Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 6548–6555
Marseille, 20-25 June 2022
© European Language Resources Association (ELRA), licensed under CC-BY-NC-4.0

Improving Large-scale Language Models and Resources for Filipino

Jan Christian Blaise Cruz and **Charibeth Cheng**
Center for Language Technologies (CeLT), De La Salle University, Manila
2401 Taft Ave., Malate, Manila, Philippines
{jan_christian_cruz,charibeth.cheng}@dlsu.edu.ph

Abstract

In this paper, we improve on existing language resources for the low-resource Filipino language in two ways. First, we outline the construction of the TLUnified dataset, a large-scale pretraining corpus that serves as an improvement over smaller existing pretraining datasets for the language in terms of scale and topic variety. Second, we pretrain new Transformer language models following the RoBERTa pretraining technique to supplant existing models trained with small corpora. Our new RoBERTa models show significant improvements over existing Filipino models in three benchmark datasets with an average gain of 4.47% test accuracy across three classification tasks with varying difficulty.

Keywords: Transformers, Corpus Creation, Benchmarking

1. Introduction

Unlike High-resource Languages (HRL) such as English, German, and French, Low-resource Languages (LRL) suffer from a lack of benchmark datasets, databases, linguistic tools, and pretrained models that impede the progress of research within those languages.

Despite the growing success of methods that intrinsically learn from little data (Deng et al., 2020; Lee et al., 2021), creating “more data” remains a very significant fundamental task in NLP. Given the data-hungry nature of the neural networks that are prevalent in NLP today, creating new datasets to train from is the most efficient way to improve model performance. In addition, cleverly-constructed datasets also reveal new insights into the models we commonly use, letting us gauge their true performance and expose hidden weaknesses (Maudslay and Cotterell, 2021).

In this paper, we improve upon the existing resources for Filipino, a low-resource language spoken in the Philippines. We create a larger, more topically-varied large-scale pretraining dataset that improves upon the existing WikiText-TL-39 (Cruz and Cheng, 2019) that is too small and too topically-narrow to create robust models that perform well in modern NLP. We also produce new RoBERTa pretrained models using our pretraining dataset that supplant existing models trained with less data (Cruz and Cheng, 2020).

2. Resource Creation

In this section, we outline our full methodology for resource creation. First, we introduce the construction of our new large-scale pretraining dataset. Next, we detail the pretraining steps for our new RoBERTa models. Lastly, we introduce the task datasets that we use to benchmark performance for our new pretrained models.

2.1. The TLUnified Dataset

To effectively pretrain a large transformer for downstream tasks, we require an equally large pretraining corpus of high-quality Filipino text. We construct our pretraining corpus by combining a number of available Filipino corpora, including:

- **Bilingual Text Data** – Bitext datasets are used for training Machine Translation models and contain crawled and aligned data from multiple sources. We collected multiple bitexts, extracted Filipino text, then deduplicated the extracted data to add to our pretraining corpus. Datasets we collected from include bible-uedin (Christodouloupoulos and Steedman, 2015), CCAligned (El-Kishky et al., 2020), ELRC 2922[1], MultiCCAligned (El-Kishky et al., 2020),ParaCrawl [2], TED2020 (Reimers and Gurevych, 2020), WikiMatrix (Schwenk et al., 2019), tico-19, Ubuntu, OpenSubtitles, QED, Tanzil, Tatoeba, GlobalVoices, KDE4, and WikiMedia (Tiedemann, 2012).

- **OSCAR** – The Open Super-Large Crawled Aggregated Corpus (OSCAR) (Ortiz Suárez et al., 2019) is a massive dataset obtained from language identification and filtering of the Common Crawl dataset. We use the deduplicated version of the Filipino (Tagalog) portion of OSCAR and add it to our pretraining corpus.

- **NewsPH** – The NewsPH (Cruz et al., 2021) corpus is a large-scale crawled corpus of Filipino news articles, originally used in automatically creating the NewsPH-NLI benchmark dataset. Since we plan on using an NLI dataset derived from NewsPH for benchmarking in this paper, we opted to only use a 60% subset of the NewsPH corpus to add to TLUnified.

