Practical Transformer-based Multilingual Text Classification

Cindy Wang
Sentropy Technologies
cindy@sentropy.io

Michele Banko
Sentropy Technologies
mbanko@sentropy.io

Abstract

Transformer-based methods are appealing for multilingual text classification, but common research benchmarks like XNLI (Conneau et al., 2018) do not reflect the data availability and task variety of industry applications. We present an empirical comparison of transformer-based text classification models in a variety of practical monolingual and multilingual pretraining and fine-tuning settings. We evaluate these methods on two distinct tasks in five different languages. Departing from prior work, our results show that multilingual language models can outperform monolingual ones in some downstream tasks and target languages. We additionally show that practical modifications such as task- and domain-adaptive pretraining and data augmentation can improve classification performance without the need for additional labeled data.

1 Introduction

While the development of natural language understanding (NLU) applications often begins with high-resource languages such as English, there is a need to create products that are accessible to speakers of the world's nearly 7,000 languages. Only 5% of the world's population is estimated to speak English as a first language.[1]

The growth of NLU-centric products within diverse language markets is evidenced by the increase in language support for popular consumer applications such as virtual assistants, Web search, and social media platforms. As of mid-2020, Google Assistant supported 44 languages on smartphones, followed by Siri (21 languages) and Amazon Alexa (8 languages). At the start of 2021, Google Search and Microsoft Bing supported 149 and 40 languages respectively. Also at this time, Twitter officially supported a total of 45 languages with Facebook reaching over 100 languages.

[1]CIA World Factbook

Advances in multilingual language models such as multilingual BERT (mBERT; Devlin et al., 2019) and XLM-RoBERTa (XLM-R; Conneau et al., 2020) which are trained on massive corpora in over 100 languages, show promise for fast iteration and deployment of NLU applications. In theory, cross-lingual approaches reduce the need for labeled training data in target languages by enabling zero- or few-shot learning. Additionally, they enable simplified model deployment compared to the use of many monolingual models. On the other hand, evaluations show that scaling to more languages causes dilution (Conneau et al., 2020) and consequently cite the relative under-performance of multilingual models on monolingual tasks (Virtanen et al., 2019; Antoun et al., 2020).

Recent studies (Hu et al., 2020; Rust et al., 2020) have explored tradeoffs of multi versus monolingual model paradigms. However, we observe that existing multilingual text classification benchmarks are designed to measure zero-shot cross-lingual transfer rather than supervised learning (Conneau et al., 2018; Yang et al., 2019), though the latter is more applicable to industry settings. Thus, the goal of this paper is to evaluate multilingual text classification approaches with a focus on real applications. Our contributions include:

- A comparison of state-of-the-art language models spanning monolingual and multilingual setups, evaluated across five languages and two distinct tasks;
- A set of practical recommendations for fine-tuning readily available language models for text classification; and
- Analyses of industry-centric challenges such as domain mismatch, labeled data availability, and runtime inference scalability.

2 Multilingual Text Classification

We consider a series of practical components for building multilingual text classification systems.

Lang.	Model	Pretraining Corpus	Tokenizer	Param.
EN	RoBERTa (Liu et al., 2019)	Various (160GB)	BPE	125M
DE	German BERT (deepset.ai, 2019)	German Wikipedia, OpenLegalData, and news articles (12 GB)	SentencePiece	110M
ES	BETO (Cañete et al., 2020)	Various (18.4GB)	WordPiece	110M
FR	CamemBERT (Martin et al., 2020)	OSCAR (138GB)	SentencePiece	110M
JA	Japanese BERT (Suzuki and Takahashi, 2019)	Japanese Wikipedia (2.6GB)	MeCab+Wordpiece	110M
MULTI	XLM-RoBERTa (Conneau et al., 2019)	CC-100 (2.5 TB) EN (301GB), DE (67GB), ES (53GB), FR (57GB), JA (69GB)	SentencePiece	270M

Table 1: Pretraining corpora, tokenizers, and size (# parameters) of the language models used in our experiments.

