United International Journal for Research & Technology

Volume 04, Issue 08, 2023 | Open Access | ISSN: 2582-6832

Enhancement to Low-Resource Text Classification via Sequential Transfer Learning

Neil Christian R. Riego[1], Danny Bell Villarba[2], Ariel Antwaun Rolando C. Sison[3], Fernandez C. Pineda[4], and Herminiño C. Lagunzad[5]

[1,2]Student, College of Engineering - Pamantasan ng Lungsod ng Maynila

[3,4,5]Professor, College of Engineering - Pamantasan ng Lungsod ng Maynila

Abstract— Textual data on many platforms has increased dramatically in recent years. With this amount of data, anyone may do text classification, such as sentiment analysis and hatespeech recognition. However, the lack of various NLP Tools in low-resource areas such as Asia and Africa limits its ability to be leveraged. We provided three (3) contributions. First, we provided a Tagalog product review dataset as a baseline for sentiment analysis tasks. Second, we pretrained and finetuned a Tagalog variation of XLNet in two datasets, reaching 78.05% accuracy in the hatespeech dataset and 95.02% in the shopee, which is 0.33% and 3.87% higher than the benchmark RoBERTa-tagalog model, respectively. Third, in the finetuning step, an improvement using bootstrap aggregation (bagging) is implemented, which boosts accuracy by 0.16% when 70% of the data is used in finetuning three XLNet-Tagalog models. Furthermore, combining RoBERTa-Tagalog and XLNet-Tagalog finetuned in 100% of data results in an accuracy of 79.47%, a 1.26% improvement over the best-performing setup using the XLNet-Tagalog. Finally, the XLNet Tagalog degrades slower than the benchmark model by 4.53. We make all our models and datasets available to the research community.

Keywords— Bagging-based Approach, Low-resource Language, NLP Tools, Natural Lanaguage Processing RoBERTa, Sentiment Analysis, Sequential Transfer Learning, Tagalog language, Textual data, Text classification, XLNet.

I. INTRODUCTION

1.1 Background of the Study

In recent years, there has been a significant increase in the amount of textual data on different platforms. As shown in Domo's [1] released infographic, from 2013 to 2022, for every minute in a day, there were an additional 247k tweets shared on Twitter, 1 million of content transmitted through Facebook, and 452 hours uploaded on Youtube. This textual data comprises personal opinions on specific topics, product reviews, and video comments which any individual can leverage in creating their social media strategy using sentiment analysis [2].

1.1.1 Text Data Mining

Text is any unit of meaning composed of sentences and paragraphs, which criteria can distinguish. Furthermore, based on Types of Text - Classification, characteristics and examples - Daily Concepts [3], it could be classified according to its purpose, area of interest, and physical support. In its raw form, textual data is hard to comprehend and use for analysis. In the book of Zong et al., "Text Mining" [4], The field of text mining involves several different technologies encompassing a variety of content. The term "mining" typically refers to the processes of "discovery, search, and induction." The goal results being sought are frequently not visible but hidden and concealed in the text or cannot be located and summarized in an extensive range since discovery and refining are necessary. Text classification, topic modeling, sentiment analysis, and opinion mining are some tasks under text data mining.

1.1.2 Sentiment Analysis and Text Classification

Sentiment analysis determines the emotional tone behind a text. It can identify the writer's attitude toward a particular topic and aims to determine whether a text's overall sentiment is positive, negative, or neutral [5]. Aside from social media strategy [2], sentiment analysis is also utilized in the Philippines context through consumer reviews [6], consumer satisfaction [7], customer recommendations [8], performance evaluation [9][10], news objectivity [11], Filipino tweets with backward slang [12], and Jejemon Slang [13].

1.1.3 Traditional NLP Tools

Nevertheless, there are problems with the different NLP models and techniques, like making and labeling datasets. Most researchers labeled and annotated data points by hand, except for Vader annotation [14] and POS tagging [13]. Also, there is a need for more available Filipino subjectivity lexicon, and Keen et al. [15] say that making one is needed to make lexicon-based sentiment analysis tasks even better. The scarcity of both of these NLP tools indicates low-resource NLP, which is hard to process and a big problem for rulebased and lexicon-based approaches and specific NLP tasks [16], as well as for supervised ML algorithms like Naive Bayes, Decision Trees, and Regression [9][7][13][14].

1.1.4 Transformer-Based NLP Tools

However, with the introduction of deep-learning algorithms, the traditional approach of supervised ML and lexicon-based sentiment analysis was proven to perform poorly compared to neural networks [12]. Neural networks are layers of interconnected "neurons" that process and transmit information through weights adjusted during training. As demonstrated by BERT [17], this produces state-of-the-art performance, a transformer-based pre-trained language model (PLM).

1.1.5 Low-Resource Languages

The lack of low-resource NLP tools is further amplified by low-resource languages. This limits the capability of getting insights from textual data using text mining. In an article by Laumann [18], low-resource languages are those that lack extensive monolingual or parallel corpora, as well as manually generated linguistic resources necessary for developing statistical NLP applications, where the majority of the languages spoken in Asia and Africa are considered as low-resource language.