[1] https://elrc-share.eu/

[2] https://www.paracrawl.eu/

Since a large portion of our corpus is crawled and artificially aligned, we expect that out-of-the-box data quality would be low. To clean our dataset, we apply a number of preprocessing filters to it, including:

1. Non-latin Filter – We filter out sentences whose characters are composed of more than 15% non-latin letters.

2. Length Filter – We remove sentences that have a number of tokens N where $4 <= N <= 150$.

3. Puncutation Filter – All sentences that have tokens composed of too many succeeding punctuations (eg. "///") are all removed.

4. Average Word Length Filter – If a sentence has tokens that are significantly longer than the other tokens in the sentence, we remove the sentence entirely. We first take the sum of the character lengths of each token, then divide it by the number of tokens to get a ratio r. Only sentences with ratio $3 <= r <= 18$ are kept in the corpus.

5. HTML Filter – All sentences with HTML and URL-related tokens (e.g. ".com" or "http://") are removed.

After filtering the dataset, we perform one additional deduplication step to ensure that no identical lines are found in the dataset. The final result is a large-scale pretraining dataset we call **TLUnified**.

We then train tokenizers using TLUnified, limiting our vocabulary to a fixed 32,000 BPE subwords (Sennrich et al., 2015). Our tokenizers are trained with a character coverage of 1.0. We also do not remove casing to ensure that capitalization is kept after tokenization.

2.2. Pretraining

We then pretrain transformer language models that can serve as bases for a variety of downstream tasks later on. For this purpose, we use the RoBERTa (Liu et al., 2019) pretraining technique. Previous pretrained transformers in Filipino (Cruz and Cheng, 2020; Cruz et al., 2021) used BERT (Devlin et al., 2018), and ELECTRA (Clark et al., 2020) as their method of choice.

We choose RoBERTa as it retains state-of-the-art performance on multiple NLP tasks while keeping its pretraining task simple unlike methods such as ELECTRA. As a reproduction study of BERT, RoBERTa optimizes and builds up on the BERT pretraining scheme to improve training efficiency and downstream performance.

Two size variants are trained in this study following the original RoBERTa paper: a Base model (110M parameters) and a Large model (330M parameters). Both size variants use the same BPE tokenizer trained with TLUnified. Our hyperparameter choices also follow the original RoBERTa paper closely. A summary of our models' hyperparameters can be found in Table 1.

	Base	Large
Hidden Size	768	1024
Feedforward Size	3072	4096
Max Sequence Length	512	512
Attention Heads	12	16
Hidden Layers	12	24
Droput	0.1	0.1

Table 1: Base and Large RoBERTa hyperparameters.

During training, we construct batches by continually filling them with tokens until we reach a maximum batch size of 8192 tokens. Both variants are trained using the Adafactor (Shazeer and Stern, 2018) optimizer with $\beta_2 = 0.98$ and a weight decay of 0.01. The base model is trained for 100,000 steps with a learning rate of 6e-4, while the large variant is trained for 300,000 steps with a learning rate of 4e-4. We also use a learning rate schedule that linearly warms up for 25,000 steps, then linearly decays for the rest of training. All experiments are done on a server with 8x NVIDIA Tesla P100 GPUs.

3. Experiments

3.1. Benchmark Datasets

We test the efficacy of our RoBERTa models on three Filipino benchmark datasets:

- **Filipino Hatespeech Dataset** – 10,000 tweets labelled as "hate" and "non-hate" collected during the 2016 Philippine Presidential Elections. Originally published in Cabasag et al. (2019) and benchmarked with modern Transformers in Cruz and Cheng (2020).

- **Filipino Dengue Dataset** – Low-resource multiclass classification dataset with 4000 samples that can be one or many of five labels. Originally published in Livelo and Cheng (2018) and benchmarked in Cruz and Cheng (2020) using pretrained Transformers.

- **NewsPH-NLI** – An automatically-generated dataset constructed by exploiting the "inverted-pyramid" structure of news articles, causing every sentence to naturally entail the sentence that came before it. Originally created in Cruz et al. (2021).

For this study, we **do not use the original NewsPH-NLI** created in Cruz et al. (2021) as it has significant overlap with the subset of the NewsPH corpus that we used for pretraining. We instead re-generated a version of NewsPH-NLI (which we call "NewsPH-NLI Medium") using 40% of the NewsPH corpus, using the other 60% as part of the TLUnified pretraining data. This ensures that no test data is present in the training data, which will significantly inflate the benchmark scores.

