Bernice: A Multilingual Pre-trained Encoder for Twitter

Alexandra DeLucia, Shijie Wu,* **Aaron Mueller, Carlos Aguirre,** and **Mark Dredze**
Center for Language and Speech Processing
Johns Hopkins University
{aadelucia,shijie.wu,amueller,caguirr4,mdredze}@jhu.edu

Philip Resnik
Linguistics/UMIACS
University of Maryland
resnik@umd.edu

Abstract

The language of Twitter differs significantly from that of other domains commonly included in large language model training. While tweets are typically multilingual and contain informal language, including emoji and hashtags, most pre-trained language models for Twitter are either monolingual, adapted from other domains rather than trained exclusively on Twitter, or are trained on a limited amount of in-domain Twitter data. We introduce Bernice, the first multilingual RoBERTa language model trained from scratch on 2.5 billion tweets with a custom tweet-focused tokenizer. We evaluate on a variety of monolingual and multilingual Twitter benchmarks, finding that our model consistently exceeds or matches the performance of a variety of models adapted to social media data as well as strong multilingual baselines, despite being trained on less data overall. We posit that it is more efficient compute- and data-wise to train completely on in-domain data with a specialized domain-specific tokenizer.

https://github.com/JHU-CLSP/Bernice-Twitter-encoder

	Multilingual	Twitter pre-training only	Twitter tokenizer
BERTweet	✗	✓	✓
XLM-R	✓	✗	✗
XLM-T	✓	✗	✗
TwHIN-BERT*	✓	✓	✗
Bernice	✓	✓	✓

Table 1: A comparison of commonly used models for tweet representation in downstream tasks. Bernice is the first multilingual model trained from scratch on Twitter data with a custom tokenizer.

1 Introduction

BERT (Devlin et al., 2019) has become a core part of modern natural language processing (NLP) pipelines. Performance on downstream tasks depends on the quality of the included BERT-style model, driven in large part by model size, vocabulary, and relevance of pre-training data. Studies have shown that pre-training data should match the intended deployment domain to achieve optimal task performance (Barbieri et al., 2021; Nguyen et al., 2020; Huang et al., 2020a; Gururangan et al., 2020; Gu et al., 2021; Zhang et al., 2022).

One of the most commonly studied domains for NLP research is Twitter, which has led to numerous downstream applications. The unique nature of Twitter data has prompted specialized applications, including sentiment prediction around elections and vaccines (Broniatowski et al., 2018), stance prediction on gun control (Benton and Dredze, 2018), civil unrest forecasting (Chinta et al., 2021; Sech et al., 2020; Alsaedi et al., 2017), and various mental health applications (Harrigian et al., 2021).

While several different BERT-style models have been developed for Twitter (Barbieri et al., 2021, 2020; Nguyen et al., 2020; Zhang et al., 2022), compared to BERT models for other domains, these Twitter-based models are smaller, use much less training data, and are often adapted from other domains instead of for Twitter initially. This is especially problematic for the tokenizer, creating a mismatch that limits its efficacy on representing Twitter data. Furthermore, these models are primarily for English, despite the depth of multilingual work on Twitter.

We introduce Bernice, the first multilingual pre-trained encoder trained exclusively on Twitter data with a custom tokenizer.[1] Bernice uses more Twitter data than most BERT style models to date (2.5 billion tweets)[2] and utilizes a tokenizer trained exclusively on Twitter data.

We evaluate Bernice on monolingual and mul-

*Currently at Bloomberg. Work performed while at Johns Hopkins.

[1] Bernice is the name of Bert's pet pigeon on Sesame Street.

[2] TwHIN-BERT (Zhang et al., 2022), released subsequent to our submission, uses 7B tweets, but does not use a custom Twitter tokenizer.

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 6191 - 6205
December 7-11, 2022

tilingual Twitter benchmarks: TweetEval (Barbieri et al., 2020), Unified Multilingual Sentiment Analysis Benchmark (UMSAB) (Barbieri et al., 2021), and Multilingual Hate Speech (Aluru et al., 2021). We also evaluate Bernice's coverage on low-resource and Twitter-specific language (e.g., hashtags) through an analysis of our tokenizer. We show that across these tasks, Bernice does better than comparable models despite having much less data overall.[3] We discuss our results' implications for training in-domain models from scratch versus adapting out-of-domain models.

2 Bernice

2.1 Pre-training Data

Bernice is trained only on Twitter data, collected from the 1% public Twitter stream via the Twitter API between January 2016 and December 2021. We filtered the data by removing tweets with less than three tokens, excluding user mentions and URLs. We follow Nguyen et al. (2020) and replace user mentions and URLs with special symbols `@USER` and `HTTPURL`, respectively, but do not convert emoji to descriptive text.

Our dataset contained 2.5 billion tweets with 56 billion subwords in 66 languages. While more languages have been identified on Twitter (Scannell, 2022a), the Twitter API only supports identification of 66 languages, including an "undefined" category. The most common languages are English and Spanish, and least common are Tibetan and Uyghur (see Appendix Figure 1). We created two training datasets to encourage exposure to low-resource languages: (1) **Presampled** and (2) **Language-sampled**. **Presampled** reflects the language distribution of the full dataset.

Language-sampled samples the data to increase the prevalence of less common languages, which has been shown to benefit multilingual model training (Lample and Conneau, 2019; Chi et al., 2021; Barbieri et al., 2021; Xue et al., 2021a). For language-specific sampling, we use exponential resampling (Lample and Conneau, 2019): $p_\ell \propto (\frac{n_\ell}{n})^\alpha$. The probability of selecting a tweet p_ℓ in language ℓ is proportional to the ratio of tweets in ℓ. The low-resource upsampling is controlled by α, where a smaller value upsamples low-resouce languages and downsamples high-resource languages.[4]

We develop a tokenizer on Twitter data using SentencePiece (Kudo and Richardson, 2018) on a subset of **Language-sampled** with a Unigram LM (Kudo, 2018) character coverage of 99.995% and special symbols `@USER` and `HTTPURL`. The subset is the first 120M characters of shuffled tweets from each month, resulting in 78M tweets with 8.4B characters. Following XLM-R and other multilingual models, we chose a vocabulary size of 250K (Xue et al., 2021b; Conneau et al., 2020).

2.2 Training

Bernice follows the same architecture and pre-training as BERTweet (Nguyen et al. (2020); L=12, H=768, A=12, 270M params), which is a $\text{BERT}_{\text{base}}$ model (Devlin et al., 2019) trained with the RoBERTa Masked Language Modeling (MLM) objective (Liu et al., 2019). Bernice has a max sequence length of 128 subword tokens to accommodate the short nature of tweets (280 characters). We use the `fairseq` (Ott et al., 2019) implementation of the RoBERTa pre-training objective and "pack" tweets to fill the maximum sequence length. Our hyperparameters are in Appendix Table 5.