1.2 Statement of the Problem

The following problems were identified in a low-resource NLP and on the transformer model used:

a) There is a scarcity of effective NLP tools for low-resource sentiment analysis (text classification).
b) Limitations in compatible language for pretrained language models are imminent.
c) There is a demand for resources for transformer-based language models.

1.3 Objective of the Study

The study aims to enhance low-resource text classification for the Filipino language. The following are the specific objectives to improve the model:

a) Build a comprehensive Filipino text classification dataset and identify its performance in terms of accuracy against a benchmark dataset for finetuning classification tasks.
b) Determine the improvement in accuracy after implementing bootstrap aggregation as an adaptation technique in sequential transfer learning.
c) Determine the performance of pretrained XLNet compared to language models pretrained in the same Filipino corpora when subjected to low resource setting using accuracy degradation, degradation percentage, and degradation speed.

II. RELATED WORKS

2.1 Lack of NLP Benchmarks Datasets

There are numerous approaches to sentiment analysis, with lexicon-based and machine-learning-based models being the most common. However, each of these models has issues with datasets.

2.1.1 Lexicon-based Datasets

Lexicon-based models rely on a corpus with each word labeled as positive, negative, or neutral; however, in a study conducted by Keen et al. [19], there is no existing subjectivity lexicon for the Filipino language, which makes it challenging to build a lexicon-based model. This is supported further by the study of Sagum et al. [20], wherein WordNet corpora used by the natural language toolkit (NLTK) are developed using forty languages, but still no variations for the Filipino language.

Even though lexicons like FICOBU and FilCon are getting better, the study by Kotelnikova et al. and Boquiren et al. [21][12] shows that deep learning models still do a better job of analyzing sentiment than lexicon-based models. However, due to the nature of these models, there is a need for large, high-quality training datasets.

2.1.2 Expertly Labeled Corpora

According to Cruz and Cheng [22], low-resource languages such as Filipino need more expert-annotated corpora that hinder the production of classifiers, which is further expanded by the same authors, Cruz and Cheng [23] by establishing baselines for text classification. Further advancement is presented in a recent study by Canon et al. [24] that focuses on creating a corpus for multi-domain Philippine English to support further the need for benchmark datasets for pre-trained language models (PLMs).

2.2 Low-resource NLP

According to Firsanova [16], low-resource natural language processing (NLP) primarily pertains to the application of machine learning techniques to languages that are generally considered to be "under-described." The scarcity of linguistic structures in such languages poses many challenges from both a linguistic and machine-learning perspective. It is worth noting, however, that not all languages that fall under the "well-described" classification are necessarily high-resource languages for NLP; certain conditions must be met for a language to be considered high-resource for NLP.

Recent research in natural language processing has identified challenges related to using the BERT model, specifically its preference for English, a high-resource language. A comprehensive examination of these difficulties was conducted in a study by Papadimitriou et al. [25]. It was found that while multilingual models have the potential to improve performance on low-resource languages by utilizing resources from higher-resource languages, they also have a negative effect of reducing overall performance across all languages, which is known as the "curse of multilinguality." Additionally, the study uncovered a novel issue with multilingual models, referred to as "grammatical structure bias," where grammatical structures from higher-resource languages bleed into lower-resource languages. The findings revealed that a multilingual BERT model exhibited a bias towards an English-like setting, as indicated by its preference for direct pronouns the model can treat compared to monolingual control models. The authors recommend using more linguistically-aware fluency evaluations in future research.

2.3 Transfer Learning in Large Language Models

Using large-scale language models gives a state-of-the-art performance but takes more hardware resources for training and inference. This is also a concern when considering low-resource equipment [23].

2.3.1 Sequential Transfer Learning

Transfer learning has emerged as a powerful approach in the field of Natural Language Processing (NLP), enabling models to leverage data from additional domains or tasks to improve generalization properties [26]. The classic supervised machine learning paradigm, which relies on learning in isolation, often requires many training examples and is more effective for well-defined and narrow tasks. In contrast, transfer learning methods extend this paradigm by capitalizing on the knowledge learned from related domains or tasks, leading to enhanced model performance and generalization.

Ruder et al. [26] provide a comprehensive overview of modern transfer learning methods in NLP. They highlighted the significant improvements achieved by these methods on a wide range of NLP tasks, paralleling the successes observed with pretrained word embeddings and ImageNet pretraining in computer vision. The authors discussed the pretraining of models and the information captured by the learned representations. They also presented examples and case studies showcasing these models' integration, sequencing, and adaptation in downstream NLP tasks.

By exploring transfer learning in NLP, this study contributes to understanding how sequential transfer learning can be leveraged to enhance low-resource text classification. The insights gained from Ruder et al. [26] and other relevant literature will inform the development of our approach, enabling us to effectively utilize pre-trained models and improve the accuracy and adaptability of low-resource text classifiers.

2.3.2 Fine Tuning

Transfer learning, mainly through finetuning, has been widely employed to address the challenge of limited datasets and improve the performance of deep learning models in various domains. In the study by Grega Vrbančič and Vili Podgorelec [27], the authors explored the application of transfer learning with adaptive finetuning in medical imaging, specifically for identifying osteosarcoma.