2.1 Pretrained Transformer Language Models

Transfer learning using pretrained language models (LMs) which are then fine-tuned for downstream tasks has emerged as a powerful technique for NLU applications. In particular, models using the now-ubiquitous transformer architecture (Vaswani et al., 2017), such as BERT (Devlin et al., 2019) and its variants, have obtained state of the art results in many monolingual and cross-lingual NLU benchmarks (Wang et al., 2019a; Raffel et al., 2020; He et al., 2021).

One drawback of data-hungry transformer models is that they are time- and resource-intensive to train. In our experiments, we consider LMs pretrained on both monolingual and multilingual corpora, and analyze the effects of combining these models with other NLU system components.

For monolingual LMs, we use BERT models pretrained on corpora in each target language. The one exception is English, where we use RoBERTa, a BERT reimplementation that exceeds its performance on an assortment of tasks (Liu et al., 2019).

For multilingual LMs, we use XLM-R, which significantly outperforms mBERT on cross-lingual benchmarks and is competitive with monolingual models on monolingual benchmarks such as GLUE (Wang et al., 2019b). All of the pretrained models used are accessible from the Hugging Face (Wolf et al., 2020) model hub, and their details are summarized in Table 1.

2.2 Domain-Adaptive and Task-Adaptive Pretraining

Though pretrained language models have hundreds of millions of parameters and are trained on diverse corpora, they are not guaranteed to generalize to all tasks and domains. For downstream tasks, a second phase of pretraining on a smaller domain- or task-specific corpus has been shown to provide performance improvements. Gururangan et al. (2020) compare domain-adaptive pretraining (DAPT), which uses a large corpus of unlabeled domain-specific text, and task-adaptive pretraining (TAPT), which uses only the training data of a particular task. The primary difference is that the task-specific corpus tends to be much smaller, but also more task-relevant. Therefore, while DAPT is helpful in both low- and high-resource settings, TAPT is much more resource-efficient and outperforms DAPT when sufficient data is available.

In our experiments, we evaluate both approaches, using the classification task training data as the TAPT corpus and in-domain unlabeled data as the DAPT corpus (see Section 3 for details). BERT and RoBERTa are pretrained with a *masked language modeling* (MLM) objective, a cross-entropy loss on randomly masked tokens in the input sequence. We similarly use the MLM objective when performing DAPT and TAPT.

2.3 Supervised Fine-Tuning

We consider three settings for supervised fine-tuning of language models for downstream classification tasks (N is the number of target languages).

- *mono-target* (N final models): Fine-tune a monolingual LM on the training data in each target language
- *multi-target* (N final models): Fine-tune XLM-R on the training data in each target language
- *multi-all* (one final model): Fine-tune XLM-R on the concatenation of all training data

To represent sequences for classification, we use the final LM hidden vectors $B \in \mathbb{R}^{l \times H}$ corresponding to each of the l input tokens.[2] We then compute average and max pools over the sequence length

[2]Though only the hidden vector for the first (`[CLS]`) token is typically used (Devlin et al., 2019), we find that the pooled sequence summary attains better results on our tasks.

Dataset	Task	Lang.	Unlab.	Train	Test
CLS (AMAZON)	Sentiment	EN	105k	6k	6k
		DE	317k	6k	6k
		FR	58k	6k	6k
		JA	294k	6k	6k
HATEVAL (TWITTER)	Hate speech	EN	-	10k	3k
		ES	-	5k	1.6k

Table 2: The target tasks, languages, and number of training and test examples in each dataset.

Hashtag	Train	Test	Test†
#NoDACA	99.36	34.26	99.60
#EndDACA	98.31	33.87	98.39
#BuildThatWall	100.0	24.89	95.99
#BuildTheDamnWall	100.0	62.07	100.0
#NoAmnesty	100.0	48.25	100.0
#SendThemBack	82.02	68.29	87.80
#DeportThemAll	100.0	83.15	99.46

Table 3: Percentage of *hateful* class by anti-immigrant hashtags in HATEVAL (non-exhaustive list). †Denotes the relabeled test set.

layer and concatenate them to create the aggregate representation $C \in \mathbb{R}^{2H}$. Finally, the summary vector C is passed to a classification layer where we compute a standard cross-entropy loss.