Preprocessing for the downstream benchmark datasets is kept simple and non-destructive to preserve the linguistic structures and information present in the original data.

For the Hatespeech and the Dengue datasets, we follow the preprocessing used in Cruz and Cheng (2020), with a number of changes. Since both are datasets composed mainly of tweet data, the following preprocessing steps are done:

- Moses detokenization (Koehn and Hoang, 2010) was applied on all Moses-tokenized text.
- All HTML meta text and link texts are collapsed into a special `[LINK]` token. This is to reduce the noise in the dataset as images in the tweets are naturally converted into links.
- All substrings that start with an `@` character that are greater than length 1 are automatically treated as a “mention” and are replaced with a `[MENTION]` special token.
- All substrings that start with a # character that are greater than length 1 are automatically treated as a “hashtag” and are replaced with a `[HASHTAG]` special token.
- We renormalize apostrophes (e.g. `it ’s` → `it’s`) and punctuation that were spaced out (e.g. `one - two` → `one-two`) during the preprocessing in the Cruz and Cheng (2020) paper.
- Characters that were converted into unicode (e.g. `&amp;`) are converted back into their encoded form (e.g. `&`).

For the Dengue dataset, we transform the multilabel, multiclass classification setup into a multiclass classification problem by concatenating an example’s labels and converting the resulting binary number into an integer. For example, a sentence with the labels `1, 1, 0, 1, 1` for `absent`, `dengue`, `healthclasses`, `mosquito`, and `sick` will be converted into `27` (`11011` → `27`). This results in 32 possible labels and increases the difficulty of the task.

For the NewsPH-NLI Medium dataset, we opted to not do any further preprocessing as the released data from Cruz et al. (2021) is already preprocessed and clean.

Sample preprocessed data from the Hatespeech, Dengue, and NewsPH-NLI Medium datasets can be found in Figure 1.

Hatespeech Dataset

BREAKING: VCM Inside Novotel Cubao owned by Mar Roxas [LINK].
LABEL: NOT HATE

RT: [MENTION] : Sa laki ng ginastos ni Binay tapos sa laki din ng talo niya sa Mayo ,sya pa din tameme sa ending ng kwento. Yun na! [LINK].
LABEL: HATE

Dengue Dataset

Ang sama ng pakiramdam ko ? kung pwede lang um-absent bukas ee ?
LABEL: 21

Di ako nagfan ngayong gabi , ni-midnight snack naman ako ng mga lamok!!! ! Hahahhahaa
LABEL: 2

NewsPH-NLI Medium Dataset

Premise: ”Dahil dito, gagamitin ng mga militanteng grupo sa kanilang kampanya ang #bantrumpPH bilang pagpapahayag ng pagtutol sa pagtungo sa bansa ni Trump at laban sa mga dikta nito sa bansa.”

Hypothesis: Wala aniyang kalaban-laban ang mga Pinoy sakaling sumiklab ang kaguluhan sa pagitan ng Amerika at ng North Korea.
LABEL: ENTAILMENT

Premise: Ito’y matapos niyang mapikon sa patutsada ng isang miyembro ng National Union of Students in the Philippines (NUSP) tungkol kanyang reaksiyon sa napipintong paglaya ni ex-Calauan Mayor Antonio Sanchez.

Hypothesis: Inamin din ng negosyante na wala siyang deed of sale at iba pang papeles na magpapatunay na nabili niya talaga ang 350-hektaryang lupain.
LABEL: CONTRADICTION

Figure 1: Examples from our preprocessed datasets. The top box contains examples from the Hatespeech dataset. The middle contains examples from the Dengue dataset. The bottom contains examples from the NewsPH-NLI Medium dataset.

	Base	Large
Max. Seq. Length	128	256
Learning Rate	2e-5	1e-5
Warmup Ratio	0.1	0.06

Table 2: Unique finetuning hyperparameters for Base and Large transformer variants.

3.2. Finetuning Setups

We then finetune for the downstream benchmark tasks using our pretrained RoBERTa models. Since the NewsPH-NLI version and the setup of the Dengue dataset task is different from the previous benchmarking paper, we also finetuned Tagalog BERT (Cruz and Cheng, 2019) and Tagalog ELECTRA (Cruz et al., 2021) to serve as baseline models against the new RoBERTa model.