The authors emphasized the importance of having sufficient datasets with reliable ground truth, which are often difficult to obtain in medical applications. They proposed the Differential Evolution-based FineTuning (DEFT) method, which addresses the problem of selecting fine-tunable layers for a target dataset under specific constraints. By finetuning only distinct layers of a pre-trained model, they aimed to achieve better performance on the target task.

To evaluate the effectiveness of their approach, the authors compared the classification accuracy of their proposed DEFT method with a conventionally trained convolutional neural network, a pre-trained model, and a finetuning approach with manually selected fine-tunable layers. Their results demonstrated that the DEFT method outperforms the compared methods, achieving a margin of improvement in classification accuracy ranging from 4.45% to 32.75%.

Although the study by Vrbančič and Podgorelec [27] focuses on medical imaging, their exploration of transfer learning with finetuning offers valuable insights that can be applied to low-resource text classification. The challenges of limited datasets and the selection of fine-tunable layers are shared across domains, and the strategies proposed in this study can guide enhancing the effectiveness of sequential transfer learning in the context of your thesis.

2.3.3 Bootstrap Aggregating (Bagging)

Deep learning models have showcased remarkable performance in various text classification tasks, surpassing classical machine learning models. However, identifying the most appropriate deep learning classifier remains challenging in text classification. Ensemble learning techniques, such as bootstrap aggregation (bagging), have been explored to enhance performance, minimize errors, and mitigate overfitting.

An article by Mohammed and Kora [28] proposes an effective ensemble deep-learning framework for text classification. Their framework incorporates a meta-learning ensemble method that combines multiple baseline deep-learning models using two tiers of meta-classifiers. While the study does not explicitly mention bootstrap aggregation, the concept aligns with using ensemble learning to boost performance and mitigate overfitting

Ensemble methods, such as bagging, involve training multiple models on different subsets of the training data and combining their predictions through voting or averaging.

This approach leverages the diversity among individual models, reducing the impact of outliers and improving the overall classification accuracy. Although Mohammed and Kora [28] do not explicitly reference bagging, their ensemble framework likely incorporates elements of this technique to enhance the performance and robustness of the deep learning models.

Incorporating bagging within the sequential transfer learning framework in low-resource text classification can further enhance classification accuracy and generalization capabilities. By leveraging the strengths of multiple models trained on different subsets of the data, bagging can help mitigate the limitations of low-resource datasets and improve the overall performance of the text classification system.

While the article by Mohammed and Kora [28] does not explicitly discuss bootstrap aggregation (bagging), the proposed ensemble deep learning framework aligns with leveraging ensemble methods to enhance text classification. Incorporating bagging techniques within the sequential transfer learning approach can improve low-resource text classification systems by mitigating overfitting and enhancing overall classification accuracy.

III. METHODOLOGY

The proposed method for the enhancement is represented in Fig 3.1.

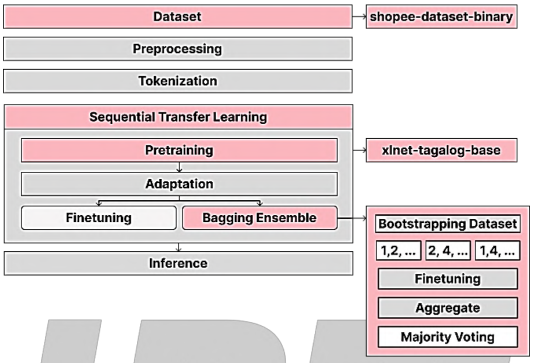

***Fig 3.1** Overview Diagram of the Methodology*

Figure 3.1 shows the summarized method of the study. It details the different contributions of the study. First is creating a shopee-dataset-binary dataset that directly addresses the study's problem 1, the lack of Filipino NLP benchmark datasets.

Second, preprocessing is essential to limit the noise in the collected dataset, while tokenization is essential to convert the dataset into a format that models can understand.

Third, sequential transfer learning will answer the study's problems 2 and 3. In the pretraining part, a Tagalog variation of XLNet will be trained in Tagalog Corpora and be adapted in text classification by finetuning and bagging methods. Lastly, the models will be tested to check their accuracy during inference.

3.1 Dataset Construction

A robust dataset is required to construct a model. In sequential transfer learning, pretraining requires large text corpora in the source task. While in the adaptation task, a relatively small to medium text corpora in the target task is a requirement to adapt.

3.1.1 Pretraining Dataset

The dataset will be utilized to pretrain a Tagalog variation of XLNet using self-supervision permutative language modeling.

3.1.1.1 Wikitext-TL-39

This study will use the dataset constructed in the study of [22] to pretrain a Tagalog variation of XLNet, which will be used for later experiments.

3.1.2 Benchmark Datasets

The dataset will be utilized in the sequential transfer learning adaptation stage to transfer learning from the Tagalog variation of XLNet to downstream tasks and compare it with the base model (Multinomial Naive Bayes) and benchmark model (RoBERTa Tagalog).

