CS410 Tech Review: XLM-T Language Model

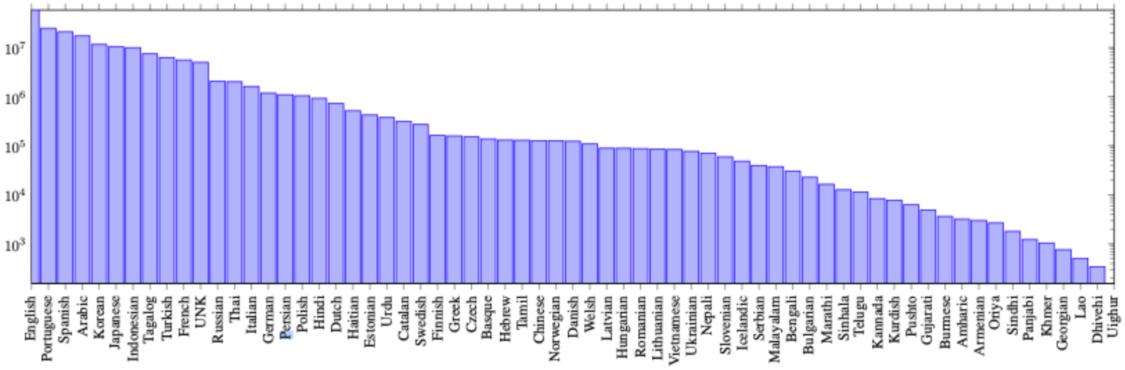
Xin Jin (xinj3)

Introduction

Language models (LM) are one of the most important aspects in NLP researches. It can often reveal the characteristics of the textual contents and is fundamental for various text retrieval and text mining tasks. However, this is often under the assumption that the texts are written in one single language, and that is far away from real world scenario. Nowadays, there are hundreds of languages that are being frequently used in the world. Social media platforms allow users with different language background to share textual information altogether. This becomes a challenge if the LM only focus on one single language. Naturally, multilingual LM becomes necessary for tasks involving multiple languages, which makes multilingual textual data processing possible, without breaking it down into monolingual single tasks. The multilingual LM can often be built based of a well-known monolingual model, such as XLM-R (Conneau et al., 2020) was built based of RoBERTa (Liu et al., 2019), and mBERT was built based of BERT (Devlin et al., 2019). This review is going to cover XLM-T (Barbieri et al., 2022), one of the latest multilingual language models released in May, 2022. It was built based on the state-of-the-art XLM-R language model and was fine-tuned specifically for the Tweet messages.

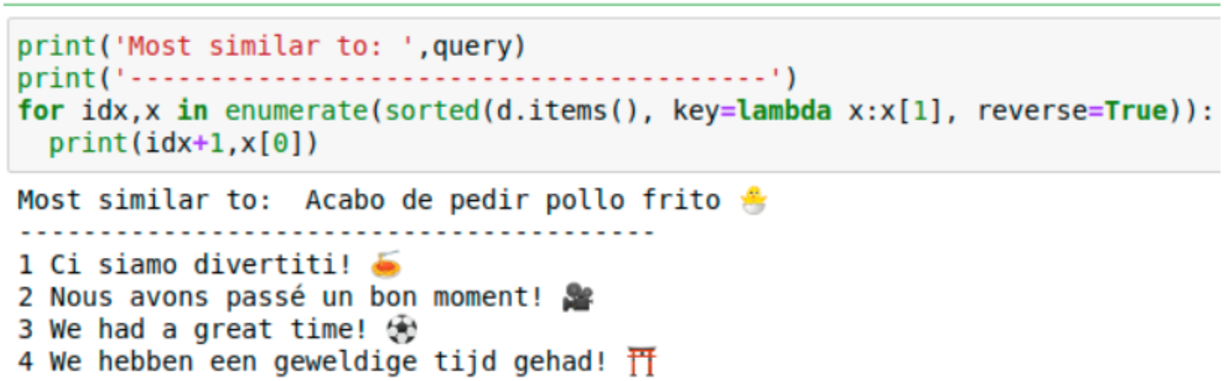
XLM-T Language Model

The dataset for XLM-T was retrieved from Twitter directly which consists of 198M tweet messages across more than 50 different languages (Barbieri et al., 2022). The figure below shows the distribution of the languages from the dataset ranked by frequency. The top five most popular languages are: English, Portuguese, Spanish, Arabic, and Korean. The language model was trained based on the existing XLM-R checkpoint until converging on tweet dataset.



