Automatic Readability Assessment for Closely Related Languages

Joseph Marvin Imperial[Ω,Λ] **Ekaterina Kochmar**[ϒ]
[Ω]National University, Philippines [Λ]University of Bath, UK
[ϒ]MBZUAI, UAE
jmri20@bath.ac.uk ekaterina.kochmar@mbzuai.ac.ae

Abstract

In recent years, the main focus of research on automatic readability assessment (ARA) has shifted towards using expensive deep learning-based methods with the primary goal of increasing models' accuracy. This, however, is rarely applicable for low-resource languages where traditional handcrafted features are still widely used due to the lack of existing NLP tools to extract deeper linguistic representations. In this work, we take a step back from the technical component and focus on how linguistic aspects such as *mutual intelligibility* or *degree of language relatedness* can improve ARA in a low-resource setting. We collect short stories written in three languages in the Philippines – Tagalog, Bikol, and Cebuano – to train readability assessment models and explore the interaction of data and features in various cross-lingual setups. Our results show that the inclusion of CROSSNGO, a novel specialized feature exploiting n-gram overlap applied to languages with high mutual intelligibility, significantly improves the performance of ARA models compared to the use of off-the-shelf large multilingual language models alone. Consequently, when both linguistic representations are combined, we achieve state-of-the-art results for Tagalog and Cebuano, and baseline scores for ARA in Bikol.

We release our data and code at github.com/imperialite/ara-close-lang

1 Introduction

Automatic readability assessment (ARA) is the task that aims to approximate the difficulty level of a piece of literary material using computer-aided tools. The need for such application arises from challenges related to the misalignment of difficulty labels when humans with various domain expertise provide annotations, as well as to the difficulty of manual extraction of complex text-based features (Deutsch et al., 2020). At the same time, readability assessment tools often use different definitions of complexity levels based on (a) age level (Vajjala and Meurers, 2012; Xia et al., 2016), (b) grade level (Imperial and Ong, 2020, 2021a), or on established frameworks such as (c) the Common European Framework of Reference for Languages (CEFR)[1] (François and Fairon, 2012; Pilán et al., 2016; Xia et al., 2016; Reynolds, 2016; Vajjala and Rama, 2018).

In recent years, deep learning methods and large language models (LLMs) have gained popularity in the research community. Often studies using these methodologies focus primarily on improving the performance across various metrics. This is particularly manifest in ARA research in languages with a high number of accessible and publicly-available readability corpora such as English (Heilman et al., 2008; Flor et al., 2013; Vajjala and Lučić, 2018) and German (Hancke et al., 2012; Weiss et al., 2021; Weiss and Meurers, 2022) to name a few. At the same time, existing studies focusing on low-resource languages such as Cebuano (Imperial et al., 2022) and Bengala (Islam et al., 2012; Islam and Rahman, 2014) are still at the stage of primarily using traditional features such as word and sentence lengths to train predictive models.

We identify two problems that are related to the use of complex neural-based approaches: the success of such models depends on (a) whether there is enough available data to train a model using a customized deep neural network, and (b) in the case of LLMs, whether there exists an available off-the-shelf pre-trained model for a low-resource language of interest. Imperial et al. (2022) have recently shown that merely integrating extracted embeddings from a multilingual BERT model as features for Cebuano, a low-resource Philippine language, *does not outperform* models trained with

[1] https://www.coe.int/en/web/common-european-framework-reference-languages/level-descriptions

Findings of the Association for Computational Linguistics: ACL 2023, pages 5371–5386
July 9-14, 2023

orthographic features such as syllable patterns customized for the language. These challenges provide motivation for researchers to further explore methods that do not rely on the availability of large amounts of data or complex pre-trained models and investigate simpler, more interpretable models instead of black box architectures.

In this paper, we take a step back and focus on the data available for low-resource Philippine languages and the features extracted from them rather than on the algorithmic aspects. Specifically, we explore a scenario where small readability corpora are available for languages that are *closely related* or belong to one major language family tree. To the best of our knowledge, incorporating the degree of language closeness or relatedness has not been explored before in any cross-lingual ARA setup. In this study, we make the following contributions:

1. We conduct an extensive pioneer study on readability assessment in a cross-lingual setting using three closely related Philippine languages: Tagalog, Bikolano, and Cebuano.

2. We extract various feature sets ranging from linguistically motivated to neural embeddings, and empirically evaluate how they affect the performance of readability models in a singular, pairwise, and full cross-lingual setup.

3. We introduce cross-lingual Character N-gram Overlap (CROSSNGO), a novel feature applicable to readability assessment in closely related languages.

4. We also introduce and release a new readability corpus for Bikolano, one of the major languages in the Philippines.

5. Finally, we set a baseline for ARA in Bikol and report state-of-the-art results for Tagalog and Cebuano.

2 Background

2.1 The Philippine Linguistic Profile

The Philippines is a linguistically diverse country in Southeast Asia (SEA) with over 180 languages spoken by over 100 million people. Languages in the Philippines can be best described as morphologically rich due to their free-word order structures and high number of possible inflections, full and partial duplications, and compound words (Go and Nocon, 2017). In addition, following lexicostatistical studies, languages are divided into two subgroups, *northern* and *central*, wherein the major languages Ilokano, Pangasinan, and Kapampangan belong to the northern subgroup, and Tagalog, Bikol, Hiligaynon, and Cebuano are allocated to the central subgroup (Walton, 1979; Constantino, 1998). Figure 1 illustrates the central subgroup of the Philippine language family tree.

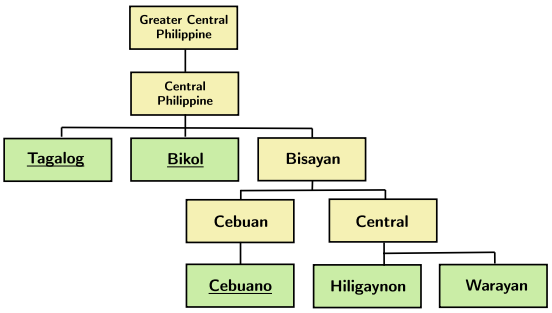


Figure 1: The central subgroup of the Philippine language family tree highlighting the origins of Tagalog, Bikol, and Cebuano.