All models are trained using the Adafactor (Shazeer and Stern, 2018) optimizer with a learning rate scheduler that linearly increases from zero after a ratio of steps-to-total-training-steps has reached, then linearly decays afterwards. We use a batch size of 32 sentences for all models and use a weight decay of 0.1. We opted to use a larger maximum sequence length for the Large RoBERTa models as it has more capacity due to its deeper encoder stack. Hyperparameters that are different between Base and Large variants of the pretrained Transformers used are found in Table 2.

We add the `[LINK]`, `[MENTION]`, and `[HASHTAG]` special tokens during finetuning for the Hatespeech and Dengue datasets to the vocabularies of the Transformers used, averaging the vectors of all subword embeddings in the embedding layer to serve as initialization for the three added tokens.

Despite our RoBERTa having a full maximum sequence length allowance of 512, we opted to use smaller maximum sequence lengths during finetuning. This speeds up training (approximately 4x for the Base models and 2x for the Large models) while losing zero information since no sentence or sentence pair in any task reaches 256 subwords in length.

All experiments are done on a server with 8x NVIDIA Tesla P100 GPUs.

3.3. Degradation Tests

Like in Cruz and Cheng (2020), we perform a number of Degradation Tests as a form of “stress test”. This test simulates training in low-data environments and aims to measure how much the performance gained from pretraining will degrade (and conversely, how much “performance is retained”) as the number of training examples diminish.

To perform a degradation test, we finetune a model using a smaller sample of a benchmark dataset, then test with the full test set. All our degradation test use three data percentages: 50%, 10%, and 1%. For finetuning, we use the same hyperparameters used during normal finetuning for the model tested

We use two main metrics for this experiment. First is **Accuracy Degradation (AD)**, which refers to the difference in accuracy between a model trained with 100% of data, and a model trained with a fraction of the data. Formally:

$$\mathrm{AD}_{p\%} = \mathrm{Acc}_{100\%} - \mathrm{Acc}_{p\%} \tag{1}$$

where $\mathrm{Acc}_{p\%}$ refers to the accuracy of the model trained with $p\%$ of data. Second, we also measure the **Degradation Percentage (DP)**, which measures how much performance from the full model is lost when we reduce the training data at a certain data percentage $p\%$. Formally:

$$\mathrm{DP}_{p\%} = \frac{\mathrm{AD}_{p\%}}{\mathrm{Acc}_{100\%}} \times 100 \tag{2}$$

where $\mathrm{AD}_{p\%}$ is the Accuracy Degradation of the model at a certain data percentage $p\%$.

In addition to these two metrics, we also report the **Degradation Speed (DS)** of the model, which is simply the average of the reported Degradation Percentages of a model for all tests done. In this case, this is the average of $\mathrm{DP}_{50\%}$, $\mathrm{DP}_{10\%}$, and $\mathrm{DP}_{1\%}$.

We first perform model comparative degradation tests using the Hatespeech dataset to compare our RoBERTa models with the BERT and ELECTRA models in terms of performance retention. Afterwhich, we perform size comparative degradation tests using all three benchmark datasets to compare RoBERTa Base and RoBERTa Large to identify differences in performance between size variants.

4. Results

4.1. Benchmark Results

We report the results for our finetuning for the three benchmark datasets in terms of validation and test accuracy. A summary of the results can be found on Table 3.

Our RoBERTa models outperformed both the BERT and the ELECTRA models across all tasks. For the Hatespeech task, RoBERTa Large outperformed the best previous model (BERT Base) by +4.07% test accuracy. RoBERTa large also had a gain in performance in the Dengue dataset (+5.3% test accuracy over BERT Base) and the NewsPH-NLI Medium dataset (+4.04% test accuracy over ELECTRA Base).

While marginally inferior to the Large variant, the Base RoBERTa variant still outperforms the baseline models in all tasks. RoBERTa Base has an improvement of +3.9% against BERT Base on the Hatespeech task, +4.4% against BERT Base on the Dengue task, and +3.95% against ELECTRA Base on the NewsPH-NLI Medium task.

The difference in performance between the Base and Large RoBERTa variants is marginal in the current benchmarks. Large outperforms Base only by +0.17% for Hatespeech, +0.9% for Dengue, and +0.09% for NewsPH-NLI Medium. We hypothesize that this is due to the size of the pretraining dataset. While the size of TLUnified is much larger than the previous WikiText-TL-39, it may still not be enough to make full use of the capacity of a Large-variant Transformer. We surmise that RoBERTa Large may need to be trained with more data to show significant, non-marginal improvements in performance.