2.4 Data Augmentation

In real applications, labeled data is often available in high resource languages such as English but sparse or nonexistent in others. We experiment with machine translation[3] as a form of cross-lingual data augmentation, which has been shown to improve performance on multilingual benchmarks (Singh et al., 2019). In single target language settings, we translate training data from other languages into the target language, yielding N times the number of training examples. In the *multi-all* setting, we translate data from every language into every other language, yielding $N(N-1)$ times the number of training examples. At training time, we directly include the translated examples in the training corpus. Following the pretraining convention of XLM-R, we do not use special markers to denote the input language.

3 Data

We choose sentiment analysis and hate speech detection as evaluation tasks due to their relevance to industry applications and the availability of multilingual datasets. An overview of the datasets is shown in Table 2.

3.1 Sentiment Analysis

The Cross-Lingual Sentiment dataset (CLS; Prettenhofer and Stein, 2010)[4] consists of AMAZON product reviews in four languages and three product categories (BOOKS, DVD, and MUSIC). Each review includes title and body text, which we concatenate to create the input example. The dataset contains training and test sets with balanced binary sentiment labels, as well as 50-320k unlabeled examples per language. We sample 10k unlabeled examples from each language for DAPT.

3.2 Hate Speech Detection

The HATEVAL dataset (Basile et al., 2019) contains tweets in English and Spanish annotated for the presence of hate speech targeting women and immigrants. Examples were collected by querying Twitter for users with histories of sending or receiving hateful messages, as well as keywords related to women and immigrants.

Relabeling English Test Data During experimentation, we found that English example labels were inconsistent across the training and test sets. For instance, many test examples containing anti-immigration hashtags were mislabeled as *non-hateful* while similar examples were labeled as *hateful* in the training set (see Table 3). We manually relabeled 641 examples in the test set and release the relabeled data for future research.[5,6]

Unlabeled Twitter Data Since no unlabeled corpus is provided, we collected a sample of 10k random tweets per language from November 2020, which we use for DAPT.

4 Experimental Setup

Preprocessing and Tokenization We apply minimal preprocessing to both datasets, replacing URLs and Twitter usernames with `<url>` and `<user>` tokens. At all stages of training, we use the default tokenizers associated with each pretrained

[3] https://cloud.google.com/translate

[4] We use the processed version of this dataset provided by Eisenschlos et al. (2019).

[5] Prior work (Stappen et al., 2020) has also noted this discrepancy and proposed repartitioning the train and test sets. We instead relabeled the test set due to the large number of mislabeled examples.

[6] https://github.com/sentropytechnologies/hateval2019-relabeled

Model	DE	FR	JA
mBERT	84.3	86.6	81.2
MultiFiT	92.2	91.4	86.2

Model	EN	ES
Majority label	36.7	37.0
SVM + tf-idf	45.1	70.1
1st place submissions	65.1	73.0

Table 4: Prior results (macro-F1) for CLS (Eisenschlos et al., 2019, top) and HATEVAL (Basile et al., 2019, bottom).

LM (see Table 1) and truncate sequences with more than 512 tokens.

Training We use 80% of each training set for training and the rest for validation. During DAPT and TAPT, we train using the MLM objective for 10 epochs. During supervised fine-tuning, we train for 5 epochs. We use the default hyperparameters for all pretrained LMs and apply dropout of 0.4 to the final classification layer.

Evaluation We report the test set macro-averaged F1 score for both datasets. (For CLS, this is equivalent to accuracy since the classes are balanced.) For reference, prior results on CLS and HATEVAL are shown in Table 4.

5 Results and Analysis

We report results for all experiments in Table 5. For both datasets, **(1)** TAPT and DAPT and **(2)** data augmentation with machine translations improve model performance. These strategies, which require no additional labeled data, improve macro-F1 score by between 0.6-1.5% for CLS and between 0.3-4.3% for HATEVAL. Even without DAPT, which is often the most expensive step, applying TAPT and/or data augmentation alone improves performance in all settings and languages except HATEVAL EN.