In this study, our readability experiments focus on three major Philippine languages, *Tagalog*, *Cebuano*, and *Bikol*, which we refer to further in the paper with their corresponding ISO-639-2 language codes as TGL, CEB, and BCL, respectively.

2.2 Mutual Intelligibility

Preliminary linguistic profiling studies of the main Philippine languages such as by McFarland (2004) show that Tagalog, Bikol, and Cebuano are more closely related to one another than any languages in the northern family tree. A language's closeness or its *degree of relatedness* to another language from the same family (sub)tree is commonly referred to as *mutual intelligibility* (Bloomfield, 1926). Such similarities can be seen across multiple aspects, including, for example (a) syllable patterns where all three languages have similar three case-marking particles – *ang* (En: *the*), *ng* (En: *of*), and *sa* (En: *at*) for Bikol and Tagalog, and *ug* instead of *sa* for Cebuano; and (b) shared words, e.g. *mata* (En: *eye*) and *tubig* (En: *water*).

For languages belonging to one greater subgroup in the case of Central Philippine for Tagalog, Bikol, and Cebuano, showing stronger quantitative evidence of mutual intelligibility may provide additional proof that these languages are indeed, at some level, closely related to each other. Thus,

to contribute towards further understanding of mutual intelligibility in the Philippines language space, we apply two linguistic similarity-based measures using character n-gram overlap and genetic distance which we discuss in the sections below.

	TGL	BCL	CEB	ENG
TGL	1.000	0.810	0.812	0.270
BCL	0.810	1.000	0.789	0.263
CEB	0.812	0.789	1.000	0.213
ENG	0.270	0.263	0.213	1.000

(a) Bigram Character Overlap

	TGL	BCL	CEB	ENG
TGL	1.000	0.588	0.628	0.121
BCL	0.588	1.000	0.533	0.144
CEB	0.628	0.533	1.000	0.090
ENG	0.121	0.144	0.090	1.000

(b) Trigram Character Overlap

Table 1: Mutual Intelligibility using Bigram and Trigram Character N-Gram Overlap

Character N-Gram Overlap. For our first measure, we use the overlap in character bigrams and trigrams for every pair from the selected set of languages. To do this, we simply extract and rank the top occurring character bigrams and trigrams for a given language and calculate the Rank-Biased Overlap (RBO)[2] (Webber et al., 2010). RBO provides a measure of similarity between two lists while preserving the ranking. We also add English (ENG) as an unrelated control language not belonging to the Philippine family tree for comparison. We use the CommonCore readability dataset (Flor et al., 2013) for English as it also has three readability levels, and the level distribution is the most similar to the dataset of the three Philippine languages. Further information on the datasets in Tagalog, Bikol, and Cebuano can be found in Section 3. For all languages, we extract the top 25% of the most frequently occurring bigrams and trigrams for analysis. The top 40 most frequent bigrams and trigrams can be found in the Appendix.

Table 1 presents character overlap for bigrams and trigrams in a pairwise manner. These results show that all three Philippine languages have character overlap greater than 75% for bigrams among themselves while overlap with English is below 27%. This pattern is observed again in trigrams with the overlap levels of 53.3% to 62.8% between Tagalog, Bikol, and Cebuano and those below 15% for English. These ranges of mutual intelligibility values for bigram and trigram overlap serve as an estimate of the degree of relatedness between the three Philippine languages, with the values for English serving as a baseline for an unrelated language.

Genetic Distance. As a secondary measure of mutual intelligibility, we calculate the genetic distance score (Beau and Crabbé, 2022) for each pair of languages studied in this work. Similar to the character n-gram overlap analysis, we add English for comparison purposes. Genetic distance (Beaufils and Tomin, 2020) is an automatic measure for quantifying the distance between two languages without the need for human judgments. This metric requires a list of words and their equivalent translations for any two languages of interest and calculates the number of exact consonant matches using the following formula:

$$\text{GeneticDistance} = 100 - \left(\frac{match(l_1, l_2)}{n}\right) \quad (1)$$

where l_1 and l_2 are a pair of languages, n is the total number of words for analysis (usually 100), and $match(\cdot)$ is a function for extracting the consonant patterns for each word from the list as described in Beaufils and Tomin (2020). The metric is measured as a distance; thus, the values closer to 100 denote higher dissimilarity or non-relatedness.

	TGL	BCL	CEB	ENG
TGL	0.000	37.083	24.846	95.690
BCL	37.083	0.000	31.933	70.735
CEB	24.846	31.933	0.000	90.970
ENG	95.690	70.735	90.970	0.000

Range	Meaning
Between 1 and 30	Highly related languages
Between 30 and 50	Related languages
Between 50 and 70	Remotely related languages
Between 70 and 78	Very remotely related languages
Between 78 and 100	No recognizable relationship

Table 2: Mutual Intelligibility using Genetic Distance with Mapping from Beaufils and Tomin (2020).

Table 2 shows the calculated genetic distance scores for each pair of languages including English. The mapping provided in the table is the prescribed guide from Beaufils and Tomin (2020).

[2] https://github.com/changyaochen/rbo

Judging by these results, the Philippine languages have genetic distance scores within the **related** and **highly related languages** range with the Tagalog–Cebuano pair showing the closest language distance of 24.846. Meanwhile, genetic distance scores between all considered Philippine languages and English fall within the **very remotely related** to **no recognizable relationship** categories, with the Tagalog–English pair showing the highest distance from each other. Similar to the character n-gram overlap, these results strengthen our initial observation and provide empirical evidence for mutual intelligibility between Tagalog, Bikol, and Cebuano languages which, beyond this study, may also be used in future linguistic research.

3 Readability Corpora in Philippine Languages

We have compiled open source readability datasets for Tagalog, Cebuano, and Bikol from online library websites and repositories. Each data instance in this study is a fictional short story. Table 3 shows the statistical breakdown and additional information on the levels in each readability dataset across different languages.

Tagalog and Cebuano. Datasets in these languages have already been used in previous research, including Imperial et al. (2019); Imperial and Ong (2020); Imperial (2021); Imperial and Ong (2021a); Imperial et al. (2022). We use the same datasets as in previous research and incorporate them into this study for comparison. For Tagalog, we have assembled 265 instances of children's fictional stories from Adarna House[3] and the Department of Education (DepED)[4]. For Cebuano, we use the dataset collected by Imperial et al. (2022) from Let's Read Asia[5] and Bloom Library[6], which were funded by the Summer Institute of Linguistics (SIL International) and BookLabs to make literary materials in multiple languages available to the public.