Overall, our new models show significant improvements over older pretrained Filipino Transformer models. This is likely due to the improved pretraining corpus, with TLUnified being larger and of more varied topics and sources than the previous WikiText-TL-39.

4.2. Model-comparative Degradation Tests

We perform a degradation test using the Hatespeech dataset to measure the performance retention of the four transformer models when subjected to low-data setups. A summary of the results can be found on Table 4.

Overall, our RoBERTa Large model degrades the slowest with a degradation speed of 11.97. This is fol-

	Hatespeech		**Dengue**		**NewsPH-NLI Med.**	
Model	Val. Acc	Test Acc.	Val. Acc	Test Acc.	Val. Acc	Test Acc.
BERT Base Cased	74.79%	74.17%	77.20%	75.80%	88.38%	88.74%
ELECTRA Base Cased	74.91%	72.50%	74.00%	69.20%	90.94%	91.06%
RoBERTa Base	78.66%	78.07%	81.80%	80.20%	94.92%	95.01%
RoBERTa Large	**78.97%**	**78.24%**	**82.81%**	**81.10%**	**94.99%**	**95.10%**

Table 3: Finetuning results for all Transformer variants on the three benchmark datasets. Our RoBERTa models outperform both the BERT and ELECTRA models with the same number of parameters (Base variant). The Large RoBERTa model marginally outperforms the Base variant for all three benchmark datasets.

		Hatespeech			
Model	Data%	Test Acc	AD$_{p\%}$	DP$_{p\%}$	Degradation Speed
BERT Base	100%	74.17%			
	50%	71.93%	-2.24%	3.02%	
	10%	68.53%	-5.64%	7.60%	
	1%	52.95%	-21.22%	28.61%	13.08
ELECTRA Base	100%	72.50%			
	50%	70.94%	-1.56%	2.15%	
	10%	65.22%	-7.28%	10.04%	
	1%	54.44%	-18.06%	24.91%	12.37
RoBERTa Base	100%	78.07%			
	50%	76.32%	-1.75%	2.24%	
	10%	72.78%	-5.29%	6.78%	
	1%	56.26%	-21.81%	27.94%	12.32
RoBERTa Large	100%	78.24%			
	50%	75.78%	-2.46%	3.15%	
	10%	72.12%	-6.12%	7.82%	
	1%	58.72%	-19.52%	24.95%	11.97

Table 4: Degradation test results for the Hatespeech dataset. Our RoBERTa models' performance degrade slower compared to the BERT and ELECTRA models in simulated low-data setups. This is likely due to the improved pretraining dataset, giving RoBERTa better priors to work with in the absence of training data.

lowed by RoBERTa Base with 12.32, a marginal difference from the degradation speed of ELECTRA Base at 12.37. The highest degradation speed was from BERT Base at 13.08. We surmise that RoBERTa Large was able to maintain most of its performance due to its size and capacity compared to the Base models.

It is interesting to note that the speed between RoBERTa Base and ELECTRA Base is only marginally small (0.05). While RoBERTa Base's pretraining corpus (TLUnified) is much larger and more topically varied than ELECTRA Base's pretraining corpus, we surmise that ELECTRA's pretraining technique being more data-efficient allowed it to maintain more performance despite the gap in pretraining data volume. ELECTRA's pretraining method allows it to "see" more of the dataset as compared to RoBERTa, which only "sees" as much as the tokens used for masked language modeling.

BERT, having a similar pretraining mechanism with RoBERTa, expectedly performed worse than the other models due to the small size of its pretraining corpus.

In addition to the final degradation speed, we also look at the degradation percentage at each data percentage level (100%, 50%, 10%, and 1%). A graph of the model degradation percentages can be found on Figure 2.

At 50% training data, we can see that the Base transformers all exhibit the same relative amount of performance drop, but the Large RoBERTa model started degrading faster than the others. This may be due to the size of the model requiring more data to be effectively finetuned to comparable performance.

At 10% training data, the ELECTRA Base model speeds up significantly, dropping to 65.22% accuracy from its original 72.50%.

