Hybrid Neural-Rule Based Architectures for Filipino Stemming with Fine-Tuned BERT Variants

Angelica Anne A. Naguio
University of the Philippines Los Baños
Los Baños, Laguna
aanaguio@up.edu.ph

Rachel Edita O. Roxas
University of the Philippines Los Baños
Los Baños, Laguna
roroxas2@up.edu.ph

Abstract

This paper introduces a novel hybrid neural-rule based architecture for Filipino stemming, combining a comprehensive rule-based stemmer with fine-tuned BERT variants. We systematically compare untrained models, rule-based models, and fine-tuned BERT models, demonstrating significant performance improvements with our hybrid approach. The RoBERTa Tagalog variant achieves 98.61% Exact Accuracy and 98.23% F1-score, outperforming both untrained and purely rule-based methods. Our findings suggest that integrating domain-specific linguistic rules with neural networks is essential for effective NLP in morphologically complex, low-resource languages like Filipino, offering a framework adaptable to similar languages.

1 Introduction

Filipino, the national language of the Philippines, is derived from Tagalog and belongs to the Austronesian language family. It shares linguistic features with other languages in Southeast Asia and the Pacific, characterized by its rich morphological structure, including a complex system of affixation, reduplication, and the use of clitics (Blust, 2009; Rubino, 2002; Roxas et al., 2009). These linguistic characteristics, while offering expressive versatility, present unique challenges for natural language processing (NLP) tasks such as stemming, lemmatization, and morphological analysis (Katamba, 1993; Yambao, 2021; Roxas and Mula, 2008).

Filipino's agglutinative grammar employs a complex system of affixes (prefixes, infixes, suffixes, and circumfixes) to express grammatical functions like tense, aspect, and voice (Blake, 1917; Cheng and See, 2006; Roxas et al., 2009). For instance, the verb root *bili* (to buy) transforms into *bumili* (bought), *binili* (was bought), *bibili* (will buy), and *pinagbibili* (being sold), demonstrating how a single root can generate multiple forms with distinct grammatical implications that challenge computational processing (Roxas and Mula, 2008).

Reduplication is another prominent feature of Filipino morphology, involving the repetition of a whole or partial root to convey grammatical functions such as plurality, intensity, or reciprocity (Blake, 1917; Roxas et al., 2009). For example, the root *takbo* (run) may become *tatakbo* (will run), indicating future tense, or *takbo-takbo* (running around), indicating repetitive action. The ability to accurately parse and handle reduplication is essential for any effective stemming or lemmatization algorithm in Filipino (Roxas and Mula, 2008).

Furthermore, Filipino extensively uses clitics—unstressed particles that attach to a preceding word to convey syntactic or phonological nuances (Bloomfield, 1917; Roxas et al., 2009). Common clitics include *ng* (of), *na* (already), and *pa* (still), which often need careful handling during preprocessing to ensure accurate linguistic analysis. The complex interplay of clitics, affixation, and reduplication makes Filipino a challenging language for NLP systems developed primarily for languages like English, which have comparatively simpler morphological structures (Roxas et al., 2009).

While recent NLP advancements favor subword tokenization and end-to-end methods, morphological analysis remains indispensable for morphologically rich languages (MRLs) like Filipino, where morphological markers encode grammatical functions that influence word meaning and syntax (Tsarfaty et al., 2013; Erkaya, 2022). In contrast to languages like English, where grammatical roles are defined by word order, Filipino relies on affixation, reduplication, and compounding to convey these functions, thus enabling flexible word order and presenting unique challenges for standard NLP models (Roxas et al., 2009).

The need for explicit morphological processing is particularly evident in applications like information retrieval, where stemming improves search relevance by matching root forms rather than exact terms (Adriani et al., 2007). For Filipino, with limited annotated data and high morphological variation, stemming becomes essential to reduce lexical sparsity and enhance performance in tasks such as document classification and sentiment analysis (Boquiren et al., 2022; Bonus, 2003).

The task of stemming—reducing words to their root form—is particularly challenging in Filipino due to its extensive use of affixes and the necessity of correctly interpreting these morphological variations (Adriani et al., 2007; McNamee and Mayfield, 2004; Roxas and Mula, 2008). Traditional rule-based approaches have historically been employed to address this challenge, leveraging handcrafted linguistic rules to strip affixes and reduce words to their base forms. However, while effective in specific cases, rule-based systems often suffer from limitations in scalability, adaptability, and the ability to generalize to unseen data (Roxas et al., 2009).

The advent of neural network-based models, particularly those utilizing the Transformer architecture (Devlin et al., 2019), has shifted the focus of NLP towards data-driven approaches. Despite their success in many domains, purely neural models often struggle with morphologically rich languages like Filipino, where complex linguistic rules must be implicitly learned from data (Pires et al., 2019; Lample and Conneau, 2019). Without extensive, language-specific training data, these models can underperform, highlighting the need for hybrid approaches that combine the strengths of rule-based systems with the generalization capabilities of neural networks (Gatt and Krahmer, 2018; Malmasi and Dras, 2014).