Bikol. There are no pre-compiled datasets available for readability assessment in Bikol yet. For this, we collected all available Bikol short stories from Let's Read Asia and Bloom Library totaling 150 instances split into 68, 27, and 55 for levels 1 to 3 respectively.

All collected data for this study follows the standard leveling scheme for early-grade learners or the first three grades from the K-12 Basic Curriculum in the Philippines.[7] Each instance has been annotated by experts with a level from 1 to 3 as seen in Table 3. We use these annotations as target labels in our experiments. Finally, all datasets used in this study can be manually downloaded from their respective websites (see footnotes for links) under the Creative Commons BY 4.0 license.

4 Experimental Setup

4.1 ML Setup

In this study, our primary focus is on the depth of analysis of the traditional and neural features used in a cross-lingual setting applied to closely related languages. Thus, we use a vanilla Random Forest model which has been previously shown to be the best-performing monolingual-trained model for ARA in Tagalog and Cebuano (Imperial and Ong, 2021a; Imperial et al., 2022). We leave the technical breadth of exploring other supervised algorithms to future work.

We use a stratified k-fold approach with k=5 to have well-represented samples per class for a small-dataset scenario used in this study. We report accuracy as the main evaluation metric across all experiments for the ease of performance comparison with previous work (see Section 5). We use WEKA 3.8 (Witten et al., 1999)[8] for all our modeling and evaluation and set hyperparameters of the Random Forest algorithm to their default values as listed in the Appendix.

4.2 Linguistic Features

We extract and consider a wide variety of features inspired by: (a) handcrafted predictors from previous work, (b) representations from a multilingual Transformer-based model (mBERT), and (c) CROSSNGO, a novel feature applicable to readability assessment in closely related languages. We discuss each feature group below.

Traditional Handcrafted Features (TRAD). We integrate available traditional surface-based and

[3] https://adarna.com.ph/
[4] https://lrmds.deped.gov.ph/
[5] https://www.letsreadasia.org/
[6] https://bloomlibrary.org/
[7] https://www.deped.gov.ph/k-to-12/about/k-to-12-basic-education-curriculum/
[8] https://www.cs.waikato.ac.nz/ml/weka/

Source	Language	Level	Doc Count	Sent Count	Vocab
Adarna and DepED	TGL (265)	L1	72	2774	4027
		L2	96	4520	7285
		L3	97	10957	12130
Let's Read Asia and Bloom Library	BCL (150)	L1	68	1578	2674
		L2	27	1144	2009
		L3	55	3347	5509
Let's Read Asia and Bloom Library	CBL (349)	L1	167	1173	2184
		L2	100	2803	4003
		L3	82	3794	6115

Table 3: Statistics on the readability corpora in Tagalog, Cebuano, and Bikol used in this study. The numbers in the brackets provided in the second column are the total number of documents per language broken down in the third and fourth columns per grade level. **Doc Count** and **Sent Count** denote the number of short story instances and the number of sentences per story. **Vocab** is the size of the vocabulary or of the accumulated unique word lists per level.

syllable pattern-based features in this study as predictors of text complexity. These features have been widely used in previous research on ARA in Tagalog and Cebuano (Imperial and Ong, 2020; Imperial et al., 2022). For Bikol, this is the first-ever study to develop a readability assessment model. In the case of low resource languages similar to those used in this study, these predictors are still the go-to features in ARA and have been empirically proven effective for Tagalog and Cebuano (Imperial and Ong, 2021b). We have extracted a total of 18 traditional features for each language, including:

1. The total number of words, phrases, and sentences (3).
2. Average word length, sentence length, and the number of syllables per word (3).
3. The total number of polysyllable words of more than 5 syllables (1).
4. Density of consonant clusters or frequency of consonants without intervening vowels in a word (e.g. Tagalog: *sastre*, En: *dressmaker*) (1).
5. Densities of syllable patterns using the following templates {v, cv, vc, cvc, vcc, ccv, cvcc, ccvc, ccvcc, ccvccc}, where v and c are vowels and consonants respectively (10).

Multilingual Neural Embeddings (mBERT). In addition to the surface-based features, we explore contextual representations from a multilingual Transformer-based large language model via mBERT (Devlin et al., 2019). Previous research on probing BERT has shown convincing evidence that various types of linguistic information (e.g. semantic and syntactic knowledge) are distributed within its twelve layers (Tenney et al., 2019; Rogers et al., 2020). Applying this to ARA, Imperial (2021) showed that BERT embeddings could act as a *substitute* feature set for lower-resource languages such as Filipino, for which NLP tools like POS taggers are lacking.

For this study, we specifically chose mBERT as this particular model has been trained using Wikipedia data in 104 different languages including Tagalog and Cebuano. Bikol is not included in any available off-the-shelf Transformer-based language models due to extremely limited online resources not large enough for training. Nonetheless, we still used the representations provided by mBERT noting its high intelligibility with Tagalog and Cebuano. Feature-wise, we use the mean-pooled representations of the entire twelve layers of mBERT via the `sentence-transformers` library (Reimers and Gurevych, 2019). Each instance in our readability data has an mBERT embedding representation of 768 dimensions.

Cross-lingual Character N-Gram Overlap (CROSSNGO). N-gram overlap has been used previously in various NLP tasks applied to Philippine language data such as language identification (Oco et al., 2013a; Cruz et al., 2016), spell checking and correction (Cheng et al., 2007; Octaviano and Borra, 2017; Go et al., 2017), and clustering (Oco et al., 2013b). Drawing inspiration from this fact and from the quantitative evidence of mutual intelligibility between Philippine languages presented in Section 2, we posit that a new feature designed specifically for closely related language data might improve the performance of the readability assess-

ment models. Thus, we introduce CROSSNGO, which quantifies linguistic similarity using character overlap from a curated list of high-frequency n-grams within languages of high mutual intelligibility. We propose the following formula for calculating this metric:

$$\text{CrossNGO}_{L,n} = \frac{m(L) \bigcap m(d)}{\text{count}(m(d))} \quad (2)$$

where $n \in \{2, 3\}$ denotes bigrams and trigrams, and $m(\cdot)$ is a function that extracts unique n-grams from a document instance d and compares them to a list of top n-grams from a specific language L. For each instance in a dataset, a vector containing three new features will be added representing the overlap between the text and the top n-grams from each of the three languages. We apply this calculation to both bigrams and trigrams using the n-gram lists for Tagalog, Bikol, and Cebuano obtained from the preliminary experiments, which results in a total of 6 new features.