At the lowest data setups (1%), we note that the Large RoBERTa model is the slowest to degrade, likely owing to its inheretly large capacity compared to the other models. ELECTRA interestingly degrades to around the same amount as RoBERTa Large (24.91% DP$_{p\%}$ for ELECTRA Base and 24.95% DP$_{p\%}$ for RoBERTa Large.) BERT Base and RoBERTa Base both degrade much faster in the 1% data setup compared to the other models, the similar degradation amount likely caused by the similarity in pretraining method and model size.

We hypothesize that the ELECTRA Base model degrades slow in the extreme 1% setup, comparable to the RoBERTa Large model, again due to its pretraining

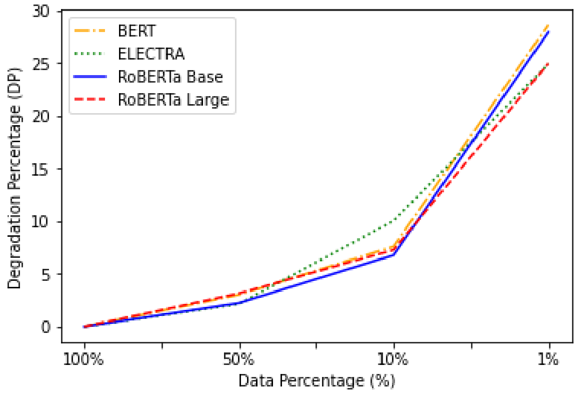

Figure 2: Degradation test results on the Hatespeech dataset plotted in four data percentages. At half data (50%), all four models degrade comparatively to each other, with ELECTRA speeding up at the 10% data setup. At 1% data, RoBERTa and BERT Base degrade fastest, while ELECTRA and RoBERTa Large degrades the slowest.

method. Being more data-efficient in pretraining, it is able to leverage better learned information when there is little data to learn from.

Overall, while the RoBERTa models have the slowest degradation speed out of all the models, transformers trained with other techniques, ELECTRA especially, may be more beneficial at different data volumes despite the smaller pretraining corpus used. If an abundance of data is available for a task, the best performing RoBERTa Large model may be leveraged for best performance. In the absence of such data, the other transformers may still be utilized considering data availability as well as model size constraints.

4.3. Size-comparative Degradation Tests

We also investigate more on the performance differences of the two RoBERTa variants, Base and Large, with varying amounts of training data via a size-comparative degradation test. For this purpose, we perform degradation tests using all three benchmark datasets using RoBERTa Base and RoBERTa Large. A summary of the results can be found on Table 5.

Result show that the Large variant degrades much slower compared to the Base variant for all tasks except for the Dengue dataset. RoBERTa Large has a degradation speed of 13.64, a full 2.38 points higher than RoBERTa Base's 11.26. We hypothesize that this is due to the inherently small size and large number of target labels of the Dengue dataset.

Given that RoBERTa large has more capacity, it may be the case that it requires more data to be finetuned effectively for small-data tasks. This corroborates the findings in the model-comparative degradation tests where RoBERTa Large is the first to degrade as the number of training examples diminish, only performing better once all the models are trained with extremely low amounts of data (1% setup).

To further investigate the difference between the two sizes, we plot the degradation percentage with respect to the data percentage for both RoBERTa Base and RoBERTa Large. This plot can be found in Figure 3.

It is interesting to note that the degradation percentage trend of the RoBERTa Base and RoBERTa Large models differ for the three benchmark datasets.

For the Hatespeech dataset, RoBERTa Base degrades slower at first, then faster then RoBERTa Large as the number of training examples approach 1%. This behavior has already been noted in the model-specific degradation tests above.

For the Dengue dataset, RoBERTa Large degrades much faster on all data percentages compared to RoBERTa Base. On the other hand, the degardation curves of RoBERTa Base and RoBERTa Large are very similar when tested on the NewsPH-NLI Medium dataset, in all data percentages.

From these observations, we hypothesize that degradation percentage is directly affected by the amount of training data provided, and that the degradation between size variants of the same model type will be more similar the more training data is provided. Conversely difference between degradation percentages between size variants only show when trained in setups with very little data.

NewsPH-NLI Medium is an inherently large task dataset compared to the other two, and even when only 1% of the training examples are provided, it still gives both models enough information to draw out good performance. This results in the Base and Large variant having very close degradation percentage curves.