CLS For languages where extremely high-resource monolingual LMs are available (EN and FR), models perform best in the *mono-target* setting, in which a monolingual LM is fine-tuned on target language data. This is consistent with prior findings that XLM-R suffers from fixed model capacity and vocabulary dilution (Conneau et al., 2019). However, for DE and JA, which are not low-resource languages but whose monolingual LM pretraining corpora are relatively limited in size and domain (see Table 1), XLM-R models perform better.

HATEVAL On average, XLM-R models perform better on HATEVAL than those fine-tuned from monolingual LMs. Unlike for CLS, this is true even in EN, suggesting that for some classification tasks, the LM pretraining corpus is not as important for downstream task performance as XLM-R's larger model capacity and cross-lingual transfer. Though scores were much higher for the relabeled EN dataset than the original, the effects of LM fine-tuning, TAPT, DAPT, and data augmentation were consistent.

5.1 Not All Classification Tasks Are Created Equal

The two text classification tasks we evaluate are significantly different from both an annotation and a modeling perspective. Sentiment is a well-defined facet of language, and language model representations have even been shown to encode semantic information about it (Radford et al., 2017). Meanwhile, defining and identifying hate speech is much more nuanced, even for humans. Hate speech detection is confounded by many factors that require not only immediate context of the input but also cultural and social contexts (Schmidt and Wiegand, 2017). The difference in the types of information that models need to encode for each task may explain why monolingual LMs, which tend to encode better lexical information than multilingual LMs (Vulić et al., 2020), can outperform XLM-based models when fine-tuned for sentiment analysis but not for hate speech detection.

5.2 Cross-lingual Transfer

Prior work has established that multilingual LMs benefit from the addition of more languages during pretraining up to a point, after which limited model capacity and vocabulary dilution cause performance to degrade on downstream tasks – this is referred to as the *curse of multilinguality* (Conneau et al., 2019). Though this is reflected in the results of CLS EN and FR, other models fine-tuned from XLM-R exhibit gains from cross-lingual transfer. In particular, for CLS JA and HATEVAL EN, the best-performing models benefit not only from multilingual pretraining corpora but also from multilingual task training data.

These results suggest that when fine-tuning LMs for downstream tasks, XLM-R is a robust baseline.