We trained on AWS EC2 with a p3.16xlarge instance with 8 NVIDIA Tesla V100 GPUs. Total train time was 330 hours for 405K steps. We trained the model on **Language-sampled** for the first 330K steps and then continued training on **Presampled** for the final 75K steps.[5] We start with **Language-sampled** to expose the model to significant amount of low-resource languages. However, since this sample is static, the model could overfit on the same subset of high-resource languages. We switch to sampling **Presampled** on the fly to prevent this overfitting. We track Bernice's training progress by checking the loss and perplexity on a validation set, a concatenation of 5K tweets randomly sampled from each month data.

3 Background: Twitter BERT Models

Several different BERT models have been trained on Twitter data. In this work we compare to BERTweet (Nguyen et al., 2020), XLM-T (Barbieri et al., 2021), TwHIN-BERT (Zhang et al., 2022), and TwHIN-BERT-MLM (Zhang et al., 2022). While not a Twitter-specific model, we also include XLM-R (Conneau et al., 2020) to compare to a model without any Twitter pre-training.[6] We

[3] Other models use large amounts of out of domain data to supplement training.

[4] We follow Lample and Conneau (2019) and set $\alpha = 0.5$.

[5] We trained until our budget was depleted.

[6] XLM-R might have seen tweets in the Common Crawl pre-training data, but still was not *specifically* trained on tweets.

provide a brief overview and compare them with respect to multilinguality, pre-training data, and tokenizers, (see Table 1 for a summary).

Besides BERTweet, which is trained on 850M English tweets, all comparison models are multilingual. Also, all models besides XLM-R have been pre-trained on tweets. XLM-T is unique because Barbieri et al. (2021) continued training the last XLM-R_{base} checkpoint on 198M multilingual tweets. Thus, out of all the Twitter models, XLM-T is domain-adapted to Twitter and not trained completely on in-domain data. XLM-T has also seen the most data overall, when taking into consideration XLM-R's pre-training on 2.5TB of CommonCrawl (CC-100). XLM-R has seen the second-most amount of data, followed by both TwHIN-BERT models (referred to as TwHIN-BERT*) with 7B tweets, and then Bernice with 2.5B tweets. The TwHIN-BERT model is also notable because of its combined novel social pre-training objective and RoBERTa MLM objective. TwHIN-BERT-MLM is the same as TwHIN-BERT, except only trained with the standard MLM objective. We compare Bernice against both models in Section 4 to evaluate the effects of more Twitter pre-training data and the social objective.

BERTweet and Bernice are the only models with tokenizers trained on tweets. XLM-T and TwHIN-BERT* models use the XLM-R tokenizer, which was trained on out-of-domain CommonCrawl data and has the same vocabulary size as Bernice (250K tokens), but covers 100 languages versus 66.[7] While trained on tweets, BERTweet is monolingual and only has a vocabulary size of 64K tokens. In §4, we explore the difference in token coverage between the XLM-R and Bernice tokenizers, and how it could impact downstream task performance.

In addition, Bernice, BERTweet, and TwHIN-BERT* models all have a smaller max sequence length to accommodate the short nature of tweets. Their max sequence length is 128, a quarter of the length of XLM-T/XLM-R's max length of 512,[8] which enables faster training and inference.

In summary, Bernice is the only multilingual model trained from scratch on Twitter data with a custom tokenizer.

[7] See Section 2.1.

[8] Nguyen et al. (2020) found an average of 25 subword tokens per tweet.

4 Evaluation

We benchmark Bernice's performance on three datasets: TweetEval, UMSAB, and Hate Speech. We also compare tokenizer coverage of hashtags and emoji, which are unique to Twitter.

Given our multilingual focus, we compare to XLM-T (Barbieri et al., 2021), XLM-R (Conneau et al., 2020), TwHIN-BERT (Zhang et al., 2022), and TwHIN-BERT-MLM (Zhang et al., 2022) which were previously shown to have competitive performance to monolingual English models for tweets (Conneau et al., 2020; Zhang et al., 2022).

4.1 TweetEval

TweetEval is an English Twitter model benchmark that contains seven heterogeneous Twitter-specific classification tasks (Barbieri et al., 2020). For each task, we fine-tuned all layers of the models in addition to training the classification head. Summary results are shown in Table 2 with comparative model scores from Barbieri et al. (2021). See Appendix Table 9 for individual task scores. Hyperparameter settings for tuning are in Appendix §B.

BERTweet is the best performing model on all tasks other than Offensive Language Identification, where it is outperformed by Bernice and RoBERTa-RT by 1.0 F1 and 0.5 F1, respectively. Bernice performs similarly to XLM-T and TwHIN-BERT-MLM, typically within 1.0 F1 of either model. These results are consistent with previous work that found multilingual models perform worse than monolingual models for high resource languages (Mueller et al., 2020). Bernice is trained on 66 languages, while BERTweet is trained on only English. Further pre-training on English could be beneficial for English-only tasks. For the remaining benchmarks, we only evaluate the multilingual models.

4.2 Unified Multilingual Sentiment Analysis Benchmark (UMSAB)

We now turn to the more important topic of multilingual evaluation, the focus of Bernice training. We first consider the Unified Multilingual Sentiment Analysis Benchmark (UMSAB), developed by Barbieri et al. (2021) to evaluate XLM-T on multilingual Twitter data. The benchmark is a collection of eight monolingual sentiment analysis datasets (Arabic, English, French, German, Hindi, Italian, Portuguese, and Spanish). We fine-tuned Bernice for the multilingual setting – the model is trained and evaluated on all languages – and the zero-shot

	Bernice	BERTweet	XLM-R	XLM-T	TwHIN-BERT-MLM	TwHIN-BERT
TweetEval	64.80	**67.90**	57.60	64.40	64.80	63.10
UMSAB	**70.34**	-	67.71	66.74	68.10	67.53
Hate Speech	**76.20**	-	74.54	73.31	73.41	74.32

Table 2: Average results on TweetEval, UMSAB, and Hate Speech benchmarks. Results are averaged across tasks for TweetEval and across languages for UMSAB and Hate Speech. See Appendix B for detailed benchmark scores. BERTweet is monolingual and excluded from the UMSAB and Hate Speech benchmarks.

setting – the model is trained on one language and evaluated on another. We evaluate across all language pairs, as in the original paper.