3.1.2.1 Hate Speech Datase

The Hate Speech dataset [29] is a set of tweets mined in real-time during the 2016 Philippine Presidential Election debates and from tweets connected to the 2016 election hashtags. The dataset is presented as a binary classification task benchmark in Filipino, with each tweet being assigned a value of 0 (non-hate) or 1 (hate). 10,000 instances in the training set have labels. To be evaluated are 4,232 samples for testing and 4,232 samples for validation, distributed equally. The training set is reasonably balanced, with 5,340 non-hate tweets and 4,660 hate tweets. All links, mentions, hashtags, obscenities, and other abnormalities are there in the dataset's raw splits that they offer. No additional characteristics are taken from the provided data, and no preprocessing has been done. Sample entries from the hate speech dataset can be found in Table 3.1.

***Table 3.1**. Sample data from the Hate Speech dataset.*

Text	**Label**
GASTOS NI VP BINAY SA POLITICAL ADS HALOS P7-M NA Inaasahan na ni Vice President Jejomar Binay na may mga taong . [LINK]	0
Mar Roxas TANG INA TUWID NA DAAN DAW . . EH SYA NGA DI STRAIGHT	1
Salamat sa walang sawang suporta ng mga taga makati ! Ang Pagbabalik Binay In Makati [HASHTAG] [LINK]	0
Nognog ? Pero nognog din ang nag malasakit ? Wtf ? Tangina mo Binay nagpapaawa kapa ! Hahahahaha [HASHTAG] ? ?	1

3.1.2.2 Shopee-Reviews-TL-Binary

The proponents constructed a new benchmark dataset of 40,000 product reviews from an e-commerce platform, Shopee. The following steps construct the dataset;

1. A web crawler built using JavaScript crawls 672 products from the Shopee website;
2. The links are then fed to a custom script that connects with Shopee API to get all the reviews for each product, resulting in 1,204,473 reviews;
3. After removing reviews without comments, data cleaning, and transformation by removing template tags and non-ASCII characters, and getting only Tagalog and English comments, results in 369,335 reviews;
4. Finally, the proponents utilized the star ratings, VADER, and langdetect to build the final dataset;
   a) All comments are once filtered again using langdetect. All detected English comments with a lower than 60% confidence level are still included in the dataset.
   b) VADER was then applied to comments to classify their sentiments weakly. Afterward, star ratings of four (4) and five (5) with positive sentiments from VADER are given a label of 1 (positive), and ratings of one (1), two (2), and three (3) with negative sentiments from VADER are given a label of 0 (negative).
   c) Finally, the dataset was split into a balanced dataset of positive and negative labels of 20k each. The corpus was split into training, validation, and test sets with a ratio of 70%-15%-15%, respectively.

3.2 Preprocessing and Tokenization

In NLP, textual data is preprocessed and tokenized to be input for the model. Preprocessing focuses on removing the dataset's noise, correcting errors, and cleaning to improve the models' quality. Tokenizing is the process of splitting text into tokens, the basic units of analysis in NLP. Tokens can be words, phrases, or even individual characters.

3.2.1 Preprocessing

Wikitext-TL-39 [22] and Hate Speech datasets [29] are already preprocessed by their respective proponents.

The Shopee-Reviews-TL-Binary were preprocessed by removing emojis, and links, then normalized into Unicode. Then it is pre-tokenized using the NLTK package and gets the average length of the words in the documents. Word with a length greater than 17 is removed from the text to limit the noise. The proponents decide not to discard the casing of the dataset. Sample entries from the Shopee-Reviews-TL-Binary dataset after preprocessing can be found in Table 3.2.

Table 3.2. *Sample data from the Shopee-Reviews-TL-Binary Dataset*

Text	Label
Hindi maganda sira yong speaker,sayang lang pera ko	0
Ok nmn sya tuwang tuwa mother in law ko hehe	1
Ang cute nya, at ang lakas NG sound thanks din Kay suking rider na apakabait	1
ETONG SELLER MAKABENTA NALANG KAHIT DAMAGED ANG ITEM PINAPADALA PA	0

3.2.2 Tokenization

Prior to pretraining an XLNet model, it is necessary to add padding tokens to odd-length lines and generate the LineByLineTextDataset using an XLNet Tokenizer that uses SentecePiece. According to the requirements of Google's XLNet models, a vocabulary of 32,000 tokens is generated is used. The Hate Speech and Shopee-Reviews-TL-Binary dataset was preprocessed using term frequency–inverse document frequency for Multinomial Naive Bayes. RoBERTa and XLNet utilized their tokenizers from the huggingface library.

3.3 Sequential Transfer Learning

Sequential transfer learning is divided into two parts: First, pretraining a Tagalog variation of XLNet using Wikitext-TL-39. Then, adapting it to downstream tasks with finetuning and bagging using the benchmark datasets.

3.3.1 Pretraining

For XLNet, the researchers used Google's provided pretraining scripts and permutative language modeling as its sole pretraining task to train an XLNet Transformer model with an embedding dimension of 768, 12 layers and 12 attention heads (a total of about 110M parameters) on the prepared WikiText-TL-39 [22] corpus and SentencePiece vocabularies. The proponents choose a 0.15 likelihood of a word being masked for the permutative language model pretraining objective. The model has 512 maximum sequence lengths, 2 batch sizes, 16 gradient accumulation stages, and a learning rate of 5e-5.

The model was trained for one epoch with 47627 maximum steps, reaching a final train loss of 2.3912. The model was pretrained for 40 hours on a machine with one NVIDIA GTX 1060 6GB GPU.