While we presented two quantitative methods of mutual intelligibility in Section 2, only CROSSNGO is applied as a metric *and* a feature for this study. Staying faithful to the work of Beaufils and Tomin (2020), we did not use Genetic Distance to generate another set of features as it was originally developed as a language-to-language metric. Thus, we use it only as additional secondary evidence of language similarity. At the same time, we note that the proposed CROSSNGO bears certain conceptual similarities to Genetic Distance as it measures the frequency of n-gram overlap with other languages. We perform an ablation study and demonstrate the contribution of individual feature sets in Section 5.

5 Results and Discussion

Table 4 shows the accuracy values obtained when training Random Forest models with various combinations of feature groups for each language of interest. The experiments were divided into three setups: (a) *singular cross-lingual* ($l_1 \rightarrow l_2$), (b) *pairwise cross-lingual* ($[l_1 + l_2] \rightarrow l_3$), and (c) *full cross-lingual* ($[l_1 + l_2 + l_3] \rightarrow l_1$), each corresponding to a separate subsection of Table 4. We use the term cross-lingual in this context when a model is trained with a readability corpus from a chosen language l_n or a combination of languages and evaluated with a test set from another language l_m as is demonstrated in Table 4. Similar to our preliminary experiments (Section 2), we include English using the CommonCore dataset as counter-evidence for comparison with closely related languages.

5.1 Low-Resource Languages Benefit from Specialized Cross-lingual Features

For the singular cross-lingual experiments, the effectiveness of exploiting the bigram and trigram overlap via CROSSNGO is demonstrated by high scores for Bikol and Cebuano (75.862 and 78.270) and comparable performance for Tagalog (50.100). Moreover, only for this setup, there is an observed trend where traditional features combined with CROSSNGO outperform mBERT embeddings or the combination of all features for the respective language pair $l_1 \rightarrow l_2$. For Tagalog, this results in 50.100 vs. 26.921 and 23.077; for Bikol – 75.862 vs. 68.965 and 69.000; for Cebuano – 78.270 vs. 71.015 and 73.913. In terms of cross-linguality, in the case of Tagalog, using a model trained with Bikol data proves to be more effective than training with the original Tagalog data with approximately 5.8-point difference in accuracy. However, we still recommend the Tagalog model using all features with 50.000 accuracy since the 0.1 difference is not a significant improvement. Consequently, this trend is not observed in the Bikol and Cebuano experiments where the best-performing models of readability assessment are trained on the data from the same language $l_1 \rightarrow l_1$.

To further confirm if the addition of the CROSSNGO feature statistically improves models' performance as compared to the representations from mBERT for low-resource languages, we aggregate the scores from the TRAD+CROSSNGO group and compare them with the scores obtained when we use mBERT embeddings only, conducting a t-test. We did not include the scores using the combination of all types of features as it would confound the significance test. We achieve statistical significance at $\alpha = 0.01$ level ($p = 0.006$) which shows that using traditional handcrafted features extended with CROSSNGO significantly improves ARA models for low-resource languages, *provided* the availability of data in a closely related language in the case of non-availability of multilingual LLMs (e.g., lack of mBERT model in Bikol).

5.2 Inclusion of a Closely Related Language in Data Produces More Confident Predictions

For pairwise cross-lingual experiments, we investigate the effect of adding a closely related language

Model	TGL				BCL				CEB			
	TRAD	TRAD + CrossNGO	mBERT Embdng	ALL	TRAD	TRAD + CrossNGO	mBERT Embdng	ALL	TRAD	TRAD + CrossNGO	mBERT Embdng	ALL
TGL	43.153	44.231	46.100	50.000	55.172	41.379	20.689	24.137	53.623	57.971	47.826	50.725
BCL	50.000	50.100	26.921	23.077	74.620	75.862	68.965	69.000	63.768	62.320	60.869	66.667
CEB	32.692	38.462	34.615	42.308	51.720	65.517	48.276	44.823	74.058	78.270	71.015	73.913
ENG*	26.923	44.230	28.846	26.923	48.275	37.681	48.250	48.275	46.375	62.018	43.478	43.376
TGL+BCL	51.101	51.923	40.384	**57.692**	72.441	69.965	69.000	68.966	56.521	60.869	62.318	69.565
BCL+CEB	48.077	50.000	42.307	48.076	68.956	72.414	75.611	75.862	74.400	75.362	75.362	**79.710**
CEB+TGL	44.230	36.538	48.076	48.100	52.720	55.172	41.379	34.483	77.711	76.811	73.913	74.464
ALL	50.000	52.910	46.153	32.692	72.413	79.113	65.517	**79.328**	77.710	78.000	78.261	75.630

Table 4: The accuracy of cross-lingual modeling per language with various iterations using different combinations of traditional and neural-based features. The underlined values correspond to the best model for each of the three setups while the **boldfaced** values correspond to the overall highest-performing model for each language across all setups. We included English as counter-evidence only for the singular cross-lingual setup.

on a model's performance using confusion matrices. As the middle section of Table 4 demonstrates, there are three possible pairwise combinations of Tagalog, Bikol, and Cebuano tested on each individual language. As there can be numerous ways to analyze the table, we highlight the results of the cross-lingual models with the top-performing pair and their utilized feature groups and compare them to their equivalent models in the singular cross-lingual experiment. Figure 2 illustrates this method of comparison for each language.

In the case of the Tagalog–Tagalog pair, most misclassifications occur between grades 1 and 2 in both training and test data using all features. This, in turn, is alleviated by incorporating the Bikol dataset in the training data, which reduces the level of confusion by approximately 7%. The inclusion of Bikol also improves classification between grades 2 and 3 by three instances. In the case of the Bikol test data, the same finding is observed for the combined Bikol and Cebuano model using all features, where confusion in classifying grades 1 and 3 is reduced by two instances. Lastly, for Cebuano, the top-performing model in the pairwise cross-lingual setup includes Bikol data and uses all features. For this model, misclassifications in predicting grade 1 against the other two levels are reduced, and performance for predicting grade 3 is improved.