Barbieri et al. (2021) use adapters to fine-tune XLM-T and XLM-R on UMSAB. For consistency with Bernice, we re-train these models with classification layers using the HuggingFace library. Hyperparameter settings are in Appendix Table 5. For the multilingual task (shown in Table 2), Bernice performs the best across all languages. From the individual language scores in Appendix Table 12, we see Bernice has the highest F1 score on all but one language over XLM-R and TwHIN-BERT-MLM, and every language over XLM-T. XLM-T and TwHIN-BERT* models outperform Bernice on English, indicating that further pre-training on English tweets could be beneficial. In the zero-shot task (see Appendix Table 13), Bernice has the highest score on all languages over TwHIN-BERT, five languages over XLM-T, and all but one for XLM-R (Hindi) and TwHIN-BERT-MLM (German).

4.3 Multilingual Hate Speech

The second multilingual task we consider is the Multilingual Hate Speech benchmark (Aluru et al., 2020), a curated collection of hate speech detection datasets for nine languages: Arabic, English, French, German, Indonesian, Italian, Polish, Portuguese, and Spanish. We follow their data collection steps for each Twitter dataset, skipping the Facebook and Stormfront datasets, resulting in 121,887 total examples.[9] We fine-tuned Bernice, XLM-T, TwHIN-BERT*, and XLM-R on the combined datasets, as in Aluru et al. (2021), with the same hyperparameter search as for UMSAB.

As shown in Table 2, all models with Twitter pre-training outperform XLM-R. Still, overall Bernice outperforms XLM-T, TwHIN-BERT-MLM, and TwHIN-BERT by roughly $1.5 - 2.0$ F1. See Appendix Table 10 for individual language scores.

4.4 The Twitter Tokenizer

As shown in Table 1, Bernice is the model multilingual Twitter model with a custom tokenizer built for tweets. This customization for tweets could come at a disadvantage for representing more languages, as the XLM-R tokenizer was explicitly trained on 100 languages and Bernice only on the 66 languages identified by Twitter metadata, including an "undefined" category. While trained on very different datasets, the XLM-R and Bernice tokenizer vocabulary has 43.53% overlap of 108,821 tokens.

We compare the models' language representation power by evaluating the subword token coverage of the tokenizers. Prior work demonstrates that models with better subword token data coverage perform better on downstream tasks (Wu and Dredze, 2020). A model has better coverage of data than another when *less* subwords are needed to represent the text and the subwords are *longer*. We evaluate tokenizer coverage on Twitter-specific features, primarily hashtags and emoji.[10] Since XLM-T, XLM-R, and TwHIN-BERT* models all use the XLM-R tokenizer, we only compare the Bernice and XLM-R tokenizers.

Hashtags To evaluate hashtag representation, an important feature of tweets, we use a collection of hashtags from Twitter trending topics (see Appendix C). In Table 3, we see Bernice uses less subwords on average to represent hashtags, and the subwords are longer, implying better coverage for hashtags.

As shown in Appendix Table 7, the Bernice tokenizer has learned pop culture (e.g., "Netflix"), sports (e.g., "Draft"), and political (e.g. "GOP") references commonly discussed on Twitter, as shown by keeping those tokens intact as subwords.

The XLM-R tokenizer could possibly provide better coverage if hashtags are first identified and

[9] Tweets are deleted or removed over time, so we were only able to recover 87.7% of the data from the original paper. See Appendix Appendix C.2.

[10] See Appendix C.2 for coverage analysis across languages.

	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)

Table 3: Average number of subwords and length of subwords (characters) to represent the 6,125 hashtags and sampled tweets. The standard deviation is shown in parenthesis.

Tokenizer	Emoji Subwords	Avg. Length	Med. Length
Bernice	3405	2.3 (1.8)	2.0
XLM-R	562	1.0 (0.16)	1.0

Table 4: The number of tokens in the vocabulary that contain an emoji, and the average and median lengths of those subword tokens. The standard deviation is shown in parenthesis.

split into their constituent words, which is straightforward for hashtags written in Camel case or with words split with underscores, but difficult for hashtags without these obvious separations.

Emoji Emoji are also an essential part of the "language" on Twitter. Both tokenizers contain hundreds of subwords that contain, or are solely, emoji. We use the `emoji` Python package to identify subwords that contain at least one emoji.[11] The Bernice tokenizer is unique because it learned frequent grouping of emoji, as shown by the average emoji-subword length in Table 4. The groupings are mostly repeated emoji, but some are groupings that have a specific meaning in pop culture (see Appendix Table 8).

5 Discussion

Across three benchmark datasets Bernice consistently outperforms or matches its closest competitors XLM-T and TwHIN-BERT-MLM, providing the community with a pre-trained multilingual encoder exclusively for Twitter. What have we learned through the development of this model?

First, Bernice was much cheaper (faster) to train than its closest competitors. For XLM-T, Barbieri et al. (2021) continued training from the final XLM-R checkpoint by adding 198M multilingual tweets. While roughly 12x less Twitter data than Bernice, it is far more overall data as it includes the 2.5TB of CC-100 pre-training data for XLM-R (Conneau et al., 2020). The total cost in terms of data and compute to get to a better result was less for Bernice. We attribute this to the efficiency of in-domain pre-training data and smaller context size. Despite using orders of magnitude less data, we have enough Twitter data to learn a better model than through adaptation.

A similar story is seen for the TwHIN-BERT* models, where Zhang et al. (2022) trained a model from scratch on 7B tweets. While still considerably less data than for XLM-T, TwHIN-BERT* models were trained on almost 3x the number of tweets as for Bernice. Despite this large increase in in-domain data, TwHIN-BERT* models perform similarity to Bernice on the benchmarks evaluated here. This similar performance could be due to Bernice's Twitter-specific tokenizer.

Second, we made modeling choices for Twitter specifically. We followed Nguyen et al. (2020) and used a max sequence length of 128 subwords; longer contexts (512 for XLM-T) are unnecessary for individual tweets (maximum 280 characters). The limited context led to more efficient training and inference.[12]

Our other modeling choice was in how we developed the vocabulary. We demonstrated in Section 4.4 that Bernice's tokenizer is better suited for Twitter data than XLM-R's, which is used by XLM-T and TwHIN-BERT* models.[13]

Finally, we purposely used a random selection of tweets for pre-training, rather than selection by location/topic, to produce a general purpose multilingual Twitter encoder with utility for as many tasks as possible. We encourage users to fine-tune on specific tasks and to continue pre-training on relevant Twitter data (Gururangan et al., 2020).[14]

6 Conclusion

Despite the prominence of NLP for Twitter data, there are few language models designed especially for this domain. Bernice is the first BERT style multilingual model trained from scratch for Twitter on 2.5B tweets with a custom tokenizer. Bernice outperforms comparative models with significantly less overall training by focusing exclusively on in-domain Twitter pre-training data.