While there is a difference in degradation percentages between the two size variants on the Hatespeech dataset, the overall degradation speed is close (12.32 for RoBERTa Base and 11.97 for RoBERTa Large, a marginal difference of 0.35). This shows that as we reach smaller data domains, the degradation percentages between different size variants of the same model will start to differentiate.

This difference is fully shown when we look at the results of the Dengue dataset. The Dengue dataset is a very small dataset to begin with, and as we further reduce the number of training examples to the extreme case of 1%, we see that the Large variant struggles to retain performance. This is likely due to the Large model needing more data to be effectively finetuned without risk of catastrophic forgetting.

Overall, while both variants of RoBERTa have slower degradation speeds compared to the previous BERT and ELECTRA models, there is a difference in their performance retention when it comes to how much data they encounter during training. In cases where is abundant data to train with, larger models may be employed for better direct performance. On the other hand, in cases of extreme data scarcity, it may be more beneficial to use smaller models.

RoBERTa Base					
Dataset	Data%	Test Acc	$\text{AD}_{p\%}$	$\text{DP}_{p\%}$	Degradation Speed
Hatespeech	100%	78.07%			
	50%	76.32%	-1.75%	2.24%	
	10%	72.78%	-5.29%	6.78%	
	1%	56.26%	-21.81%	27.94%	12.32
Dengue	100%	80.20%			
	50%	79.11%	-1.09%	1.36%	
	10%	74.27%	-5.93%	7.39%	
	1%	60.12%	-20.08%	25.04%	11.26
NewsPH-NLI Medium	100%	95.01%			
	50%	92.47%	-2.54%	2.67%	
	10%	87.93%	-7.08%	7.45%	
	1%	71.56%	-23.45%	24.68%	11.60

RoBERTa Large					
Dataset	Data%	Test Acc	$\text{AD}_{p\%}$	$\text{DP}_{p\%}$	Degradation Speed
Hatespeech	100%	78.24%			
	50%	75.78%	-2.46%	3.15%	
	10%	72.12%	-6.12%	7.82%	
	1%	58.72%	-19.52%	24.95%	11.97
Dengue	100%	81.10%			
	50%	79.61%	-1.49%	1.83%	
	10%	71.95%	-9.15%	11.28%	
	1%	58.55%	-22.55%	27.81%	13.64
NewsPH-NLI Medium	100%	95.10%			
	50%	93.02%	-2.08%	2.19%	
	10%	88.19%	-6.91%	7.27%	
	1%	71.95%	-23.15%	24.34%	11.27

Table 5: Comparative degradation test results for both sizes of RoBERTa on all three benchmark datasets. The large RoBERTa variant degrades slower compared to the Base variant in all benchmark tasks except for the Dengue dataset.

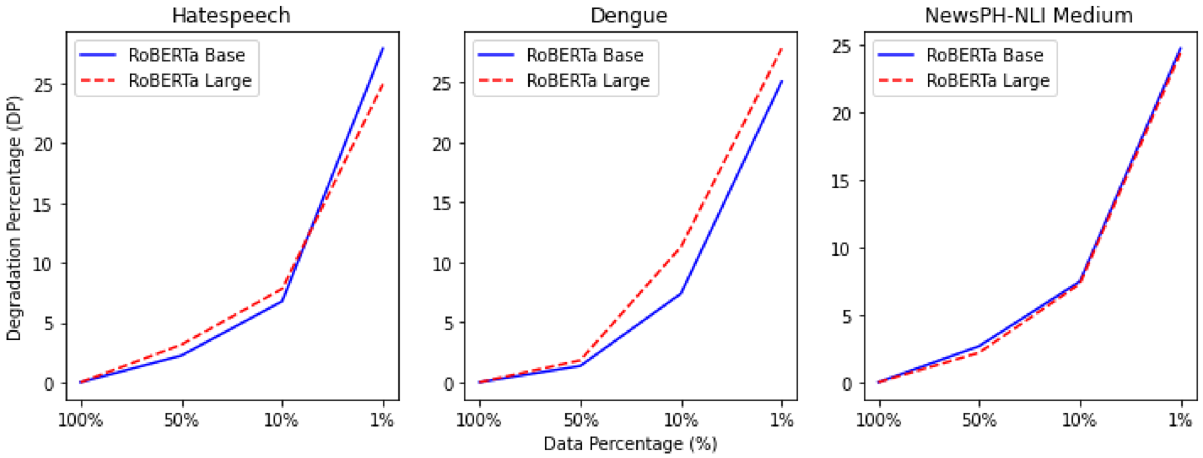

Figure 3: Comparison of degradation speeds between the Base and Large RoBERTa variants for all three benchmark datasets. Interestingly, results aren't consistent across model sizes and datasets: the Base variant degrades faster in Hatespeech and slower in Dengue. The Large variant degrades faster in Dengue and slower in Hatespeech. For the NewsPH-NLI Medium task, both models degrade similarly.