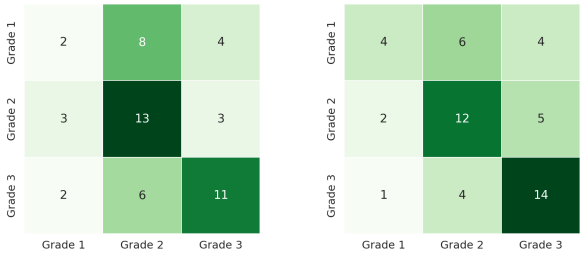
We further corroborate our observations that pairwise cross-lingual models outperform singular cross-lingual models by aggregating the scores from the two setups and running a t-test. Further to the results reported in the previous section, we observe statistically significant difference at the $\alpha = 0.01$ level ($p = 0.003$) when pairwise cross-lingual models are compared to singular cross-lingual models. Overall, our findings provide solid empirical evidence that including a closely related language in the training data for a low-resource language significantly improves performance.

5.3 Combining Specialized Cross-Lingual Features with Multilingual Neural Embeddings Achieves SOTA Results

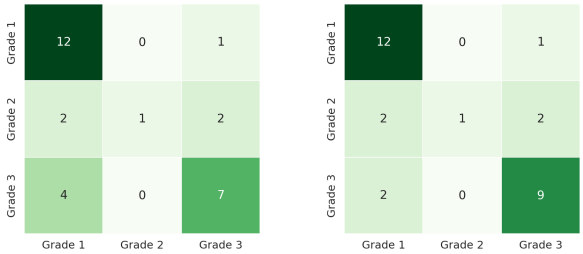
While the previous sections highlight the significant increase in performance when using traditional features with CROSSNGO as compared to mBERT embeddings only, we now discuss results and contributions when both linguistic representations are combined. As is demonstrated in Table 4, the scores obtained using the combined features applied to Tagalog and Cebuano achieve state-of-the-art results for ARA in these languages. For Tagalog, our model's accuracy of 57.692 outperforms the SVM with 57.10 accuracy and the Random Forest model with 46.70 presented in Imperial (2021). For Cebuano, our model achieves 79.710 beating the Random Forest model presented in Imperial et al. (2022) with a score of 57.485 with both models utilizing the same Cebuano dataset. Lastly, as there are no automated readability assessment models yet for Bikol, we report a baseline accuracy of 79.328, which is achieved using a model with a combination of traditional features (extended with CROSSNGO) and mBERT embeddings extracted from data in all three Philippine languages.

5.4 Conventional Fine-Tuning of mBERT Underperforms for Low Resource Cross-Lingual ARA

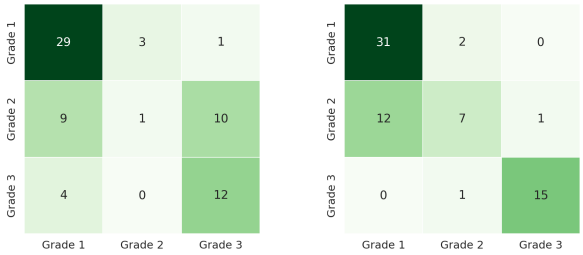
While the main focus of our work is on using traditional machine learning models with Random Forest, we explore if the standard approach for fine-



(a) TGL only trained model (left) against TGL+BCL model (right) using all features for ARA in TGL.



(b) BCL only trained model (left) against BCL+CEB model (right) using all features for ARA in BCL.



(c) BCL only trained model (left) against BCL+CEB model (right) using multilingual embeddings for ARA in CEB.

Figure 2: Confusion matrices for pairwise cross-lingual setup for all three languages. All models using an additional language dataset for ARA achieved an improved performance with an average of 10.131 across the board (7.692 for TGL, 6.862 for BCL, and 15.841 for CEB).

Model	TGL	BCL	CEB
TGL	0.420	0.500	0.333
BCL	0.420	0.633	0.575
CEB	**0.520**	0.500	**0.697**
TGL+BCL	0.440	0.566	0.469
BCL+CEB	0.400	**0.637**	0.666
CEB+TGL	0.480	0.500	0.590
***ALL**	0.460	0.633	0.636

Table 5: The accuracy scores of the conventional fine-tuning strategy applied to LLMs for various methods of cross-lingual ARA using the same uncased mBERT model for the extraction of embeddings.

tuning LLMs such as mBERT can produce comparable performance. We use the same uncased mBERT model as presented in Section 4.

Table 5 shows the performance of singular, pairwise, and full cross-lingual setups formatted similarly to Table 4. These results confirm the findings of Ibañez et al. (2022), who have applied a similar setup to monolingual Tagalog ARA using a Tagalog BERT model. Judging by their results, the conventional fine-tuning approach proved to be inferior to the traditional way of extracting linguistic features from text and training a machine learning model like SVM or Random Forest. For this study, the highest-performing setups for Tagalog and Cebuano use Cebuano data only, and that for Bikol uses the combined Cebuano + Bikol datasets. None of the fine-tuned models outperform those presented in Table 4 using combinations of traditional features and CROSSNGO. While previous work in cross-lingual ARA by Lee and Vajjala (2022) and Madrazo Azpiazu and Pera (2020) achieved relatively high performance with non-closely related languages using LLMs, we obtain less promising results which we can attribute to: (a) the use of datasets of substantially smaller sizes (a total of 13,786 documents used in Azpiazu and Pera (2019) and 17,518 in Lee and Vajjala (2022) vs. only 764 in out study), and (b) lack of diverse data sources since only Wikipedia dumps were used for Tagalog and Cebuano for training the mBERT model.

