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 Resnik[2]

[1] *Center for Language and Speech Processing, Johns Hopkins University*

[2] *Linguistics/UMIACS, University of Maryland*

Model Design

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

	Multilingual	Twitter pre-training only	Twitter tokenizer
BERTweet	🚩	✅	✅
XLM-R	✅	🚩	🚩
XLM-T	✅	🚩	🚩
TwHIN-BERT*	✅	✅	🚩
Bernice	✅	✅	✅

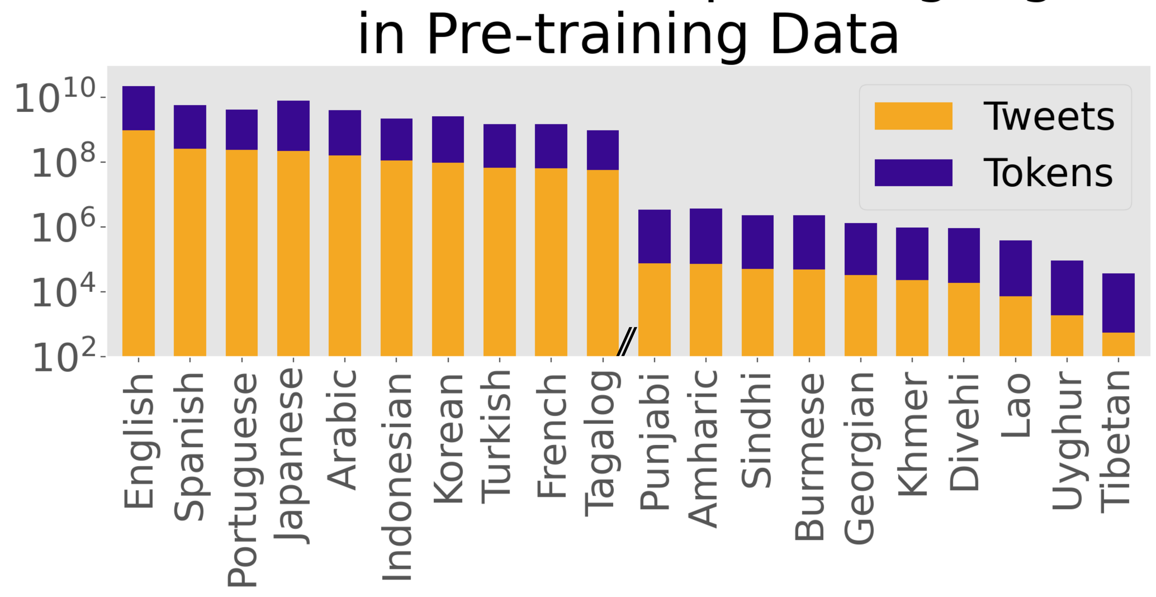
Architecture

- $\text{BERT}_{\text{Base}}$ 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

Tweets and Tokens per Language in Pre-training Data



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

	Tokenizer	Subwords	Subword length
Tweets	Bernice	26.27 (25.48)	2.95 (1.98)
	XLM-R	27.86 (25.69)	2.78 (1.85)
Hashtags	Bernice	5.0 (2.0)	3.0 (1.9)
	XLM-R	6.5 (2.5)	2.3 (1.1)

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

Summary

- **Bernice** is a $\text{BERT}_{\text{Base}}$ model customized for multilingual tweet representation
  - Trained on 2.5B tweets
  - Custom tokenizer for hashtag, emoji, and slang
- Benchmark performance is **better** than non-Twitter models and **on-par** with Twitter models, but trained on **significantly less data**
- Domain-specific models are **more efficient** to train overall than domain-adapted models

Try our Model

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

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 TwHIN-BERT models, which have seen 7B 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

	Emoji	Emotion	Hate	Irony	Offensive	Sentiment	Stance	All (TE)
BERTweet	**33.4**	**79.3**	**56.4**	**82.1**	79.5	**73.4**	**71.2**	**67.9**
RoBERTa-RT	31.4	78.5	52.3	61.7	80.5	72.8	69.3	65.2
RoBERTa-Tw	29.3	72.0	49.9	65.4	77.1	69.1	66.7	61.4
XLM-R	28.6	72.3	44.4	57.4	75.7	68.6	65.4	57.6
XLM-T	30.9	77.0	50.8	69.9	79.9	72.3	67.1	64.4
TwHIN-BERT-MLM	30.5	**79.3**	50.5	71.6	80.0	72.5	69.4	64.8
TwHIN-BERT	30.5	77.5	45.6	69.1	79.1	72.8	67.3	63.1
Bernice	31.2	78.3	50.2	71.5	**81.0**	73.3	68.2	64.8

