Bernice: A Multilingual Pre-trained Encoder for Twitter
Alexandra DeLucia,1 Shijie Wu,1 Aaron Mueller,1 Carlos Aguirre,1 Mark Dredze,1 and Philip Resnik2
1 Center for Language and Speech Processing, Johns Hopkins University
2 Linguistics/UMIACS, University of Maryland

Motivation

• A multilingual model tailored for Twitter
• Existing models are either not multilingual, do not use a Twitter-specific tokenizer, or are only secondarily trained on tweets
• We introduce Bernice®, the first multilingual model trained exclusively on tweets with a Twitter-trained tokenizer

*Named after Bert’s pet pigeon

Architecture

• BERT Base with 270M parameters
• 128 maximum sequence length, which covers 99.96% of tweets

Training Data

• 2.5 billion tweets in 66 languages from 1% public stream from Jan 2016–Dec 2021
• 56 billion subwords
• Two pre-training datasets with different distributions over language. Both are sampled with exponential resampling with α=0.5

1. Presampled: language distribution of the full dataset. Resampled on-the-fly during training.
2. Language-sampled: static exponentially-sampled dataset with high exposure to low-resource languages

Tokens and Tokens per Language in Pre-training Data

<table>
<thead>
<tr>
<th>Language</th>
<th>Tweets</th>
<th>Tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arabic</td>
<td>1,000</td>
<td>900</td>
</tr>
<tr>
<td>English</td>
<td>1,000</td>
<td>900</td>
</tr>
<tr>
<td>French</td>
<td>1,000</td>
<td>900</td>
</tr>
<tr>
<td>German</td>
<td>1,000</td>
<td>900</td>
</tr>
<tr>
<td>Italian</td>
<td>1,000</td>
<td>900</td>
</tr>
<tr>
<td>Portuguese</td>
<td>1,000</td>
<td>900</td>
</tr>
<tr>
<td>Spanish</td>
<td>1,000</td>
<td>900</td>
</tr>
</tbody>
</table>

2. Unified Multilingual Sentiment Analysis

• 8 individual language datasets
• Label tweet as positive, negative, or neutral
• Reported score is Macro-F1

3. Multilingual Hate Speech

• Combined 16 datasets across 9 languages
• Label tweets as hate speech or normal
• Reported score is Macro-F1

Tokenizer Analysis

• Compare coverage of non-Twitter tokenizer (XLM-R) to custom tokenizer
• Bernice tokenizer has better coverage of Twitter-specific vocabulary, hashtags, and emoji

Pre-training

• RoBERTa masked language modeling (MLM) objective
• AWS EC2 with a p3.16xlarge instance with 8 NVIDIA Tesla V100 GPUs
• 405K training steps over 330 hours

Evaluation

• Bernice consistently outperforms XLM-R, a model without any Twitter pre-training
• Also, Bernice consistently performs on-par with models that have seen significantly more data, including TwiHIN-BERT models, which have seen 78 tweets

Benchmarks

• We compare Bernice to competitive Twitter language models on 3 benchmarks
• All models are fine-tuned on each benchmark task after performing a hyperparameter search

1. TweetEval

• English-only
• 7 Twitter-specific tasks
• Reported score is Macro-F1 for all except sentiment, which is Macro-Recall

2. Unified Multilingual Sentiment Analysis

• English
• French
• German
• Italian
• Portuguese
• Spanish
• All

3. Multilingual Hate Speech

• English
• French
• German
• Italian
• Portuguese
• Spanish
• All

Tokenizer

• A Twitter-specific tokenizer (250K vocabulary) using unigram SentencePiece model trained on language-sampled 78M tweets
• Replace user mentions and URLs with special symbols @USER and HTTPURL

Try our Model

https://huggingface.co/jhu-clsp/bernice