6 Conclusion

In this work, we took a step back from the trend of exploring various technical components of the complex, deep learning models and, instead, focused on studying the potential effectiveness of linguistic characteristics such as mutual intelligibility for ARA in closely related Philippine languages – Tagalog, Bikol, and Cebuano. We implemented three cross-lingual setups to closely study the effects of interaction between the three languages and proposed a new feature utilizing n-gram overlap, CROSSNGO, which is specially developed for cross-lingual ARA using closely related languages. Our results show that: (a) using CROSSNGO combined with handcrafted features achieves significantly higher performance than using mBERT embeddings, (b) the inclusion of another closely related Philippine language reduces model confusion, and (c) using the conventional fine-tuning for LLMs like mBERT in this setup still does not

outperform models with traditional features. Consequently, we come to the conclusion that using languages with high intelligibility is more suited for cross-lingual ARA. This is demonstrated in experiments with English added as an example of a non-related language, in which we do not achieve a substantial increase in performances for Tagalog, Cebuano, and Bikol.

Our results agree with the findings of previous studies in cross-lingual ARA such as those of Madrazo Azpiazu and Pera (2020) using English, Spanish, Basque, Italian, French, Catalan, and Weiss et al. (2021) using English and German, that also showed that the inclusion of additional language data can improve ARA results on other languages. However, our work is primarily motivated by the degree of language relatedness: we show that better results can be achieved for ARA in low-resource languages if we use closely related languages rather than any language, including non-related ones like English. Our study also provides an encouragement for researchers to consider approaches grounded in linguistic theories which can potentially be used to improve the performance in NLP tasks rather than always resorting to models that are expensive to train and hard to interpret.

7 Limitations

We discuss some limitations of our current work which can be further explored in the future.

On Data Format. We specifically use fictional short stories as our primary data for the study since we require gold standard labels for this document classification task. Moreover, fictional short stories are easier to find as they often come with a specified grade level compared to other types of literary texts such as magazines or web articles written in any of the three Philippine languages. We do not claim that our models are able to generalize on these other types of literary materials or on other types of closely related language pairs unless a full study is conducted which is outside the scope of this work.

On Handcrafted Features. We were only able to use traditional handcrafted features covering count-based predictors such as sentence or word count and syllable pattern-based features for training the Random Forest models. We did not extract other feature sets one may find in the previous work on English such as lexical density or discourse-based features since such features require NLP tools that are able to extract POS, named entities, relations, and discourse patterns that do not yet exist for all three Philippine languages used in this study. The work of Imperial and Ong (2021b) covered a small set of lexical features such as *type–token ratio* and *compound word density* for readability assessment in Tagalog. Still, we cannot use this approach since all languages would need to have the same number of features as is a standard practice in model training.

On Model Training. Our choice of the Random Forest algorithm for training the ARA models is based on the substantial amount of previous work supporting the application of this method to low-resource ARA, e.g., to Tagalog and Cebuano in a monolingual setup (Imperial and Ong, 2020, 2021a; Imperial, 2021; Imperial et al., 2022), where it achieved better results than other algorithms such as SVM or Logistic Regression. One can consider these algorithms for comparison but the analysis of each ARA model trained with various algorithms to the same level of depth and focus that we have given to the Random Forest classifier in the present study would require a considerable amount of time as well as a higher page limit.

On Current Measures of Mutual Intelligibility. The majority of existing literature in linguistics, specifically on the topic of mutual intelligibility in Philippine languages, discusses examples in the context of speech communication. As such, one might claim that Cebuano and Tagalog are *not* mutually intelligible by giving an example where a Tagalog speaker may not fully comprehend (or only recognize a few common words) another speaker if they are talking in Cebuano. While this is certainly true, in this study, we specifically focus on the mutual intelligibility of languages at a word and character level via written texts such as children's fiction books. From this, we see a substantial degree of *closeness* between Tagalog, Cebuano, and Bikol compared to English. Thus, based on our results, we posit that mutual intelligibility may be used as an additional feature (see CROSSNGO in Section 4) for text-based tasks such as readability assessment. We leave the exploration of our proposed novel feature in the speech communication

area to future work.

8 Ethical Considerations

We foresee no ethical issues related to the study.

Acknowledgements

We thank the anonymous reviewers and area chairs for their constructive and helpful feedback. We also thank the communities and organizations behind the creation of open-source datasets in Philippine languages used in this research: DepED, Adarna House, Bloom Library, Let's Read Asia, SIL, and BookLabs. JMI is supported by the UKRI CDT in Accountable, Responsible, and Transparent AI of the University of Bath and by the Study Grant Program of the National University Philippines.

References

Ion Madrazo Azpiazu and Maria Soledad Pera. 2019. Multiattentive recurrent neural network architecture for multilingual readability assessment. *Transactions of the Association for Computational Linguistics*, 7:421–436.

Nathanaël Beau and Benoit Crabbé. 2022. The impact of lexical and grammatical processing on generating code from natural language. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2204–2214, Dublin, Ireland. Association for Computational Linguistics.

Vincent Beaufils and Johannes Tomin. 2020. Stochastic approach to worldwide language classification: the signals and the noise towards long-range exploration. SocArXiv.

Leonard Bloomfield. 1926. A set of postulates for the science of language. *Language*, 2(3):153–164.

Charibeth Cheng, Cedric Paul Alberto, Ian Anthony Chan, and Vazir Joshua Querol. 2007. SpellChef: spelling checker and corrector for Filipino. *Journal of Research in Science, Computing and Engineering*, 4(3):75–82.

Ernesto A Constantino. 1998. Current topics in Philippine linguistics. In *Revised version of the paper read at the meeting of the Linguistic Society of Japan held in Yamaguchi University, Yamaguchi, Japan, on 31 0ctober*.

Angelica Dela Cruz, Nathaniel Oco, Leif Romeritch Syliongka, and Rachel Edita Roxas. 2016. Phoneme inventory, trigrams and geographic location as features for clustering different philippine languages. In *2016 Conference of The Oriental Chapter of International Committee for Coordination and Standardization of Speech Databases and Assessment Techniques (O-COCOSDA)*, pages 137–140. IEEE.

Tovly Deutsch, Masoud Jasbi, and Stuart Shieber. 2020. Linguistic features for readability assessment. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–17, Seattle, WA, USA → Online. Association for Computational Linguistics.

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.

Michael Flor, Beata Beigman Klebanov, and Kathleen M. Sheehan. 2013. Lexical tightness and text complexity. In *Proceedings of the Workshop on Natural Language Processing for Improving Textual Accessibility*, pages 29–38, Atlanta, Georgia. Association for Computational Linguistics.

Thomas François and Cédrick Fairon. 2012. An "AI readability" formula for French as a foreign language. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 466–477.