[11] https://github.com/carpedm20/emoji/

[12] A smaller max sequence length allows for more examples to fit in a batch for GPU training.

[13] An alternative would be vocabulary adaptation (Sato et al., 2020; Nayak et al., 2020), which we do not explore in this work.

[14] We note that it remains an open question as to whether task-agnostic pre-training is superior to task-focused training schemes (e.g. multi-task learning) (Dery et al., 2022).

7 Limitations

In this work, we introduced Bernice, the first multilingual pre-trained encoder trained exclusively on Twitter data with a custom tokenizer, and evaluate the model on English and multilingual benchmarks. The limitations of this work concern choices with regards to model training and evaluation.

As with all social media data, there exists spam and hate speech. We cleaned our data by filtering for tweet length, but the possibility of this spam remains. Hate speech is difficult to detect, especially across languages and cultures (Ousidhoum et al., 2019a; Huang et al., 2020b), thus we leave its removal for future work. Also, we chose to train the base-size transformer model instead of large, due to base models being more accessible because of compute power.

Limitations in model evaluation are due to only evaluating performance across certain tasks and demographics. We looked at minority populations by language speaker in Appendix C.2, but not by other demographics. Previous work suggests that models do poorly on some demographics (Aguirre et al., 2021). Within languages, even with language sampling during training (see §2.1), Bernice is still not exposed to the same variety of examples in low-resource languages as high-resource languages like English and Spanish. It is unclear whether enough Twitter data exists in these languages, such as Tibetan and Telugu, to ever match the performance on high-resource languages. Only models more efficient at generalizing can pave the way for better performance in the wide variety of languages in this low-resource category.

Also, model performance may vary across applications not covered in this work, or on other datasets. There is a possibility that other models, such as XLM-T and TwHIN-BERT*, might have an advantage in applications focused on more formally-written tweets, such as those from news organizations, whereas Bernice might have an advantage in the "average" Twitter conversation.

Acknowledgements

We thank the anonymous EMNLP reviewers, Suzanna Sia, Farnaz Yousefi, and Rachel Wicks for their time and helpful suggestions. We also thank Xinyang Zhang for his help with the TwHIN-BERT comparisons.

References

Carlos Aguirre, Keith Harrigian, and Mark Dredze. 2021. Gender and racial fairness in depression research using social media. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2932–2949, Online. Association for Computational Linguistics.

Ika Alfina, Rio Mulia, Mohamad Ivan Fanany, and Yudo Ekanata. 2017. Hate speech detection in the indonesian language: A dataset and preliminary study. In *2017 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, pages 233–238.

Nasser Alsaedi, Pete Burnap, and Omer Rana. 2017. Can we predict a riot? disruptive event detection using twitter. *ACM Trans. Internet Technol.*, 17(2).

Sai Saketh Aluru, Binny Mathew, Punyajoy Saha, and Animesh Mukherjee. 2020. A Deep Dive into Multilingual Hate Speech Classification. volume 5, pages 423–439, Ghent, Belgium. Springer Nature Switzerland AG 2021.

Sai Saketh Aluru, Binny Mathew, Punyajoy Saha, and Animesh Mukherjee. 2021. A deep dive into multilingual hate speech classification. In *Machine Learning and Knowledge Discovery in Databases. Applied Data Science and Demo Track*, pages 423–439. Springer International Publishing.

Francesco Barbieri, Luis Espinosa Anke, and Jose Camacho-Collados. 2021. XLM-T: A Multilingual Language Model Toolkit for Twitter. *arXiv:2104.12250 [cs]*. ArXiv: 2104.12250.

Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. 2020. TweetEval: Unified Benchmark and Comparative Evaluation for Tweet Classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1644–1650, Online. Association for Computational Linguistics.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Adrian Benton and Mark Dredze. 2018. Using Author Embeddings to Improve Tweet Stance Classification. In *Proceedings of the 2018 EMNLP Workshop W-NUT: The 4th Workshop on Noisy User-generated Text*, pages 184–194, Brussels, Belgium. Association for Computational Linguistics.

Cristina Bosco, Felice Dell'Orletta, Fabio Poletto, Manuela Sanguinetti, and Maurizio Tesconi. 2018.

Overview of the evalita 2018 hate speech detection task. In *EVALITA@CLiC-it*.

David A. Broniatowski, Amelia M. Jamison, SiHua Qi, Lulwah AlKulaib, Tao Chen, Adrian Benton, Sandra C. Quinn, and Mark Dredze. 2018. Weaponized health communication: Twitter bots and russian trolls amplify the vaccine debate. *American Journal of Public Health*, 108(10):1378–1384. PMID: 30138075.

Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and Ming Zhou. 2021. InfoXLM: An Information-Theoretic Framework for Cross-Lingual Language Model Pre-Training. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3576–3588, Online. Association for Computational Linguistics.

Abhinav Chinta, Jingyu Zhang, Alexandra DeLucia, Mark Dredze, and Anna L. Buczak. 2021. Study of manifestation of civil unrest on Twitter. In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pages 396–409, Online. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated Hate Speech Detection and the Problem of Offensive Language. *Proceedings of the International AAAI Conference on Web and Social Media*, 11(1):512–515.

Lucio M. Dery, Paul Michel, Ameet Talwalkar, and Graham Neubig. 2022. Should we be pre-training? an argument for end-task aware training as an alternative. In *International Conference on Learning Representations*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Jesse Dodge, Suchin Gururangan, Dallas Card, Roy Schwartz, and Noah A. Smith. 2019. Show your work: Improved reporting of experimental results. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2185–2194, Hong Kong, China. Association for Computational Linguistics.

Paula Fortuna, João Rocha da Silva, Juan Soler-Company, Leo Wanner, and Sérgio Nunes. 2019. A hierarchically-labeled Portuguese hate speech dataset. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 94–104, Florence, Italy. Association for Computational Linguistics.

Antigoni Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large Scale Crowdsourcing and Characterization of Twitter Abusive Behavior. *Proceedings of the International AAAI Conference on Web and Social Media*, 12(1).

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. *ACM Transactions on Computing for Healthcare*, 3(1):2:1–2:23.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Association for Computational Linguistics (ACL)*.

Keith Harrigian, Carlos Aguirre, and Mark Dredze. 2021. On the state of social media data for mental health research. In *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access*, pages 15–24, Online. Association for Computational Linguistics.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2020a. ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission. Number: arXiv:1904.05342 arXiv:1904.05342 [cs].