2. Unified Multilingual Sentiment Analysis
   - 8 individual language datasets
   - Label tweet as positive, negative, or neutral
   - Reported score is Macro-F1

	Bernice	XLM-T	XLM-R	TwHIN-MLM	TwHIN
Arabic	**65.77**	64.99	64.99	65.19	65.15
English	68.05	68.01	66.38	**70.36**	69.53
French	72.39	70.67	**72.46**	68.57	70.78
German	**77.21**	74.70	75.07	74.56	72.80
Hindi	**59.14**	56.38	47.86	55.34	53.09
Italian	**72.82**	66.49	68.89	68.79	68.38
Portuguese	**77.86**	73.71	72.37	74.78	74.64
Spanish	**69.48**	66.73	65.87	67.19	65.85
All	**70.34**	67.71	66.74	68.10	67.53

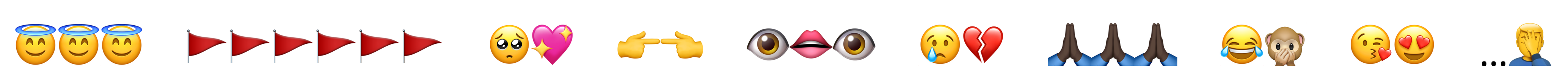
3. Multilingual Hate Speech
   - Combined 16 datasets across 9 languages
   - Label tweets as hate speech or normal
   - Reported score is Macro-F1

	Bernice	XLM-T	XLM-R	TwHIN-BERT-MLM	TwHIN-BERT
Arabic	86.67	**88.25**	83.29	86.75	87.99
English	82.24	82.02	81.49	82.18	**82.26**
French	69.51	**69.87**	68.24	67.58	62.32
German	**85.98**	71.97	71.97	70.16	81.03
Indonesian	**89.82**	87.39	87.44	86.78	89.21
Italian	66.90	**69.32**	67.99	66.84	64.50
Polish	48.99	48.96	48.99	48.96	48.99
Portugese	**73.11**	70.54	70.01	70.79	70.56
Spanish	**82.56**	82.54	80.38	80.62	82.04
All	**76.20**	74.54	73.31	73.41	74.32

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

Select emoji in Bernice vocabulary



Hashtag	Bernice	XLM-R
#DahmerNetflix	['D', 'ah', 'mer', 'Netflix']	['D', 'ah', 'mer', 'Net', 'flix']
#AsiaCup2023	['Asia', 'Cup', '2023']	['Asia', 'C', 'up', '20', '23']
#BLEACH_anime	['BLEACH', '_', 'anime']	['BLE', 'ACH', '_', 'an', 'ime']
#ToriesDestroyingOurCountry	['Tories', 'Destroying', 'Our', 'Country']	['To', 'ries', 'D', 'estro', 'ying', 'O', 'ur', 'Count', 'ry']
#MarriedAtFirstSight	['Married', 'At', 'First', 'Sight']	['Mar', 'ried', 'At', 'First', 'S', 'ight']
#NoGOPAbortionBans	['No', 'GOP', 'Abortion', 'Ban', 's']	['No', 'G', 'OPA', 'bor', 'tion', 'Ban', 's']
#SaudiNationalDay	['Saudi', 'National', 'Day']	['S', 'audi', 'National', 'Day']
#PakvsEngland	['Pak', 'vs', 'England']	['Pak', 'vs', 'Eng', 'land']
#pakvsengland	['pak', 'vs', 'england']	['pak', 'v', 'seng', 'land']
#DiaMundialDelTurismo	['Dia', 'Mundial', 'Del', 'Turismo']	['Dia', 'M', 'undi', 'al', 'Del', 'Tur', 'ismo']
#buenmiercoles	['buen', 'miercoles']	['bu', 'en', 'mier', 'cole', 's']
#يوم_المعلم	['يوم', '_', 'المعلم']	['يوم', '_', 'الم', 'علم']
#DraftKingsTNF	['Draft', 'Kings', 'TN', 'F']	['D', 'raf', 't', 'K', 'ings', 'TN', 'F']