3.3.2 Adaptation

The proponents utilized two methods: a commonly used adaptation technique called finetuning, and the other is the proposed enhancement, bootstrap aggregating-based finetuning.

3.3.2.1 Standard Classifier Finetuning

The proponents finetuned the pretrained models to the goal classification tasks by retaining the pretrained weights and adding an appended classification layer or head.

For RoBERTa, the proponents append a classification head composed of a single linear layer and a RELU transformation to the transformer model. The RoBERTa-Tagalog-Base model was then finetuned on both of the classification tasks for three epochs, using a batch size of 32, a learning rate of 2e-4, adam epsilon of 1e-6, adam beta 1 and 2 of ~.9, and a weight decay of 1e-8. The model was finetuned using a machine with one NVIDIA TESLA A1000 40GB GPU.

Similarly, the proponents also append a classification head composed of a single linear layer and a RELU transformation to the transformer model in the XLNet model. The XLNet-Tagalog model was then finetuned on both classification tasks for three epochs, using a batch size of 32 and a learning rate of 2e-5. The model was finetuned using a machine with one NVIDIA TESLA A1000 40GB GPU.

3.3.2.2 Proposed Enhancement using Bootstrap Aggregating-based Finetuning

The proponents created a script to bootstrap the dataset before applying the same finetuning adaptation discussed previously. The bootstrapping process takes a random dataset sample and uses it as input in the finetuning process.

The proponents created three experiments: (1) a 10% bootstrapped dataset with five finetuned models, (2) a 10% bootstrapped dataset with ten finetuned models, and (3) a 70% bootstrapped dataset with three finetuned models. The finetuned models are then used to make inferences from the test dataset and used majority voting in the aggregation step, as shown in Fig 3.2.

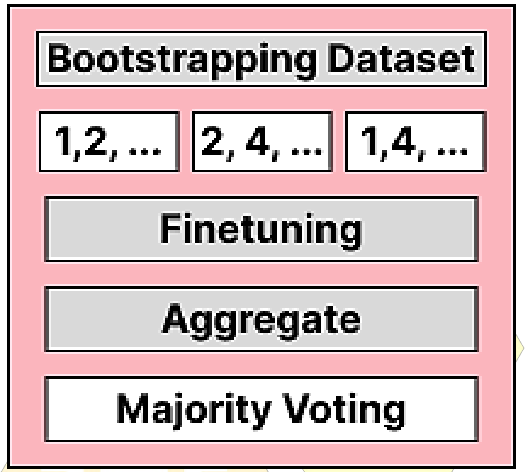

***Fig 3.2** Proposed enhancement to finetuning technique*

3.4 Evaluation Metrics

The proponents use accuracy to evaluate the finetuned models' performance. Proponents used Accuracy Degradation (AD), Degradation Percentage (DP), and Degradation Speed (DS) during the degradation test.

3.4.1 Classification Evaluation Metrics

Accuracy, recall, and precision assess how successfully a model can categorize messages into two groups, such as spam or non-spam.

Accuracy is the ratio of correct predictions to total guesses, as shown in Equation (3.1). However, it can be deceptive when the categories are unbalanced. accuracy = (TP + TN) / (TP + TN + FP + FN) (3.1)

Where,
TP = True Positives
TN = True Negatives
FP = False Positives
FN = False Negatives

3.4.2 Degradation Test

We execute a series of Degradation Tests as a "stress test," similar to Cruz and Cheng [23]. This test mimics training in low-data contexts to determine how much pretraining performance degrades (and, conversely, how much "performance is retained") as the number of training examples decreases. To execute a degradation test, we finetune a model with a smaller sample of a benchmark dataset before testing with the entire test set. All of our degradation tests use four different data percentages: 70%, 30%, 10%, and 1%. We use the same hyperparameters used in standard finetuning for the model tested for finetuning.

For this experiment, we employ two key measures. The first is Accuracy Degradation (AD), the difference in accuracy between a model trained with all the data and a model trained with only a portion of the data, calculated in Equation (3.2).

$ADp\% = Acc100\% - Accp\%$ (3.2)

where Accp% refers to the accuracy of the model trained with p% of data.

Second, we compute the Degradation Percentage (DP), which calculates how much performance from the entire model is lost when a given data percentage p% reduces the training data, calculated in Equation (4.3).

$DPp\% = (ADp\% \ / Acc100\%) \times 100$ (4.3)

where ADp% is the Accuracy Degradation of the model at a certain data percentage p%

In addition to these two metrics, we also provided the model's Degradation Speed (DS), which is just the average of the reported Degradation Percentages for all tests performed. This is the average of DP30%, DP10%, and DP1% in this example.

***Table 4.1** Final Model Test and Train Accuracy Results for Benchmark Datasets*

Model	**Hatespeech**		**Shopee-Reviews-TL-Binary**	
	Train Acc.	Test Acc.	Train Acc.	Test Acc.
Multinomial Naive Bayes (Baseline)	87.01%	74.27%	92.7%	90.05%
RoBERTa-Tagalog (Benchmark)	77.34%	77.72%	90.95%	91.15%
XLNet-Tagalog (Ours)	77.58%	78.05%	94.75%	95.02%

IV. RESULTS AND DISCUSSION

This chapter describes the findings from this paper's techniques and experiments. The outcomes of the various experiment sets outlined in the technique will be included in this chapter. This chapter will also provide a comparison of the study to other existing language models with variations of Tagalog.