5. Conclusion

Our work has two main contributions in terms of language resources for the Filipino language. First, we construct TLUnified, a new large-scale pretraining corpus for Filipino. This is an improvement over the much smaller pretraining corpora currently available, boasting much larger scale and topic variety. Second, we release new pretrained Transformers using the RoBERTa pretraining method. Our new models outperform existing baselines on three different classification tasks, with significant improvements of +4.07%, +5.03%, and +4.04% test accuracy for the Hatespeech, Dengue, and NewsPH-NLI Medium datasets respectively.

6. Bibliographical References

Cabasag, N. V., Chan, V. R., Lim, S. C., Gonzales, M. E., and Cheng, C. (2019). Hate speech in philippine election-related tweets: Automatic detection and classification using natural language processing. *Philippine Computing Journal, XIV No*, 1.

Christodouloupoulos, C. and Steedman, M. (2015). A massively parallel corpus: the bible in 100 languages. *Language resources and evaluation*, 49(2):375–395.

Clark, K., Luong, M.-T., Le, Q. V., and Manning, C. D. (2020). Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

Cruz, J. C. B. and Cheng, C. (2019). Evaluating language model finetuning techniques for low-resource languages. *arXiv preprint arXiv:1907.00409*.

Cruz, J. C. B. and Cheng, C. (2020). Establishing baselines for text classification in low-resource languages. *arXiv preprint arXiv:2005.02068*.

Cruz, J. C. B., Resabal, J. K., Lin, J., Velasco, D. J., and Cheng, C. (2021). Exploiting news article structure for automatic corpus generation of entailment datasets. In *Pacific Rim International Conference on Artificial Intelligence*, pages 86–99. Springer.

Deng, S., Zhang, N., Sun, Z., Chen, J., and Chen, H. (2020). When low resource nlp meets unsupervised language model: Meta-pretraining then meta-learning for few-shot text classification (student abstract). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13773–13774.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

El-Kishky, A., Chaudhary, V., Guzmán, F., and Koehn, P. (2020). CCAligned: A massive collection of cross-lingual web-document pairs. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020)*, pages 5960–5969, Online, November. Association for Computational Linguistics.

Koehn, P. and Hoang, H. (2010). Moses. *Statistical Machine Translation System, User Manual and Code Guide*, page 245.

Lee, H.-y., Vu, N. T., and Li, S.-W. (2021). Meta learning and its applications to natural language processing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Tutorial Abstracts*, pages 15–20.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Livelo, E. D. and Cheng, C. (2018). Intelligent dengue infoveillance using gated recurrent neural learning and cross-label frequencies. In *2018 IEEE International Conference on Agents (ICA)*, pages 2–7. IEEE.

Maudslay, R. H. and Cotterell, R. (2021). Do syntactic probes probe syntax? experiments with jabberwocky probing. *arXiv preprint arXiv:2106.02559*.

Ortiz Suárez, P. J., Sagot, B., and Romary, L. (2019). Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures. Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7) 2019. Cardiff, 22nd July 2019, pages 9 – 16, Mannheim. Leibniz-Institut für Deutsche Sprache.

Reimers, N. and Gurevych, I. (2020). Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11.

Schwenk, H., Chaudhary, V., Sun, S., Gong, H., and Guzmán, F. (2019). Wikimatrix: Mining 135m parallel sentences in 1620 language pairs from wikipedia. *arXiv preprint arXiv:1907.05791*.

Sennrich, R., Haddow, B., and Birch, A. (2015). Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

Shazeer, N. and Stern, M. (2018). Adafactor: Adaptive learning rates with sublinear memory cost. In *International Conference on Machine Learning*, pages 4596–4604. PMLR.

Tiedemann, J. (2012). Parallel data, tools and interfaces in opus. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may. European Language Resources Association (ELRA).