Xiaolei Huang, Linzi Xing, Franck Dernoncourt, and Michael J. Paul. 2020b. Multilingual Twitter Corpus and Baselines for Evaluating Demographic Bias in Hate Speech Recognition. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1440–1448, Marseille, France. European Language Resources Association.

Muhammad Okky Ibrohim and Indra Budi. 2019. Multi-label hate speech and abusive language detection in Indonesian Twitter. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 46–57, Florence, Italy. Association for Computational Linguistics.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75, Melbourne, Australia. Association for Computational Linguistics.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual Language Model Pretraining. Technical Report arXiv:1901.07291, arXiv. ArXiv:1901.07291 [cs] type: article.

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:1907.11692 [cs]*. ArXiv: 1907.11692.

David Mueller, Nicholas Andrews, and Mark Dredze. 2020. Sources of transfer in multilingual named entity recognition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8093–8104, Online. Association for Computational Linguistics.

Hala Mulki, Hatem Haddad, Chedi Bechikh Ali, and Halima Alshabani. 2019. L-HSAB: A Levantine Twitter dataset for hate speech and abusive language. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 111–118, Florence, Italy. Association for Computational Linguistics.

Anmol Nayak, Hariprasad Timmapathini, Karthikeyan Ponnalagu, and Vijendran Gopalan Venkoparao. 2020. Domain adaptation challenges of bert in tokenization and sub-word representations of out-of-vocabulary words. In *Proceedings of the First Workshop on Insights from Negative Results in NLP*, pages 1–5.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. BERTweet: A pre-trained language model for English Tweets. *arXiv:2005.10200 [cs]*. ArXiv: 2005.10200.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A Fast, Extensible Toolkit for Sequence Modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

Nedjma Ousidhoum, Zizheng Lin, Hongming Zhang, Yangqiu Song, and Dit-Yan Yeung. 2019a. Multilingual and Multi-Aspect Hate Speech Analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4675–4684, Hong Kong, China. Association for Computational Linguistics.

Nedjma Ousidhoum, Zizheng Lin, Hongming Zhang, Yangqiu Song, and Dit-Yan Yeung. 2019b. Multilingual and multi-aspect hate speech analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4675–4684, Hong Kong, China. Association for Computational Linguistics.

Juan Carlos Pereira-Kohatsu, Lara Quijano-Sánchez, Federico Liberatore, and Miguel Camacho-Collados. 2019. Detecting and monitoring hate speech in twitter. *Sensors*, 19(21).

Michal Ptaszynski, Agata Pieciukiewicz, and Paweł Dybała. 2019. Results of the poleval 2019 shared task 6 : first dataset and open shared task for automatic cyberbullying detection in polish twitter. In Maciej Ogrodniczuk and Łukasz Kobyliński, editors, *Proceedings of the PolEval 2019 Workshop*, pages 89–110. Institute of Computer Sciences. Polish Academy of Sciences, Warszawa.

Björn Ross, Michael Rist, Guillermo Carbonell, Benjamin Cabrera, Nils Kurowsky, and Michael Wojatzki. 2017. Measuring the reliability of hate speech annotations: The case of the european refugee crisis. *CoRR*, abs/1701.08118.

Manuela Sanguinetti, Fabio Poletto, Cristina Bosco, Viviana Patti, and Marco Stranisci. 2018. An Italian Twitter corpus of hate speech against immigrants. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Shoetsu Sato, Jin Sakuma, Naoki Yoshinaga, Masashi Toyoda, and Masaru Kitsuregawa. 2020. Vocabulary adaptation for domain adaptation in neural machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4269–4279.

Kevin Scannell. 2020. Universal Dependencies for Manx Gaelic. In *Proceedings of the Fourth Workshop on Universal Dependencies (UDW 2020)*, pages 152–157, Barcelona, Spain (Online). Association for Computational Linguistics.

Kevin P. Scannell. 2022a. Managing Data from Social Media: The Indigenous Tweets Project. In *The Open Handbook of Linguistic Data Management*. The MIT Press.

Kevin P. Scannell. 2022b. Managing Data from Social Media: The Indigenous Tweets Project.

Justin Sech, Alexandra DeLucia, Anna L. Buczak, and Mark Dredze. 2020. Civil unrest on Twitter (CUT): A dataset of tweets to support research on civil unrest. In *Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020)*, pages 215–221, Online. Association for Computational Linguistics.

Stefanos Stoikos and Mike Izbicki. 2020. Multilingual Emoticon Prediction of Tweets about COVID-19. In *Proceedings of the Third Workshop on Computational Modeling of People's Opinions, Personality, and Emotion's in Social Media*, pages 109–118, Barcelona, Spain (Online). Association for Computational Linguistics.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California. Association for Computational Linguistics.

Shijie Wu and Mark Dredze. 2020. Are All Languages Created Equal in Multilingual BERT? In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 120–130, Online. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021a. mT5: A massively multilingual pre-trained text-to-text transformer. Technical Report arXiv:2010.11934, arXiv. ArXiv:2010.11934 [cs] type: article.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021b. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Xinyang Zhang, Yury Malkov, Omar Florez, Serim Park, Brian McWilliams, Jiawei Han, and Ahmed El-Kishky. 2022. TwHIN-BERT: A Socially-Enriched Pre-trained Language Model for Multilingual Tweet Representations. ArXiv:2209.07562 [cs].

A Appendix

Optimizer	Algorithm	Adam
	Learning rate (LR)	5×10^{-4}
	Epsilon (eps, ϵ)	1×10^{-6}
	LR scheduler	linear decay and warmup
	Warmup steps	24000
	Gradient clip norm	1.0
	Betas	$(0.9, 0.98)$
	Weight decay	0.01
Batch	Sequence length	128
	Batch size	8192
Misc	Dropout	0.1
	Attention dropout	0.1

Table 5: All hyperparameters and `fairseq` settings for training Bernice.

B Evaluation Details

B.1 Hyperparameter Search

We report the hyperparameter search and best model settings as recommended by Dodge et al. (2019).

TweetEval We use the reported benchmark values from Barbieri et al. (2020) for BERTweet, RoBERTa-RT, RoBERTa-Tw, XLM-R, and XLM-T. For Bernice and TwHIN-BERT* we followed Barbieri et al. (2020) and performed a parameter search over learning rate values of 1×10^{-3}, 1×10^{-4}, 1×10^{-5}, and 1×10^{-6}, with a batch size of 32 for 5 epochs. The best models for all tasks for Bernice had a learning rate of 1×10^{-5}. The best learning rate for TwHIN-BERT-MLM was 1×10^{-5} for all tasks except Stance. The best learning rate for TwHIN-BERT varied as follows: 1×10^{-5} for Emoji, Emotion, Offensive, and Sentiment, and 1×10^{-4} for the remaining tasks.