Feature Extraction

The feature extraction was done by the tokenizing the given text message and using the XLM-T model to get the embedding. The embedded features of the tweet message can then be sure to find similar messages across different languages. This can be useful for tweet messages categorization and retrieval. The user can categorize the tweet messages based on embedded features. Or, classifying the tweet messages based on given presentative messages in each category. For text retrieval, the user can embed the query message and use it to find the most relevant tweet messages. The user can even use it as a ranking system based on the cosine similarity between the query message and the tweet dataset. Notice from the example below that there is no constraint on the language types.

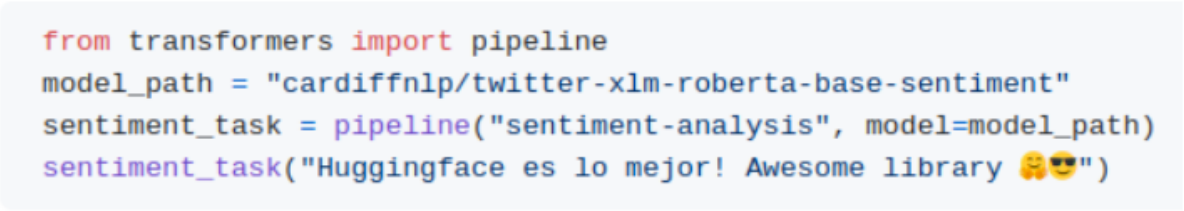


```
print('Most similar to: ',query)
print('-------------------------------------')
for idx,x in enumerate(sorted(d.items(), key=lambda x:x[1], reverse=True)):
  print(idx+1,x[0])
```

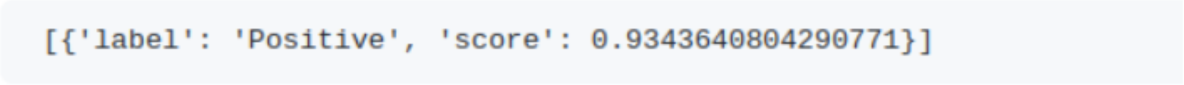
```
Most similar to:  Acabo de pedir pollo frito 🐣
---------------------------------------
1 Ci siamo divertiti! 🍝
2 Nous avons passé un bon moment! 🎥
3 We had a great time! ⚽
4 We hebben een geweldige tijd gehad! ⛩
```

Tweet Classification and Inference

XLM-T also provides the possibility for more complex task such as sentiment analysis and classification. Sentiment analysis and classification often require feature extractions on input texts, and a machine learning model is generally used to accomplish the sentiment classification job. The XLM-T shows the power of extracting features from tweet messages. The extracted features can then be used as the inputs along with the labels for the classification task. The provided example below shows that the using XLM-T for feature extraction can effectively be used for sentiment analysis.



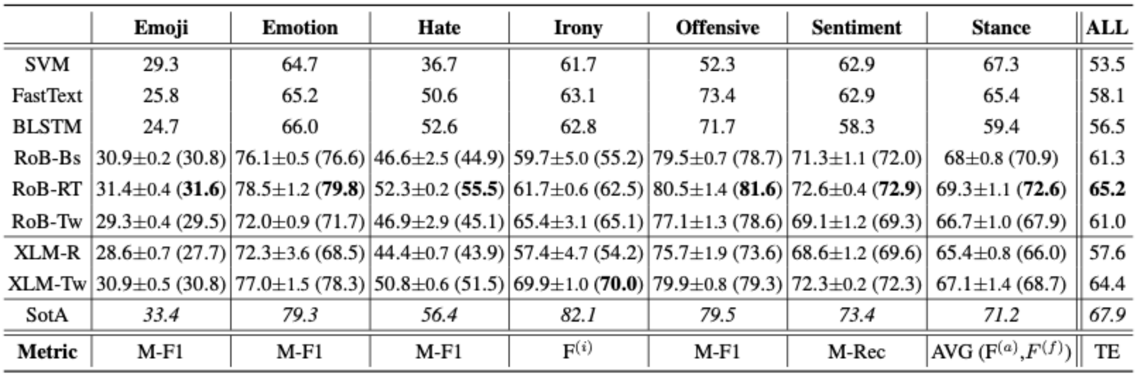
```
from transformers import pipeline
model_path = "cardiffnlp/twitter-xlm-roberta-base-sentiment"
sentiment_task = pipeline("sentiment-analysis", model=model_path)
sentiment_task("Huggingface es lo mejor! Awesome library 🤗😎")
```