Matthew Phillip Go and Nicco Nocon. 2017. Using Stanford part-of-speech tagger for the morphologically-rich Filipino language. In *Proceedings of the 31st Pacific Asia Conference on Language, Information and Computation*, pages 81–88. The National University (Phillippines).

Matthew Phillip Go, Nicco Nocon, and Allan Borra. 2017. Gramatika: A grammar checker for the low-resourced Filipino language. In *TENCON 2017-2017 IEEE Region 10 Conference*, pages 471–475. IEEE.

Julia Hancke, Sowmya Vajjala, and Detmar Meurers. 2012. Readability classification for German using lexical, syntactic, and morphological features. In *Proceedings of COLING 2012*, pages 1063–1080, Mumbai, India. The COLING 2012 Organizing Committee.

Michael Heilman, Kevyn Collins-Thompson, and Maxine Eskenazi. 2008. An analysis of statistical models and features for reading difficulty prediction. In *Proceedings of the Third Workshop on Innovative Use of NLP for Building Educational Applications*, pages 71–79, Columbus, Ohio. Association for Computational Linguistics.

Michael Ibañez, Lloyd Lois Antonie Reyes, Ranz Sapinit, Mohammed Ahmed Hussien, and Joseph Marvin Imperial. 2022. On Applicability of Neural Language Models for Readability Assessment in Filipino. In *International Conference on Artificial Intelligence in Education*, pages 573–576. Springer.

Joseph Marvin Imperial. 2021. BERT embeddings for automatic readability assessment. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 611–618, Held Online. INCOMA Ltd.

Joseph Marvin Imperial and Ethel Ong. 2020. Exploring hybrid linguistic feature sets to measure Filipino text readability. In *2020 International Conference on Asian Language Processing (IALP)*, pages 175–180. IEEE.

Joseph Marvin Imperial and Ethel Ong. 2021a. Diverse linguistic features for assessing reading difficulty of educational Filipino texts. *arXiv preprint arXiv:2108.00241*.

Joseph Marvin Imperial and Ethel Ong. 2021b. Under the microscope: Interpreting readability assessment models for Filipino. In *Proceedings of the 35th Pacific Asia Conference on Language, Information and Computation*, pages 1–10, Shanghai, China. Association for Computational Lingustics.

Joseph Marvin Imperial, Lloyd Lois Antonie Reyes, Michael Antonio Ibanez, Ranz Sapinit, and Mohammed Hussien. 2022. A baseline readability model for Cebuano. In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, pages 27–32, Seattle, Washington. Association for Computational Linguistics.

Joseph Marvin Imperial, Rachel Edita Roxas, Erica Mae Campos, Jemelee Oandasan, Reyniel Caraballo, Ferry Winsley Sabdani, and Ani Rosa Almaroi. 2019. Developing a machine learning-based grade level classifier for Filipino children's literature. In *2019 International Conference on Asian Language Processing (IALP)*, pages 413–418. IEEE.

Zahurul Islam, Alexander Mehler, and Rashedur Rahman. 2012. Text readability classification of textbooks of a low-resource language. In *Proceedings of the 26th Pacific Asia Conference on Language, Information, and Computation*, pages 545–553.

Zahurul Islam and Rashedur Rahman. 2014. Readability of Bangla news articles for children. In *Proceedings of the 28th Pacific Asia Conference on Language, Information and Computing*, pages 309–317.

Justin Lee and Sowmya Vajjala. 2022. A neural pairwise ranking model for readability assessment. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3802–3813, Dublin, Ireland. Association for Computational Linguistics.

Ion Madrazo Azpiazu and Maria Soledad Pera. 2020. Is cross-lingual readability assessment possible? *Journal of the Association for Information Science and Technology*, 71(6):644–656.

Curtis D McFarland. 2004. The Philippine language situation. *World Englishes*, 23(1):59–75.

Nathaniel Oco, Joel Ilao, Rachel Edita Roxas, and Leif Romeritch Syliongka. 2013a. Measuring language similarity using trigrams: Limitations of language identification. In *2013 International Conference on Recent Trends in Information Technology (ICRTIT)*, pages 478–481. IEEE.

Nathaniel Oco, Leif Romeritch Syliongka, Rachel Edita Roxas, and Joel Ilao. 2013b. Dice's coefficient on trigram profiles as metric for language similarity. In *2013 International Conference Oriental COCOSDA held jointly with 2013 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE)*, pages 1–4. IEEE.

Manolito Octaviano and Allan Borra. 2017. A spell checker for a low-resourced and morphologically rich language. In *TENCON 2017-2017 IEEE Region 10 Conference*, pages 1853–1856. IEEE.

Ildikó Pilán, Sowmya Vajjala, and Elena Volodina. 2016. A readable read: Automatic assessment of language learning materials based on linguistic complexity. *arXiv preprint arXiv:1603.08868*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Robert Reynolds. 2016. Insights from Russian second language readability classification: complexity-dependent training requirements, and feature evaluation of multiple categories. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 289–300, San Diego, CA. Association for Computational Linguistics.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

Sowmya Vajjala and Ivana Lučić. 2018. OneStopEnglish corpus: A new corpus for automatic readability assessment and text simplification. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 297–304, New Orleans, Louisiana. Association for Computational Linguistics.

Sowmya Vajjala and Detmar Meurers. 2012. On improving the accuracy of readability classification using insights from second language acquisition. In

Proceedings of the seventh workshop on building educational applications using NLP, pages 163–173.

Sowmya Vajjala and Taraka Rama. 2018. Experiments with universal CEFR classification. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 147–153, New Orleans, Louisiana. Association for Computational Linguistics.

Charles Walton. 1979. A Philippine language tree. *Anthropological linguistics*, 21(2):70–98.

William Webber, Alistair Moffat, and Justin Zobel. 2010. A similarity measure for indefinite rankings. *ACM Transactions on Information Systems (TOIS)*, 28(4):1–38.

Zarah Weiss, Xiaobin Chen, and Detmar Meurers. 2021. Using broad linguistic complexity modeling for cross-lingual readability assessment. In *Proceedings of the 10th Workshop on NLP for Computer Assisted Language Learning*, pages 38–54, Online. LiU Electronic Press.