Model	Adapt.	Aug.	CLS					HATEVAL				
			EN	DE	FR	JA	AVG	EN	EN†	ES	AVG	AVG†
mono-target												
RoBERTa (EN) BERT (OTHERS)	×	×	$94.7_{0.4}$	$90.9_{0.6}$	$95.2_{0.0}$	$88.7_{0.3}$	92.4	$44.4_{5.3}$	$58.5_{6.2}$	$75.6_{0.6}$	60.0	67.1
		✓	$\mathbf{95.3}_{0.3}$	$92.0_{0.2}$	$95.6_{0.3}$	$89.3_{0.02}$	93.0	$46.1_{2.6}$	$60.6_{3.2}$	$76.0_{1.7}$	61.0	68.3
	TAPT	×	$94.9_{0.1}$	$91.6_{0.1}$	$95.4_{0.1}$	$89.3_{0.3}$	92.8	$45.4_{1.9}$	$59.9_{2.7}$	$76.1_{1.1}$	60.8	68.0
		✓	$95.0_{0.4}$	$92.3_{0.4}$	$95.8_{0.2}$	$89.7_{0.4}$	93.2	$44.7_{1.5}$	$59.2_{1.7}$	$76.9_{1.4}$	60.8	68.0
	TAPT+	×	$94.9_{0.4}$	$91.8_{0.2}$	$95.5_{0.3}$	$89.5_{0.2}$	92.9	$48.0_{1.5}$	$63.1_{2.6}$	$76.3_{1.1}$	62.2	69.7
	DAPT	✓	$\mathbf{95.3}_{0.1}$	$93.0_{0.8}$	$\mathbf{95.9}_{0.1}$	$89.9_{0.4}$	**93.5**	$46.0_{4.3}$	$60.2_{4.4}$	$76.9_{0.6}$	61.4	68.5
multi-target												
XLM-RoBERTa	×	×	$92.5_{0.4}$	$93.0_{0.2}$	$92.5_{0.3}$	$90.4_{0.5}$	92.1	$47.2_{2.0}$	$61.4_{1.9}$	$74.8_{0.5}$	61.0	68.1
		✓	$93.3_{0.1}$	$94.0_{0.2}$	$93.8_{0.2}$	$90.3_{0.3}$	92.8	$45.6_{1.6}$	$59.3_{2.5}$	$77.0_{1.1}$	61.3	68.1
	TAPT	×	$92.7_{0.5}$	$93.5_{0.5}$	$93.9_{0.3}$	$90.3_{0.1}$	92.6	$47.0_{2.7}$	$62.4_{3.3}$	$76.1_{1.4}$	61.6	69.2
		✓	$93.4_{0.6}$	$94.0_{0.3}$	$93.8_{0.5}$	$90.5_{0.4}$	92.9	$47.9_{1.3}$	$63.5_{1.5}$	$77.9_{0.9}$	62.9	70.7
	TAPT+	×	$93.1_{0.6}$	$93.0_{0.5}$	$93.6_{0.1}$	$90.8_{0.3}$	92.6	$49.9_{2.5}$	$65.6_{2.4}$	$76.5_{1.0}$	63.2	71.0
	DAPT	✓	$94.0_{0.3}$	$\mathbf{94.1}_{0.4}$	$93.8_{0.3}$	$91.1_{0.4}$	93.2	$46.6_{2.1}$	$61.7_{2.5}$	$\mathbf{78.1}_{0.8}$	62.3	69.9
multi-all												
XLM-RoBERTa	×	×	$92.4_{0.3}$	$92.6_{0.4}$	$93.3_{0.4}$	$90.4_{0.4}$	92.2	$48.4_{3.5}$	$63.1_{4.5}$	$77.5_{0.4}$	62.9	70.3
		✓	$93.4_{0.3}$	$93.3_{0.2}$	$94.0_{0.2}$	$90.4_{0.5}$	92.8	$49.8_{3.5}$	$66.0_{4.6}$	$77.8_{0.9}$	63.8	71.9
	TAPT	×	$92.5_{0.4}$	$93.0_{0.3}$	$93.9_{0.3}$	$90.9_{0.3}$	92.6	$48.4_{2.7}$	$64.2_{3.5}$	$77.4_{0.9}$	62.9	70.8
		✓	$93.5_{0.4}$	$93.4_{0.5}$	$94.1_{0.2}$	$91.1_{0.2}$	93.0	$50.0_{2.2}$	$66.5_{2.6}$	$77.8_{0.6}$	63.9	72.2
	TAPT+	×	$92.7_{0.3}$	$93.3_{0.2}$	$94.0_{0.3}$	$91.2_{0.3}$	92.8	$47.1_{3.9}$	$62.7_{5.3}$	$77.4_{1.0}$	62.3	70.1
	DAPT	✓	$93.5_{0.3}$	$93.8_{0.2}$	$94.3_{0.3}$	$\mathbf{91.4}_{0.2}$	93.3	$\mathbf{50.7}_{1.1}$	$\mathbf{67.4}_{1.4}$	$77.7_{0.7}$	**64.2**	**72.6**

Table 5: CLS and HATEVAL results (macro-F1) averaged over five random seeds. The best results for each target language test set are **bolded**, and standard deviations are shown in subscripts. **Model** denotes the supervised fine-tuning setting. **Adapt.** denotes the adaptive pretraining setting: × (no adaptive pretraining), TAPT (task-adaptation only), or TAPT+DAPT (task- and domain-adaptation). **Aug.** denotes whether the training data was augmented with machine-translated examples. For HATEVAL, we report results for both the original and relabeled† test sets.

Model	Data	DE	FR	JA	ES
multi-target	target	94.1	93.8	91.1	78.1
multi-all	all	93.8	94.3	91.4	77.7
zero-shot	EN	92.7	92.6	88.5	72.1

Table 6: Zero-shot learning versus best multilingual approaches. *Data* denotes language of training data. We fine-tune XLM-R and use DAPT, TAPT, and data augmentation for all models shown.