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

Evaluation

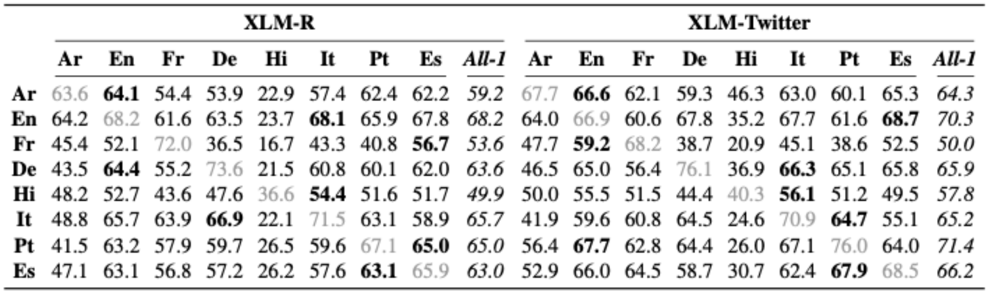
The true advantage of XLM-T language model is its powerful multilingual feature extraction ability on tweet messages. This was evaluated using the TweetEval benchmark (Barbieri et al., 2020), which is consisted of 7 classification tasks. The table below shows the comparison results. The XLM-T model outperforms the XLM-R model on every category, and it also outperforms the baseline RoBERTa and RoBERTa-Twitter (which is fine tune on tweet dataset). However, XLM-T is underperformed by the English monolingual RoBERTa model fine-tuned on Twitter dataset.



	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.

Furthermore, XLM-T also has outstanding performance on zero shot sentiment analysis task. The task was given 8 most popular languages. The zero shot was done by transferring one language to another and use the other language to do the sentiment analysis. The table below shows the zero shot cross lingual sentiment analysis results. Again, XLM-T outperforms XLM-R on every single language.



	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.

Conclusion

This review presents the XLM-T model and explores its downstream applications and evaluated on its performance on various multilingual NLP tasks. As a language model, XLM-T was built based of the XLM-R language model by specifically fine-tuning on the tweet dataset. It fills in a gap for tweet message specifically related multilingual NLP tasks. It can be used as the backbone model for text mining and text retrieval tasks. The experiments showed effective performance on cross-lingual and multilingual feature extractions. It also outperforms most of the state-of-the-art model on tweet message dataset and evaluation metric. Thus, XLM-T language model shows a lot of potentials, and more discussion and explorations on downstream tasks can be expected in the future.

Reference

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

Barbieri, F., Anke, L. E., & Camacho-Collados, J. (2022, June). Xlm-t: Multilingual language models in twitter for sentiment analysis and beyond. In Proceedings of the Thirteenth Language Resources and Evaluation Conference (pp. 258-266).

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmá n, F., Grave, E., Ott, M., Zettle- moyer, L., and Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. In Pro- ceedings of the 58th Annual Meeting of the Associa- tion for Computational Linguistics, pages 8440–8451, Online, July. Association for Computational Linguis- tics.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceed- ings of the 2019 Conference of the North American Chapter of the Association for Computational Lin- guistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Min- neapolis, Minnesota, June. Association for Computa- tional 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 pre- training approach. arXiv preprint arXiv:1907.11692.