4.1 Standard Classifier Finetuning Results

The baseline Multinomial Naive Bayes achieved an accuracy of 74.27% on the hatespeech task and 90.05% on the shopee-reviews-tl-binary classification challenge.

The benchmark RoBERTa was finetuned to a final accuracy of 77.72% for hatespeech and 91.15% for shopee datasets, representing a 3.45% and 1.10% improvement over the baseline model.

While the model of the proponents of a pretrained form of XLNet improved to a final accuracy of 78.05% for hatespeech and 95.02% for the shopee datasets, a 3.78% and 4.97% improvement over the baseline model and 0.33% and 3.87% improvement in the benchmark model, respectively. The model results are all summarized in Table 4.1.

The baseline model, Multinomial Naive Bayes, overfits in the hatespeech dataset that is characterized by a higher accuracy in training dataset compared to test dataset, as also seen in a case study by [30], whereas neither of the language models overfits. Furthermore, language models outperformed the baseline model in accuracy because the language models used to finetune the classifiers contained intact pretrained information.

In all of the fine-tuned experiments, XLNet-Tagalog has the highest accuracy, precision, and recall; additionally, the XLNet model performs better even though it is trained in a smaller corpus, Wikitext-TL-39 [22] in comparison to the benchmark model, RoBERTa-Tagalog is trained in a larger pretraining dataset TLUnified dataset [31].

4.2 Bootstrap Aggregating-based Finetuning Results

The pretrained models were finetuned using the same hyperparameters utilized in the prior experiments with the bootstrapped hatespeech dataset for the bagging-based finetuning. The best-performing experiment in both models uses 70% of the data and three fine-tuned models, as shown in Table 4.2, along with a summary of all experiments.

Table 4.2. *Final Model Results for Hatespeech Dataset using Bagging-based Finetuning*

Hatespeech			
Model	Count	Data%	Accuracy
RoBERTa-Tagalog	1*	100*	77.72%
	3	70	78.17%
	5	10	74.48%
	10	10	75.73%
XLNet-Tagalog	1*	100*	78.05%
	3	70	78.21%
	5	10	71.01%
	10	10	71.30%
XLNet-Tagalog + RoBERTa-Tagalog	1 ea	100	79.47%

Note: * indicates the base accuracy of finetuned models without any applied enhancements

The best-performing RoBERTa-Tagalog bagging-based finetuned model achieved 78.17% accuracy, a 0.45% improvement over the usual finetuning. While the XLNet-Tagalog increased to 78.21% final accuracy, a 0.16% improvement. The difference in accuracy between the five and ten finetuned models, with a difference of 0.29% in XLNet-Tagalog and 1.25% in RoBERTa-Tagalog, supports the proponents' hypothesis that the more significant data percentage and higher number of finetuned models increase the performance of the enhancement.

Table 4.2 demonstrates that when XLNet-Tagalog and RoBERTa-Tagalog are combined, the accuracy increases by 1.42% compared to the best-performing model in the XLNet setup with standard finetuning and 1.26% over the best-performing setup with the bagging-based finetuning.

4.3 Degradation Test Results

The pretrained models are then subjected to degradation tests on 30%, 10%, and 1% of the shopee dataset. The proponents calculated the accuracy degradation,

degradation percentage, and speed to assess the performance retention of the two transformer models when subjected to low-data setups. Table 4.3 summarizes the results.

Table 4.3. Degradation Results for Shopee-Reviews-TL-Binary

Shopee-Reviews-TL-Binary					
Model	Data%	Test Accuracy	ADp%	DPp%	Degradation Speed
XLNet-Tagalog (Ours)	100 10 1	95.02% 92.52% 88.33%	- -2.5% -6.69%	- 2.63% 7.04%	4.835
RoBERTa-Tagalog (Benchmark)	100 10 1	91.15% 85.68% 79.55%	- -5.47% -11.60%	- 6.00% 12.73%	9.365

In the degradation findings reported in table 4.3, XLNet Tagalog degrades 4.53 times slower than the benchmark. This demonstrates that XLNet operates well even when resources are limited.

V. CONCLUSIONS AND RECOMMENDATIONS

This chapter offers recommendations for the next steps based on the findings of the methods and experiment.

5.1. Conclusions

In this thesis, the proponents successfully contributed three contributions: creating a Tagalog benchmark dataset in the product review field. Second, pretraining a Tagalog variation of the XLNet transformer model. Lastly, an enhancement in finetuning techniques was employed using a bagging-based technique. Specifically, the following conclusions were formulated and generated according to the gathered data and findings.

5.1.1. Build a comprehensive Filipino text classification dataset and identify its performance in terms of accuracy against a benchmark dataset for finetuning classification tasks.

The Shopee-Reviews-TL-Binary dataset contains 40k product reviews and a balanced dataset of 20k positive and negative labels. The finetuned model in the shopee dataset has more than 90% accuracy compared to the models finetuned in the benchmark dataset, hatespeech, which has an average accuracy of 70% - 80%.