In cases where knowledge transfer from a monolingual LM might be difficult (e.g. due to a limited pretraining corpus or specialized downstream task), XLM-R may even outperform its monolingual competitors.

5.3 Are Target Language Labels Needed?

Zero-shot learning is a topic of significant interest in multilingual NLU research (Conneau et al., 2018, 2019; Artetxe and Schwenk, 2019). In this context, we use *zero-shot learning* to refer to learning a classification task without observing training examples in the target language. Such an approach would allow practitioners to train a classification model using labeled data in a high-resource language such as EN and deploy it in other languages for which labels are not available.

To evaluate the viability of zero-shot approaches for our tasks, we compare the best performing models from the experiments in Table 5 with models trained only on EN training data. We report the test set results for each of the non-EN target languages in Table 6. Zero-shot models are competitive with previously published baselines (Table 4), which demonstrates the effectiveness of cross-lingual transfer in models like XLM-R. However, models trained using target language labels still outperform them by a large margin. Since obtaining a small number of target language labels is straightforward and typically required for validation in real applications, the need for zero-shot learning is reduced in practical scenarios.

5.4 Speed and Memory Usage

The deployment of multilingual NLU systems varies significantly depending on the number of downstream task models trained and the model architectures used. For instance, the *mono-target* and *multi-target* settings induce one model per target

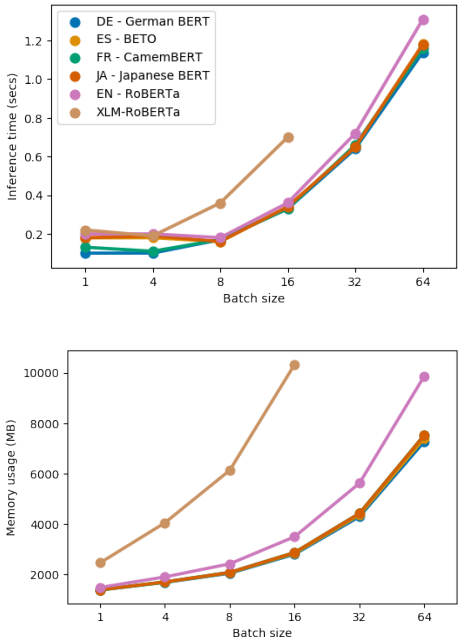

Figure 1: Inference time (top) and memory usage (bottom) benchmarks. XLM-R results not shown at batch sizes 32 and 64 due to GPU memory restraints. Environment details: `transformers v3.1.0, PyTorch v1.4.0, python v3.7.4, Linux. CPU: x86_64 (fp16=False, RAM=15GB). GPU: Tesla P100-PCIE-16GB, RAM=16GB, power=250.0W, perf. state=0).`

language. Conversely, *multi-all* models have more consistent end-task performance and do not require the added complexity and latency of language detection.

We use the Hugging Face library to benchmark the pretrained transformer models used in our experiments. We measure the inference time and memory usage of a single forward pass on a single Nvidia Tesla P100 GPU. Results are shown in Figure 1.

Monolingual BERT models in different languages are nearly identical in inference speed, but vary slightly at small batch sizes. RoBERTa has more parameters than BERT, but the impact on inference time and memory is small. XLM-R is also comparable with monolingual models at small batch sizes, but its memory usage becomes prohibitively large at batch sizes larger than 32. For certain applications such as those with real-time inference, this may not be important since the most common batch size is 1. Overall, the main tradeoff we observe is between the complexity of deploying N language-specific models and the high parameter count of a single multilingual model.

6 Related Work

6.1 Multilingual Classification Benchmarks

XNLI (Conneau et al., 2018) and PAWS-X (Yang et al., 2019) are commonly used as representative benchmarks for cross-lingual text classification (Hu et al., 2020; Conneau et al., 2019). However, both datasets are designed for evaluating zero-shot cross-lingual transfer. While useful, they do not reflect practical scenarios where (1) a small amount of labeled data obviates zero-shot approaches, and (2) target language test data are not semantically aligned.