UMSAB For each model we perform a hyperparameter search over learning rates $1 \times 10^{-5}, 2 \times 10^{-5}, 5 \times 10^{-5}$ and batch sizes $8, 16, 32$. We set the max epochs to 20 with a patience of 5 based on macro-F1 score. The best settings for all models except TwHIN-BERT* were batch size 32, learning rate of 1×10^{-5}. The TwHIN-BERT* models had the same best learning rate but different best batch sizes. TwHIN-BERT-MLM had a batch size of 8 and TwHIN-BERT had a batch size of 16.

Hate Speech For each model we perform the same hyperparameter search as for UMSAB. The best performing models had a learning rate of 1×10^{-5}. All models other than XLM-T had a batch size of 32, with XLM-T using a batch size of 16.

B.2 Detailed Benchmark Results

The task- and language-specific scores for the TweetEval, UMSAB multilingual, UMSAB zero-shot, and Hate Speech benchmarks are in Tables 9, 10, 12 and 13, respectively.

C Tokenizer Analysis

C.1 Data

Unseen 2022 Data For general general perplexity and tokenizer analyses, we evaluate on data that none of the models have seen. Bernice was trained on tweets from the 1% Twitter public stream through December 2021, and TwHIN-BERT* models were trained on data through June 2022. Since none of the models have seen data after June 2022, we gathered tweets from the 1% public stream from July 1, 2022 to October 8, 2022. To reduce the millions of tweets, we randomly sampled 10,000 tweets from each language from July to October.

Indigenous Tweets Project The project is a database of users that are known to tweet in a given low-resource, or indigenous, language. The database contains users for 185 languages. As per Twitter Terms of Service, Scannell (2022b) do not distribute the tweets directly, instead hosting a website that contains which users tweet in a specific language. We scrape the website for all users and other metadata, such as an estimate of how many tweets are in the language versus another language. From the user names (i.e., screen names) we use the Twitter API to collect their tweet history.

We were not able to recover the full dataset and number of tweets per user reported by Indigenous Tweets due to API limits (can only gather 3,200 tweets per user), accounts that were made private, and deleted tweets.

To reduce noise from misidentified languages, we restrict the database to tweets from users who tweet in a given language at least 90% of the time. This reduced the dataset to 3,318 tweets in 40 languages, 30 of which are shown in Table 6. Since the collected dataset was small, we did not remove tweets that could have been present in our training dataset.

Twitter Trending Hashtags We created a Trending Hashtag dataset by collecting trending topics

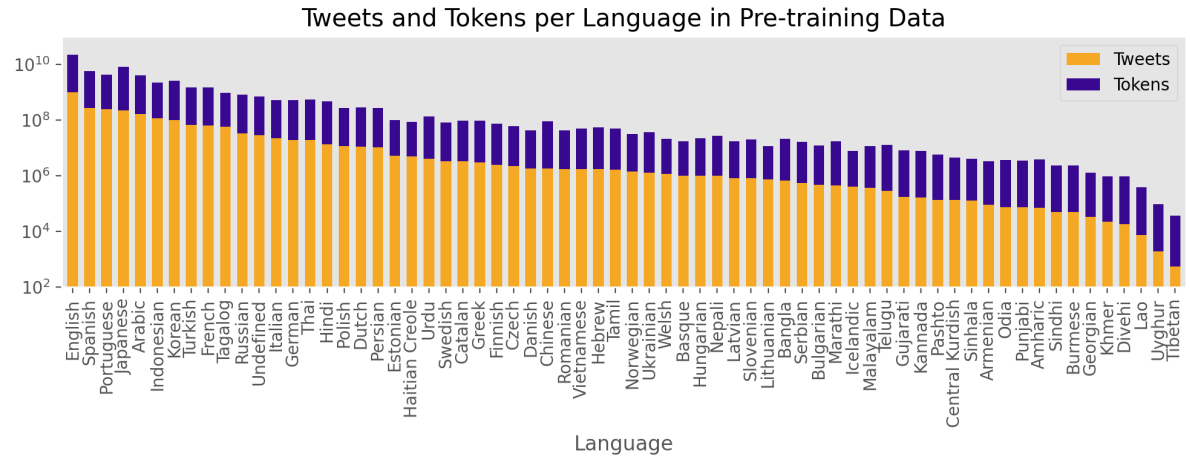

Figure 1: Language distribution of tweets and tokens for the 2.5B tweets in the **Presampled** pre-training dataset. "Undefined" refers to the tweets whose language was not identified by the Twitter API. Counts are in log scale.

around the world with the `tweepy`[15] Python package. From September 23 to October 20, we collected 32,629 unique topics from 17 countries. We reduced the topics to those containing hashtags, resulting in 6,125 hashtags. From visually inspecting the data, the hashtags appear to be in English, Spanish, Arabic, Japanese, Hindi, and Korean. Other languages may be present. See Appendix C.2 for a language-specific evaluation.

C.2 Language Coverage

The benchmarks in Section 4 focused on high-resource languages. Similar benchmarks for low-resource languages are either non-existent or focus on a single language. To approximate our model's performance on downstream tasks concerning low-resource languages, we evaluate its token coverage on tweets from the Indigenous Tweets Project (Scannell, 2020).[16] We compare against token coverage on unseen tweets in high-resource languages from 2022 for comparison. The data collection details are in Appendix C.

As in Section 4.4, we compare Bernice's and XLM-R's tokenizer. The average number of subwords in a tokenized tweet and the average length of a subword (i.e., characters) for all languages are shown in Table 3. The per-language breakdown is in Table 6. For clarity, we only include the top 10 languages in the pre-training data (**Presampled**) and the 30 least prevalent languages from the Indigenous Tweets Project (denoted with a line).

From the language coverage numbers, we see Bernice's tokenizer has greater coverage of the high resource languages than XLM-R's tokenizer. This trend changes as we approach the lower resource languages from the Indigenous Tweet project (identified with an *). For these ultra low-resource languages, we see the tokenizers' performances vary. To determine if a model has better coverage of a specific language, more analyses and performance metrics on language-specific downstream tasks are needed. We leave this as an avenue for future work. Possible tasks that do not require manual labeling are hashtag prediction (Zhang et al., 2022) and emoji prediction (Stoikos and Izbicki, 2020).