Zarah Weiss and Detmar Meurers. 2022. Assessing sentence readability for German language learners with broad linguistic modeling or readability formulas: When do linguistic insights make a difference? In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, pages 141–153, Seattle, Washington. Association for Computational Linguistics.

Ian H Witten, Eibe Frank, Leonard E Trigg, Mark A Hall, Geoffrey Holmes, and Sally Jo Cunningham. 1999. Weka: Practical machine learning tools and techniques with Java implementations.

Menglin Xia, Ekaterina Kochmar, and Ted Briscoe. 2016. Text readability assessment for second language learners. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 12–22, San Diego, CA. Association for Computational Linguistics.

A Appendix

Hyperparameter	**Value**
batchSize	100
bagSizePercent	100
maxDepth	unlimited
numIterations	100
numFeatures	int(log(#predictors) + 1)
seed	1

Table 6: Hyperparameter settings for the Random Forest algorithm used for training the models in WEKA. These are default values and the 3.8.6 version of WEKA would have these already preset.

Hyperparameter	**Value**
max seq length	300
batch size	8
dropout	0.01
optimizer	Adam
activation	ReLu
layer count	1 (768 x 256)
loss	Negative Log Likelihood
learning rate	0.002
epochs	50

Table 7: Hyperparameter settings for the mBERT model used for fine-tuning. Please refer to Ibañez et al. (2022) for more information on these values.

TGL		CEB		BCL	
bigram	**count**	**bigram**	**count**	**bigram**	**count**
ng	43215	an	15636	an	12562
an	39268	ng	14451	na	7315
na	22041	sa	8311	ng	6754
in	18449	na	7167	in	6138
ma	16501	ga	6714	sa	5753
sa	16037	ka	5951	ka	5176
la	15283	la	5638	ag	4558
ka	14263	ma	4889	ma	4452
ag	12386	ni	4701	on	3490
at	12380	ta	4692	ga	3462
pa	12171	in	4591	pa	3453
al	11521	pa	4333	ni	3416
ga	10818	ag	4247	ak	3291
ay	10771	on	4113	ar	3012
ak	10271	ay	3799	si	2957
ni	9814	si	3636	da	2920
ta	9738	ya	3603	ya	2886
si	9126	al	3406	ta	2796
ya	8724	at	3150	la	2676
on	8288	ba	3099	al	2658
ba	7402	ak	3062	ba	2613
it	7288	ha	2729	ra	2518
am	6667	iy	2634	as	2447
iy	6339	ug	2531	at	2315
as	6210	il	2511	ay	2187
ko	5928	un	2502	ab	1893
ha	5885	gi	2460	ai	1843
il	5857	li	2413	ko	1840
ar	5848	am	2327	ha	1763
li	5696	ah	2251	li	1697
ap	5190	it	2059	ad	1679
ab	5000	ad	1834	ro	1574
ra	4867	as	1801	am	1544
da	4777	da	1793	un	1316
aw	4598	us	1781	ti	1293
ti	4577	ko	1771	nd	1202
wa	4572	to	1770	ap	1172
ah	4410	aw	1767	mg	1165
um	4391	ab	1690	ah	1164
bi	4382	yo	1667	it	1160
is	4286	ki	1615	bi	1146
to	4248	hi	1589	ku	1140
mi	4179	ap	1516	aw	1139
un	4168	mg	1504	wa	1086

Table 8: Full list of the top 25% bigrams extracted from the Tagalog, Cebuano, and Bikol datasets. The same list is used for calculating overlap via CROSSNGO.

TGL		CEB		BCL	
trigram	**count**	**trigram**	**count**	**trigram**	**count**
ang	22650	ang	7941	ang	3350
ala	6120	nga	3283	nag	1721
ing	5456	iya	2547	kan	1518
ong	5036	ing	1697	aka	1507
iya	4761	ala	1534	ing	1434
lan	3880	mga	1479	nin	1389
ina	3481	ila	1474	ong	1374
aka	3266	ana	1395	ara	1210
nan	3151	lan	1317	mga	1164
ama	3021	ong	1315	man	1103
ara	3007	ata	1306	yan	979
ata	2976	usa	1286	sin	947
ila	2965	tan	1276	ala	940
mga	2867	yan	1172	iya	928
nag	2797	han	1139	asi	897
niy	2795	ali	1061	sai	853
pag	2793	nag	1043	aba	835
yan	2757	pag	982	ina	833
apa	2716	aka	975	aga	824
aga	2694	ayo	933	ini	816
ali	2622	aha	931	mag	812
man	2574	nan	928	aro	730
aha	2450	siy	916	ako	730
uma	2412	ako	868	gan	718
aki	2376	pan	863	par	705
nga	2281	ama	847	nbs	702
mag	2269	man	831	bsp	702
aba	2253	ini	830	ata	683
awa	2249	ita	827	nga	683
kan	2219	una	811	pag	639
tin	2208	ina	763	ati	605
asa	2142	aba	758	lan	582
ako	2130	kin	744	ion	576
hin	2119	nak	727	nda	574
ito	2033	ung	718	lin	569
aya	2000	kan	716	sak	567
ana	1993	san	700	ano	553
gan	1973	nah	700	ban	547
ami	1934	ngo	679	ind	538
san	1913	kat	675	ron	530
nak	1896	gan	665	apa	527
abi	1878	ula	636	ana	526
tan	1844	ano	626	ili	524
siy	1835	uot	611	ent	508
ani	1773	ahi	605	ada	502

Table 9: Full list of the top 25% trigrams extracted from the Tagalog, Cebuano, and Bikol datasets. The same list is used for calculating overlap via CROSSNGO.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Section 7

☐ A2. Did you discuss any potential risks of your work?
Not applicable. Left blank.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Section 1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

We used WEKA as discussed in Section 4.

☑ B1. Did you cite the creators of artifacts you used?
Section 4

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Section 3 for the collected Philippine language data and Section 4 for WEKA

☐ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Not applicable. Left blank.

☒ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
The dataset is not scraped from social media sites.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Section 2-3

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Section 4

C ☑ Did you run computational experiments?

Left blank.

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Appendix

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Appendix

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Section 5

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Section 4

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Not applicable. Left blank.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Not applicable. Left blank.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Not applicable. Left blank.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. Left blank.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Not applicable. Left blank.