Meanwhile, benchmarks for supervised multilingual text classification are limited. Artetxe and Schwenk (2019) propose Language-Agnostic SEntence Representations (LASER) and evaluate them on Multilingual Document Classification Corpus (MLDoc; Schwenk and Li, 2018). Eisenschlos et al. (2019) later show that their multilingual fine-tuning and bootstrapping approach, MultiFit, outperforms LASER and mBERT on CLS and MLDoc. The recently released Multilingual Amazon Reviews Corpus (MARC; Keung et al., 2020) is similar to CLS, but contains a different set of languages and large-scale training sets. Rust et al. (2020) perform a systematic evaluation similar to ours, comparing monolingual and multilingual BERT models on seven monolingual sentiment analysis datasets. Unlike our work, they do not consider multilingual test sets or cross-lingual transfer during training (as in the *multi-all* setting). None of the above evaluate practical training modifications, XLM-R, or tasks with class imbalance.

6.2 Hate Speech Detection

Due to the increased volume and consequence of online content moderation in recent years, there is a growing body of work on multilingual hate speech data and methodology. The Multilingual Toxic Comment Classification Kaggle challenge (Jigsaw, 2019) included a multilingual test set of Wikipedia talk page comments annotated for toxicity. More recently, Glavaš et al. (2020) introduced XHATE-999, an evaluation set of 999 semantically aligned test instances annotated for abusive language in five typologically diverse languages. Similar to our work, they compare state-of-the-art monolingual and multilingual transformer models. However, both the Jigsaw dataset and XHATE-999 are designed for evaluating zero-shot transfer and do not contain multilingual training data.

Other multilingual hate speech studies have largely combined separate existing monolingual datasets for evaluation (Pamungkas and Patti, 2019; Sohn and Lee, 2019; Aluru et al., 2020; Corazza et al., 2020; Zampieri et al., 2020). To avoid domain mismatch effects across languages, we use the HATEVAL dataset (Basile et al., 2019), for which all examples were collected simultaneously.

Previously evaluated approaches include LSTM architectures and feature selection (Pamungkas and Patti, 2019; Corazza et al., 2020), as well as using transformers for fine-tuning (Sohn and Lee, 2019) or feature extraction (Stappen et al., 2020). Aluru et al. (2020) show that fine-tuning from transformer-based language models generally outperforms other methods, including cross-lingual fixed representations like LASER.

7 Conclusion

We conduct an empirical evaluation of transformer-based methods for multilingual text classification in a variety of pretraining and fine-tuning settings. We evaluate our results on two multilingual datasets spanning five languages: CLS (sentiment analysis) and HATEVAL (hate speech detection). Additionally, we contribute a relabeled version of HATEVAL to address mislabeled test examples and enable meaningful comparisons in future work.

Our results and analysis show that practical methods such as task- and domain-adaptive pretraining and data augmentation using machine translations consistently improve model performance without requiring additional labeled data. We further show that multilingual model performance can vary based on task semantics, and that monolingual models are not always guaranteed to outperform massively multilingual models like XLM-R due to its large pretraining corpora and increased capacity.

Our work points to a number of future directions, including cross-domain and cross-task transfer, low-resource and few-shot learning, and practical alternatives to large multilingual models such as distillation.

Acknowledgements

We wish to thank Boya (Emma) Peng, Alexander Wang, and Thomas Boser for discussions and feedback on this work. Thanks also to the anonymous reviewers whose detailed suggestions helped improve its clarity and usefulness.

References

Sai Saket Aluru, Binny Mathew, Punyajoy Saha, and Animesh Mukherjee. 2020. Deep learning models for multilingual hate speech detection. *arXiv preprint arXiv:2004.06465*.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. AraBERT: Transformer-based model for Arabic language understanding. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 9–15, Marseille, France. European Language Resource Association.

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

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.

José Cañete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. 2020. Spanish pre-trained bert model and evaluation data. In *PML4DC at ICLR 2020*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

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.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Michele Corazza, Stefano Menini, Elena Cabrio, Sara Tonelli, and Serena Villata. 2020. A multilingual evaluation for online hate speech detection. *ACM Trans. Internet Technol.*, 20(2).

deepset.ai. 2019. Open sourcing german bert. https://deepset.ai/german-bert.