5.1.2. Determine the improvement in accuracy after implementing bootstrap aggregation as an adaptation technique in sequential transfer learning.

It is also discovered that pretrained with less pretraining data, pretrained XLNet outperforms the baseline classical machine learning approach and RoBERTa. Furthermore, proponents recommend finetuning utilizing bootstrap aggregating (bagging) ensemble, which improves the accuracy of transformer-based sequential transfer learning by 0.10% to 5%. They also proved that the approach might be used for additional binary classification applications, such as hate speech detection and sentiment analysis.

Finally, the model can be trained for low-resource devices by using a lower data percentage in the finetuning step and more models, or it can be trained for high-resource devices by using a higher data percentage in the finetuning step and fewer models.

6.1.3. Determine the performance of pretrained XLNet compared to language models pretrained in the same Filipino corpora when subjected to low resource setting using accuracy degradation, degradation percentage, and degradation speed.

The XLNet-Tagalog variant of proponents performs well in low-resource settings, which is supported by the findings of Cheng and Cruz's (2022) study, which reveals that model accuracy drops dramatically above 10%, whereas in our setting, XLNet accuracy degradation is 6.69%, which is less than 10%, indicating that it performs well in low-resource contexts.

6.2 Recommendations

To improve the contributions and proposed enhancement, future researchers could explore the following areas:

1. Future researchers can scrape more Shopee product reviews to broaden the vocabulary and consider leveraging the template tags to perform multi-label and multiclass text categorization.

2. For pretraining a Tagalog variation, use the XLNet-Large-Cased model as a base model with a larger dataset, such as the TLUnified dataset, or create own dataset.
3. Create more diversified bagging configurations by raising the counts of finetuned models or utilizing a higher data percentage. Use it in conjunction with other language model finetuning to validate its efficacy further.
4. Use other ensemble techniques in the adaptation technique of sequential transfer learning.

REFERENCES

[1] Domo. (2022, September 21). Data Never Sleeps 10.0. Domo.com. https://www.domo.com/data-never-sleeps

[2] Barnhart, B. (2022, June 15). The importance of social media sentiment analysis (and how to conduct it). Sprout Social. https://sproutsocial.com/insights/social-media-sentiment-analysis/

[3] Types of Text - Classification, characteristics and examples - Daily Concepts. (2021, October 16). Daily Concepts. https://conceptdaily.com/types-of-text-classification-characteristics-and-examples

[4] Zong, C., Xia, R., & Zong, C. (2021). Text Classification. 93–124. https://doi.org/10.1007/978-981-16-0100-2_5

[5] Gupta, S. (2018, January 7). Sentiment Analysis: Concept, Analysis and Applications. Medium; Towards Data Science. https://towardsdatascience.com/sentiment-analysis-concept-analysis-and-applications-6c94d6f58c17

[6] Mindoro, J. N., Malbog, M. A. F., Nipas, M. DS., Susa, J. A. B., Acoba, A. G., & Gulmatico, J. S. (2022). Sentiment Analysis in Customer Experience in Philippine Courier Delivery Services using VADER Algorithm Thru Chatbot Interviews. 2022 Second International Conference on Power, Control and Computing Technologies (ICPC2T). https://doi.org/10.1109/icpc2t53885.2022.9777007

[7] Contreras, J. O., Ballera, M. A., Lagman, A. C., & Raviz, J. G. (2018, December 29). Lexicon-based Sentiment Analysis with Pattern Matching Application using Regular Expression in Automata | Proceedings of the 6th International Conference on Information Technology: IoT and Smart City. ACM Other Conferences. https://dl.acm.org/doi/abs/10.1145/3301551.3301596

[8] Sagarino, V. M. C., Montejo, J. I. M., & Ceniza-Canillo, A. M. (2022). Sentiment Analysis of Product Reviews as Customer Recommendations in Shopee Philippines Using Hybrid Approach. 2022 IEEE 7th International Conference on Information Technology and Digital Applications (ICITDA). https://doi.org/10.1109/icitda55840.2022.9971379

[9] Balahadia, F. F., Fernando, Ma. C. G., & Juanatas, I. C. (2016). Teacher's performance evaluation tool using opinion mining with sentiment analysis. 2016 IEEE Region 10 Symposium (TENSYMP). https://doi.org/10.1109/tenconspring.2016.7519384

[10] Pacol, C., & Palaoag, T. (2021). Bilingual Lexicon Approach to English-Filipino Sentiment Analysis of Teaching Performance. IOP Conference Series: Materials Science and Engineering, 1077(1), 012044. https://doi.org/10.1088/1757-899x/1077/1/012044

[11] Valdeavilla, D., & Pulido, M. (2019). Bias in Filipino Newspapers? Newspaper Sentiment Analysis of the 2017 Battle of Marawi. Proceedings of the 4th International Conference on Internet of Things, Big Data and Security. https://doi.org/10.5220/0007752104080413

[12] Boquiren, A. J. V., Garcia, R. A., Hungria, C. J. D., & de Goma, J. C. (2022). Tagalog Sentiment Analysis Using Deep Learning Approach With Backward Slang Inclusion. In IEOM Society. https://ieomsociety.org/proceedings/2022nigeria/180.pdf

