021. L3cubemahasent: A marathi tweet-based sentiment analysis dataset. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. pages 213–220.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* .

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* .

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual bert? *arXiv preprint arXiv:1906.01502* .

Marco Pota, Mirko Ventura, Hamido Fujita, and Massimo Esposito. 2021. Multilingual evaluation of pre-processing for bert-based sentiment analysis of tweets. *Expert Systems with Applications* 181:115119.

Zihan Wang, Stephen Mayhew, Dan Roth, et al. 2019. Cross-lingual ability of multilingual bert: An empirical study. *arXiv preprint arXiv:1912.07840* .

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2019. Ccnet: Extracting high quality monolingual datasets from web crawl data. *arXiv preprint arXiv:1911.00359* .

Shijie Wu and Mark Dredze. 2020. Are all languages created equal in multilingual bert? *arXiv preprint arXiv:2005.09093* .

Bai Xue, Chen Fu, and Zhan Shaobin. 2014. A study on sentiment computing and classification of sina weibo with word2vec. In *2014 IEEE International Congress on Big Data*. IEEE, pages 358–363.

	Trial 1	Trial 2	Trial 3	Trial 4	Trial 5	Trial 6
Batch	32	32	32	64	64	64
Rate	1e-5	2e5	3e5	1e5	2e5	3e5
Epoch	3.00	3.00	3.00	3.00	3.00	3.00
E1 Validation Loss	69.40	82.58	68.71	34.70	33.95	35.09
E1 Validation Accuracy	0.80	0.79	0.80	0.79	0.79	0.79
E1 Validation F1	0.79	0.78	0.78	0.78	0.77	0.78
E2 Validation Loss	68.58	101.74	65.30	34.69	32.90	34.45
E2 Validation Accuracy	0.79	0.79	0.80	0.80	0.80	0.80
E2 Validation F1	0.79	0.79	0.79	0.77	0.80	0.79
E3 Validation Loss	69.94	87.96	74.00	32.91	33.86	35.72
E3 Validation Accuracy	0.80	0.79	0.81	0.80	0.80	0.81
E3 Validation F1	0.78	0.77	0.79	0.77	0.78	0.80

Table 3: Optimizing Hyperparameters

All Languages Model						Urdu Model				
Language	accuracy	f1	precision	recall		Language	accuracy	f1	precision	recall
Chinese	0.83	0.83	0.79	0.87		Chinese	0.69	0.65	0.73	0.58
English	0.77	0.77	0.76	0.79		English	0.64	0.66	0.62	0.70
Spanish	0.84	0.69	0.75	0.64		Spanish	0.62	0.52	0.41	0.72
Marathi	0.82	0.81	0.84	0.79		Marathi	0.71	0.68	0.75	0.62
Urdu	0.63	0.62	0.69	0.56		Urdu	0.58	0.46	0.62	0.36
Chinese Model						Spanish Model				
Chinese	0.82	0.83	0.82	0.85		Chinese	0.73	0.73	0.70	0.78
English	0.71	0.72	0.75	0.69		English	0.67	0.64	0.70	0.59
Spanish	0.67	0.59	0.46	0.84		Spanish	0.82	0.71	0.66	0.77
Marathi	0.73	0.69	0.77	0.63		Marathi	0.73	0.73	0.72	0.75
Urdu	0.66	0.49	0.75	0.37		Urdu	0.62	0.51	0.71	0.40
English Model						Marathi Model				
Chinese	0.76	0.79	0.71	0.90		Chinese	0.69	0.69	0.67	0.71
English	0.74	0.74	0.79	0.70		English	0.64	0.64	0.64	0.64
Spanish	0.63	0.56	0.43	0.83		Spanish	0.71	0.61	0.50	0.77
Marathi	0.68	0.62	0.65	0.60		Marathi	0.82	0.83	0.78	0.89
Urdu	0.68	0.68	0.67	0.69		Urdu	0.61	0.58	0.62	0.56

Table 4: F1 and Accuracy on the Six Models