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

Francesco Barbieri♦, Luis Espinosa-Anke♦, Jose Camacho-Collados♦
Snap Inc., ♦Cardiff NLP, Cardiff University

LREC 2022

Motivation & Background

Specializing Language Models
• augmenting with external information
• pretraining on domain-specific corpora

LMs for Twitter and English
• BerTweet (Nguyen et al., 2020), model and TweetEval (Barbieri et al., 2020), model and unified benchmark for tweet classification.
• However, not a similar set of models and frameworks for multilingual tweet classification.

XLM-T: The Model
• 198M tweets between May 2018 and March 2020
  ○ At least three tokens
  ○ No URLs
  ○ No language filtering
• Continue pretraining an XLM-R model
  ○ Start from a publicly available checkpoint
  ○ same MLM objective
  ○ pretrain until convergence in a validation set.
• About 14 days on 8 NVIDIA V100 GPUs

Sentiment Analysis Experiments

X-Lingual, zero-shot experiments

With target language training data

FINDINGS
• More training data is better, even if different languages
  ○ Better off XLM-T than XLM-R
  ○ However, smart language selection is better in ~50% of cases
  ○ Typological proximity?
  ○ Obvious trade-offs

Releases

Language Models at
https://huggingface.co/cardiffnlp
• Multilingual Twitter-specific Language Model
  ○ https://huggingface.co/cardiffnlp/twitter-xlm-roberta-base
  ○ The above, fine-tuned on sentiment analysis
  ○ https://huggingface.co/cardiffnlp/twitter-xlm-roberta-base-sentiment

Repository
https://github.com/cardiffnlp/xlm-t
• Fine-tuning interface
  ○ LM fine-tuning
  ○ Adapters!
• Notebooks with starter code
  ○ End-to-end pipeline on UMSAB
  ○ Download, predict, evaluate
  ○ Extract embeddings from tweets
  ○ Sentiment prediction
  ○ Fine-tuning on custom data

Leadership

Details:
• Pruned all datasets to the smallest language (Hindi)
• Dataset size: 24,263 tweets (3,033 per language)

UMSAB: The Dataset

Similarity with English

Findings

Repository
https://github.com/cardiffnlp/xlm-t
• Fine-tuning interface
  ○ LM fine-tuning
  ○ Adapters!
• Notebooks with starter code
  ○ End-to-end pipeline on UMSAB
  ○ Download, predict, evaluate
  ○ Extract embeddings from tweets
  ○ Sentiment prediction
  ○ Fine-tuning on custom data

Leadership

Details:
• Pruned all datasets to the smallest language (Hindi)
• Dataset size: 24,263 tweets (3,033 per language)

UMSAB: The Dataset

Similarity with English

Findings

Repository
https://github.com/cardiffnlp/xlm-t
• Fine-tuning interface
  ○ LM fine-tuning
  ○ Adapters!
• Notebooks with starter code
  ○ End-to-end pipeline on UMSAB
  ○ Download, predict, evaluate
  ○ Extract embeddings from tweets
  ○ Sentiment prediction
  ○ Fine-tuning on custom data

Leadership

Details:
• Pruned all datasets to the smallest language (Hindi)
• Dataset size: 24,263 tweets (3,033 per language)

UMSAB: The Dataset

Similarity with English

Findings

Repository
https://github.com/cardiffnlp/xlm-t
• Fine-tuning interface
  ○ LM fine-tuning
  ○ Adapters!
• Notebooks with starter code
  ○ End-to-end pipeline on UMSAB
  ○ Download, predict, evaluate
  ○ Extract embeddings from tweets
  ○ Sentiment prediction
  ○ Fine-tuning on custom data

Leadership

Details:
• Pruned all datasets to the smallest language (Hindi)
• Dataset size: 24,263 tweets (3,033 per language)