[15] https://www.tweepy.org/
[16] http://indigenoustweets.com

Language	Pre-train Data	Sample	Bernice # Subwords	XLM-R # Subwords	Bernice Lengths	XLM-R Lengths
en	966,874,161	10000	**21.24 (17.51)**	24.64 (19.97)	**3.91 (2.38)**	3.37 (2.09)
es	262,919,279	10000	**19.32 (16.50)**	22.76 (18.82)	**4.04 (2.43)**	3.43 (2.12)
pt	238,075,817	10000	**15.42 (13.43)**	18.56 (15.27)	**4.12 (2.42)**	3.42 (2.08)
ja	221,188,011	10000	**20.06 (17.05)**	26.34 (21.22)	**2.24 (1.67)**	1.71 (1.10)
ar	163,753,840	10000	**31.98 (29.79)**	38.13 (33.23)	**2.64 (2.04)**	2.22 (1.52)
in	113,542,378	10000	**16.06 (15.44)**	19.18 (17.33)	**3.78 (2.19)**	3.17 (1.92)
ko	97,851,118	10000	**20.07 (18.38)**	24.71 (21.77)	**2.06 (1.34)**	1.67 (0.96)
tr	66,774,565	10000	**21.54 (17.49)**	25.70 (19.91)	**3.92 (2.36)**	3.29 (1.94)
fr	64,201,496	10000	**22.79 (18.87)**	25.45 (20.08)	**3.70 (2.31)**	3.31 (2.08)
tl	57,654,649	10000	**13.81 (12.24)**	17.09 (14.16)	**3.71 (2.17)**	3.00 (1.70)
km	22,637	4146	17.87 (19.11)	**14.57 (13.79)**	2.09 (1.48)	**2.56 (1.59)**
*dv	18,460	20	24.00 (3.26)	**14.65 (2.10)**	2.61 (1.68)	**4.28 (3.69)**
lo	7,356	3331	19.47 (23.45)	**15.49 (16.45)**	1.85 (1.27)	**2.33 (1.31)**
ug	1,902	103	58.66 (39.06)	**46.09 (26.66)**	2.17 (1.26)	**2.76 (1.78)**
bo	552	106	45.70 (43.02)	**16.96 (13.68)**	1.63 (1.17)	**4.40 (7.08)**
*ga	0	268	38.42 (10.91)	**32.74 (8.92)**	2.64 (1.43)	**3.10 (1.65)**
*haw	0	242	33.49 (15.73)	**32.81 (15.23)**	2.21 (1.32)	**2.26 (1.24)**
*fy	0	112	38.28 (9.01)	**36.29 (8.53)**	2.98 (1.55)	**3.14 (1.49)**
*gd	0	82	43.60 (7.29)	**39.10 (6.07)**	2.71 (1.48)	**3.02 (1.52)**
*la	0	64	29.33 (11.93)	**27.16 (11.23)**	3.15 (1.87)	**3.40 (1.97)**
*kw	0	60	18.30 (13.92)	**18.27 (13.89)**	2.85 (1.58)	**2.86 (1.34)**
*mia	0	58	**19.69 (7.46)**	22.26 (8.21)	**2.80 (1.53)**	2.48 (1.18)
*an	0	39	**29.87 (11.40)**	32.00 (12.16)	**3.40 (2.06)**	3.17 (1.79)
*gv	0	26	**22.12 (11.13)**	23.15 (10.24)	**2.69 (1.45)**	2.57 (1.26)
*lkt	0	25	70.12 (21.03)	**68.92 (19.85)**	1.80 (0.89)	**1.83 (0.85)**
*fj	0	23	**43.87 (10.44)**	46.13 (10.50)	**2.61 (1.28)**	2.48 (1.13)
*mg	0	23	34.22 (11.01)	**29.52 (8.98)**	2.86 (1.70)	**3.31 (1.64)**
*gn	0	20	**38.45 (10.49)**	38.45 (11.20)	2.37 (1.34)	**2.37 (1.27)**
*ha	0	20	11.25 (3.67)	**11.10 (3.73)**	2.88 (1.43)	**2.91 (1.41)**
*chr	0	19	49.95 (17.24)	**29.79 (9.65)**	1.24 (0.91)	**2.08 (1.61)**
*hsb	0	17	45.88 (4.78)	**45.12 (4.42)**	2.76 (1.54)	**2.81 (1.50)**
*cv	0	8	**24.75 (8.55)**	24.75 (9.88)	2.26 (1.70)	**2.26 (1.49)**
*ch	0	6	**26.50 (12.53)**	29.33 (13.74)	**2.80 (1.71)**	2.53 (1.52)
*mi	0	6	**11.67 (7.18)**	13.83 (9.21)	**4.26 (2.67)**	3.59 (2.28)
*gil	0	4	**42.50 (6.02)**	43.25 (5.76)	**2.95 (1.38)**	2.90 (1.33)
*ay	0	2	**14.50 (0.50)**	16.50 (0.50)	**3.52 (1.77)**	3.09 (1.54)
*li-x-east	0	2	**15.50 (2.50)**	16.00 (1.00)	**2.97 (1.33)**	2.88 (1.24)
*ff	0	1	15.00 (0.00)	**14.00 (0.00)**	2.87 (1.09)	**3.07 (1.22)**
*kl	0	1	**42.00 (0.00)**	45.00 (0.00)	**3.00 (1.21)**	2.80 (1.17)
*oc	0	1	47.00 (0.00)	**46.00 (0.00)**	2.98 (1.54)	**3.04 (1.52)**

Table 6: The number of tweets per language in our sample from the 1% Twitter API and the average number of subwords per tweet and the average subword length for the Bernice and XLM-R tokenizers. Standard deviation is in parentheses. "Pre-train" is the number of tweets in the **Presampled** dataset. Languages with an * are from the Indigenous Tweets Project.

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_الم علم	['يوم', '_', 'الم علم']	['يوم', '_', 'الم', 'علم']
#DraftKingsTNF	['Draft', 'Kings', 'TN', 'F']	['D', 'raf', 't', 'K', 'ings', 'TN', 'F']

Table 7: Example hashtags and their tokenizations with the Bernice and XLM-R tokenizers. All hashtags are preceded with "_#".

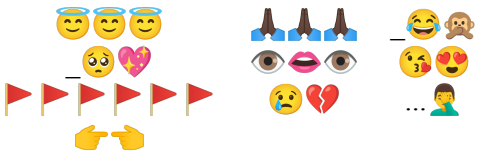

Table 8: Example subwords containing emoji from the Bernice tokenizer. Interpreting the emoji meetings are left as an exercise for the reader.