[13] Ilao, A. L., & Fajardo, A. C. (2020). SENTIPUBLIKO: Sentiment Analysis of Repost Jejemon Messages using Hybrid Approach Algorithm. IOP Conference Series: Materials Science and Engineering, 938(1), 012010. https://doi.org/10.1088/1757-899x/938/1/012010

[14] Sagarino, V. M. C., Montejo, J. I. M., & Ceniza-Canillo, A. M. (2022). Sentiment Analysis of Product Reviews as Customer Recommendations in Shopee Philippines Using Hybrid Approach. 2022 IEEE 7th International Conference on Information Technology and Digital Applications (ICITDA). https://doi.org/10.1109/icitda55840.2022.9971379

[15] Keen, D., King, M. C., Jerome Lorenzo Lopez, & Ponay, C. (2015, June 27). FilCon: Filipino Sentiment Lexicon Generation Using Word Level-

Annotated Dictionary-Based and Corpus-Based Cross Lingual Approach. ASIALEX 2015. https://www.researchgate.net/publication/279947154_FilCon_Filipino_Sentiment_Lexicon_Generation_Using_Word_Level-Annotated_Dictionary-Based_and_Corpus-Based_Cross_Lingual_Approach

[16] Firsanova, V. (2022, May 4). A Quick Guide to Low-Resource NLP. MLOps Community. https://mlops.community/a-quick-guide-to-low-resource-nlp/

[17] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. ArXiv.org. https://doi.org/10.48550/arXiv.1810.04805

[18] Laumann, F. (2022, June 10). Low-resource language: what does it mean? - NeuralSpace - Medium. Medium; NeuralSpace. https://medium.com/neuralspace/low-resource-language-what-does-it-mean-d067ec85dea5

[19] Keen, D., King, M. C., Jerome Lorenzo Lopez, & Ponay, C. (2015, June 27). FilCon: Filipino Sentiment Lexicon Generation Using Word Level-Annotated Dictionary-Based and Corpus-Based Cross Lingual Approach. ASIALEX 2015. https://www.researchgate.net/publication/279947154_FilCon_Filipino_Sentiment_Lexicon_Generation_Using_Word_Level-Annotated_Dictionary-Based_and_Corpus-Based_Cross_Lingual_Approach

[20] Sagum, R. A., Ramos, A. D., & Llanes, M. T. (2019). FICOBU: Filipino WordNet Construction Using Decision Tree and Language Modeling. International Journal of Machine Learning and Computing, 9(1), 103–107. https://doi.org/10.18178/ijmlc.2019.9.1.772

[21] Kotelnikova, A., Paschenko, D., Bochenina, K., & Kotelnikov, E. (2021). Lexicon-based Methods vs. BERT for Text Sentiment Analysis. ArXiv.org. https://doi.org/10.48550/arXiv.2111.10097

[22] Cruz, J. C. B., & Cheng, C. (2019). Evaluating Language Model Finetuning Techniques for Low-resource Languages. ArXiv.org. https://doi.org/10.13140/RG.2.2.23028.40322

[23] Cruz, J. C. B., & Cheng, C. (2020). Establishing Baselines for Text Classification in Low-Resource Languages. ArXiv.org. https://doi.org/10.48550/arXiv.2005.02068

[24] Canon, M. J. P., Sy, C. Y., Palaoag, T. D., Roxas, R. E. O., & Maceda, L. L. (2022). Language Resource Construction of Multi-Domain Philippine English Text for Pre-training Objective. 2022 International Conference on Advanced Computer Science and Information Systems (ICACSIS). https://doi.org/10.1109/icacsis56558.2022.9923429

[25] Papadimitriou, I., Lopez, K., &amp; Jurafsky, D. (2022, October 11). Multilingual Bert has an accent: Evaluating English influences on fluency in multilingual models. DeepAI. Retrieved January 11, 2023, from https://deepai.org/publication/multilingual-bert-has-an-accent-evaluating-english-influences-on-fluency-in-multilingual-models

[26] Ruder, S., Peters, M. J., Swabha Swayamdipta, & Wolf, T. (2019). Transfer Learning in Natural Language Processing. https://doi.org/10.18653/v1/n19-5004

[27] Grega Vrbančič, & Vili Podgorelec. (2020). Transfer Learning With Adaptive Fine-Tuning. 8, 196197–196211. https://doi.org/10.1109/access.2020.3034343

[28] Mohammed, A., & Kora, R. (2021). An effective ensemble deep learning framework for text classification. 34(10), 8825–8837. https://doi.org/10.1016/j.jksuci.2021.11.001

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

[30] Riego, N. C. R., & Villarba, D. B. (2023). Utilization of Multinomial Naive Bayes Algorithm and Term Frequency Inverse Document Frequency (TF-IDF Vectorizer) in Checking the Credibility of News Tweet in the Philippines. Arxiv.org. http://export.arxiv.org/abs/2306.00018v1

[31] Cruz, J. C. B., & Cheng, C. (2022). Improving Large-scale Language Models and Resources for Filipino. ArXiv.org. https://doi.org/10.48550/arXiv.2111.06053