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.

Julian Eisenschlos, Sebastian Ruder, Piotr Czapla, Marcin Kadras, Sylvain Gugger, and Jeremy Howard. 2019. Multifit: Efficient multi-lingual language model fine-tuning. 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 5706–5711.

Goran Glavaš, Mladen Karan, and Ivan Vulić. 2020. XHate-999: Analyzing and detecting abusive language across domains and languages. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6350–6365, Barcelona, Spain (Online). International Committee on Computational Linguistics.

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 *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360. Association for Computational Linguistics.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization. *CoRR*, abs/2003.11080.

Jigsaw. 2019. Jigsaw multilingual toxic comment classification. https://www.kaggle.com/c/jigsaw-multilingual-toxic-comment-classification.

Phillip Keung, Yichao Lu, György Szarvas, and Noah A. Smith. 2020. The multilingual Amazon reviews corpus. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4563–4568, Online. Association for Computational Linguistics.

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 preprint arXiv:1907.11692*.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. CamemBERT: a tasty French language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online. Association for Computational Linguistics.

Endang Wahyu Pamungkas and Viviana Patti. 2019. Cross-domain and cross-lingual abusive language detection: A hybrid approach with deep learning and a multilingual lexicon. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 363–370, Florence, Italy. Association for Computational Linguistics.

Peter Prettenhofer and Benno Stein. 2010. Cross-language text classification using structural correspondence learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1118–1127.

Alec Radford, Rafal Józefowicz, and Ilya Sutskever. 2017. Learning to generate reviews and discovering sentiment. *CoRR*, abs/1704.01444.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Phillip Rust, Jonas Pfeiffer, Ivan Vulić, Sebastian Ruder, and Iryna Gurevych. 2020. How good is your tokenizer? on the monolingual performance of multilingual language models.

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10, Valencia, Spain. Association for Computational Linguistics.

Holger Schwenk and Xian Li. 2018. A Corpus for Multilingual Document Classification in Eight Languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Jasdeep Singh, Bryan McCann, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. 2019. Xlda: Cross-lingual data augmentation for natural language inference and question answering.

Hajung Sohn and Hyunju Lee. 2019. Mc-bert4hate: Hate speech detection using multi-channel bert for different languages and translations. *2019 International Conference on Data Mining Workshops (ICDMW)*, pages 551–559.

Lukas Stappen, Fabian Brunn, and B. Schuller. 2020. Cross-lingual zero- and few-shot hate speech detection utilising frozen transformer language models and axel. *ArXiv*, abs/2004.13850.

Masatoshi Suzuki and Ryo Takahashi. 2019. Pretrained japanese bert models. https://github.com/cl-tohoku/bert-japanese.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Antti Virtanen, Jenna Kanerva, Rami Ilo, Jouni Luoma, Juhani Luotolahti, Tapio Salakoski, Filip Ginter, and Sampo Pyysalo. 2019. Multilingual is not enough: Bert for finnish. *arXiv preprint arXiv:1912.07076*.

Ivan Vulić, Edoardo Maria Ponti, Robert Litschko, Goran Glavaš, and Anna Korhonen. 2020. Probing pretrained language models for lexical semantics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7222–7240.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019a. Superglue: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems*, volume 32, pages 3266–3280. Curran Associates, Inc.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019b. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In the Proceedings of ICLR.

Thomas Wolf, Julien Chaumond, Lysandre Debut, Victor Sanh, Clement Delangue, Anthony Moi, Pierric Cistac, Morgan Funtowicz, Joe Davison, Sam Shleifer, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45.

Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. 2019. PAWS-X: A cross-lingual adversarial dataset for paraphrase identification. In *Proceedings of EMNLP 2019*, pages 3685–3690.

Marcos Zampieri, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis, and Çağrı Çöltekin. 2020. SemEval-2020 task 12: Multilingual offensive language identification in social media (OffensEval 2020). In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1425–1447, Barcelona (online). International Committee for Computational Linguistics.