	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

Table 9: Bernice results on TweetEval benchmark (Barbieri et al., 2020) along with other comparison models. TwHIN-BERT* scores were not available and were gathered by us with the same methods as for Bernice. The scores are macro-F1 for all tasks other than Sentiment, which is macro-Recall.

	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

Table 10: Macro-F1 test set results on the Multilingual Hate Speech task. The poor performance on Polish is most likely because of a large class imbalance within that language.

Language	Dataset	Hate	Non-Hate	Total
Arabic	Mulki et al. (2019)	468	3650	4118
	Ousidhoum et al. (2019b)	755	915	1670
English	Davidson et al. (2017)	1430	4163	5593
	Waseem and Hovy (2016)	2685	7394	10079
	Basile et al. (2019)	5470	7530	13000
	Ousidhoum et al. (2019b)	1278	661	1939
	Founta et al. (2018)	1929	32458	34387
German	Ross et al. (2017)	54	315	369
Indonesian	Ibrohim and Budi (2019)	5561	7608	13169
	Alfina et al. (2017)	260	453	713
Italian	Sanguinetti et al. (2018)	785	4268	5053
	Bosco et al. (2018)	1296	2704	4000
Polish	Ptaszynski et al. (2019)	329	7978	8307
Portuguese	Fortuna et al. (2019)	1788	3882	5670
Spanish	Basile et al. (2019)	2739	3861	6600
	Pereira-Kohatsu et al. (2019)	1567	4433	6000
French	Ousidhoum et al. (2019b)	399	821	1220
Total		28793	93094	121887

Table 11: Final size of datasets collected for the Multilingual Hate Speech task (Aluru et al., 2021). Our size differs from original dataset sizes due to deleted tweets over time.

	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

Table 12: Macro-F1 test set results on UMSAB multilingual task. TwHIN-BERT is shortened to TwHIN for space.

Bernice

	Ar	En	Fr	De	Hi	It	Pt	Es	AVG
Ar	65.5	65.5	55.4	54.8	56.4	53.7	61.5	66.8	**59.9**
En	65.7	66.5	60.8	64.4	56.7	61.9	65.6	71.6	**64.1**
Fr	51.7	61.0	69.3	54.6	55.1	62.0	65.5	66.4	**60.7**
De	54.6	66.5	32.3	74.2	56.2	63.4	66.4	71.4	60.7
Hi	33.1	45.4	26.6	47.1	41.4	40.7	46.2	44.1	40.6
It	58.7	67.6	47.9	65.4	58.6	71.3	63.0	67.8	**62.5**
Pt	54.5	65.3	37.0	63.2	54.5	60.7	65.1	76.4	59.6
Es	56.6	65.0	50.9	66.9	53.1	68.5	67.9	73.6	62.8

XLM-T

	Ar	En	Fr	De	Hi	It	Pt	Es	AVG
Ar	67.5	65.0	52.4	50.3	47.3	48.9	52.2	54.5	54.8
En	65.1	67.8	59.4	63.6	52.8	59.3	64.4	70.0	62.8
Fr	48.6	59.1	74.6	49.8	44.4	59.8	59.3	59.7	56.9
De	58.0	66.3	39.8	74.0	51.6	67.5	62.4	65.3	60.6
Hi	53.7	36.7	44.4	45.3	43.0	43.4	42.5	34.8	43.0
It	52.0	60.2	39.0	63.3	49.0	68.1	55.6	63.0	56.3
Pt	64.8	68.9	46.1	65.9	50.5	66.1	62.8	73.4	**62.3**
Es	65.0	68.3	56.0	66.4	49.9	59.7	66.8	71.7	**63.0**

XLM-R

	Ar	En	Fr	De	Hi	It	Pt	Es	AVG
Ar	61.8	62.9	54.1	46.7	43.8	49.2	50.9	45.9	51.9
En	61.1	66.7	63.1	60.9	51.1	60.6	62.9	61.4	61.0
Fr	49.4	60.6	74.9	54.6	46.2	57.1	62.1	59.7	58.1
De	54.1	63.9	40.0	72.2	51.8	61.9	58.7	58.6	57.7
Hi	48.8	48.6	27.0	47.2	43.5	48.0	44.2	44.1	**43.9**
It	57.3	63.7	54.2	63.3	51.7	63.9	58.9	61.4	59.3
Pt	62.4	63.0	51.2	60.2	50.9	62.9	62.7	70.2	60.4
Es	52.2	63.0	51.4	57.8	45.7	57.1	62.2	61.0	56.3

TwHIN-BERT-MLM

	Ar	En	Fr	De	Hi	It	Pt	Es	AVG
Ar	67.6	66.0	53.7	53.2	51.2	53.0	55.7	61.4	57.7
En	68.5	68.7	58.5	64.1	52.1	59.8	64.0	68.2	63.0
Fr	55.6	64.2	69.6	54.5	53.2	57.8	61.7	67.1	60.5
De	57.8	66.4	43.3	74.4	51.2	67.8	62.0	68.5	**61.4**
Hi	39.7	41.4	33.8	37.1	43.6	37.8	37.5	45.8	39.6
It	58.7	66.2	50.2	62.8	55.1	64.7	61.5	67.4	60.8
Pt	56.8	61.8	40.1	60.0	47.7	55.5	65.0	78.4	58.2
Es	63.0	66.6	58.6	63.5	48.6	61.3	68.8	71.5	62.7

TwHIN-BERT

	Ar	En	Fr	De	Hi	It	Pt	Es	AVG
Ar	65.9	63.1	47.1	37.3	45.2	55.7	44.8	50.3	51.2
En	66.7	68.5	46.3	62.7	52.4	62.2	63.4	69.6	61.5
Fr	35.6	58.8	68.9	49.7	35.5	54.4	50.9	46.0	50.0
De	61.1	65.9	31.6	74.9	50.0	57.7	60.0	68.7	58.7
Hi	28.9	39.1	17.3	38.3	39.6	38.4	30.6	38.0	33.8
It	57.1	67.9	42.6	60.2	45.0	67.4	55.2	61.6	57.1
Pt	54.2	64.2	31.8	49.5	43.8	54.1	48.1	74.8	52.6
Es	57.3	65.0	37.4	58.7	45.9	56.2	64.2	69.1	56.7

Table 13: F1 test set results on UMSAB zero-shot cross-lingual sentiment analysis. For each row/column, the model is exclusively trained on the row-language and evaluated on the column-language. For example, in row 3 column 4, the model is trained on French data but evaluated on German data. The non-cross-lingual scores (i.e., model is evaluated on same language) are in grey. The average scores across languages are bolded with respect to highest score across the three models.