Hybrid models that integrate rule-based methodologies with neural networks offer a promising solution to the challenges posed by the morphological complexity of Filipino (Gatt and Krahmer, 2018; Malmasi and Dras, 2014; Roxas and Mula, 2008). By embedding linguistic rules within a rule-based stemmer and enhancing it with the contextual understanding provided by fine-tuned BERT models, a hybrid approach can achieve higher accuracy and robustness in stemming tasks (Yambao, 2021). This synergy leverages the precise, deterministic nature of rule-based systems with the adaptive, contextual strengths of neural models, offering a more comprehensive solution to the complexities of Filipino morphology (Roxas and Mula, 2008).

This research builds on existing hybrid approaches by proposing a novel model that combines a rule-based Filipino stemmer with fine-tuned BERT variants. These variants—Multilingual BERT, RoBERTa Tagalog, and XLM-RoBERTa—have been pre-trained on large multilingual corpora but require adaptation to effectively handle the unique morphological characteristics of Filipino (Devlin et al., 2019; Cruz and Cheng, 2022; Lample and Conneau, 2019). By fine-tuning these models on a Filipino-specific dataset and integrating them with a rule-based stemmer, this study seeks to enhance the performance of NLP tasks in Filipino, particularly stemming.

The contributions of this research are twofold. First, we provide a comprehensive evaluation of the performance of various BERT variants on Filipino word stemming, both as standalone models and within a hybrid framework. Second, we demonstrate the effectiveness of the hybrid approach in handling the morphological complexity of Filipino, offering insights that may be applicable to other morphologically rich, low-resource languages. This research not only advances the state of the art in Filipino NLP but also provides a foundation for future work in developing more sophisticated and adaptable linguistic models for diverse languages worldwide.

2 Related Works

The rise of transformer-based models such as BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and XLM-RoBERTa (Lample and Conneau, 2019) has significantly advanced natural language processing (NLP) tasks. These models utilize deep bidirectional transformers to capture contextual information from large corpora, achieving state-of-the-art performance in various language understanding tasks across multiple languages. However, their application in morphologically rich languages, such as Filipino, presents unique challenges due to the complexity and variability in word formation (Cruz and Cheng, 2022).

2.1 Filipino and Morphological Challenges

Filipino's morphological complexity, involving affixation, reduplication, and compounding, challenges standard NLP approaches, necessitating model adaptations for lexical variation, which presents a challenge for standard NLP models (Roxas et al., 2009). The language's rich system of affixes can complicate computational processing, as traditional rule-based systems often struggle with the irregularities and extensive use of affixes in Filipino (Nelson, 2004).

Traditional approaches to Filipino morphology, such as rule-based systems, have been explored but often face limitations. For instance, an exhaustive rule-based affix extraction method for Tagalog was proposed to address issues of understemming and unstemmed errors by generating all possible word forms through a tree structure (Tolentino and Borra, 2018). Additionally, a morphological analyzer for Filipino verbs has been developed to produce affixes, infinitive forms, and tenses from conjugated verbs, highlighting the complexity of Filipino morphological analysis (Roxas and Mula, 2008). Despite these advancements, there is still a need for improved methods that can handle the nuances of Filipino word formation more effectively, particularly in capturing irregular forms and complex morphophonological alternations (Yambao, 2021).

The challenge is further compounded by the scarcity of high-quality annotated data, though recent initiatives like iTANONG-DS have begun addressing this limitation by providing comprehensive benchmark datasets (Visperas et al., 2023). Similarly, developing models that effectively handle Filipino's unique morphological features remains an active research challenge (Riego et al., 2023).

2.2 Stemming in Filipino NLP

Stemming plays a crucial role in various Filipino NLP applications, particularly in addressing the challenges posed by the language's rich morphological structure. In sentiment analysis, stemming helps identify sentiments by reducing morphologically complex emotional words to their root forms. For example, words like *masaya* (happy), *nagpapasaya* (making happy), *pinasaya* (made happy), and *kasiyahan* (happiness) are reduced to their root form 'saya', which is particularly important for social media text analysis where these morphological variations are common (Boquiren et al., 2022). This morphological reduction is crucial for improving text classification tasks by reducing lexical sparsity and consolidating semantically related word forms (Cruz and Cheng, 2022).

In information retrieval tasks, stemming improves search effectiveness by reducing feature space dimensionality and enabling better semantic matching (Tolentino and Borra, 2018). This is crucial for Filipino, where a single root word generates numerous forms through affixation, reduplication, and clitics, directly impacting precision and recall in search applications (Roxas and Mula, 2008).

Stemming in Filipino has been approached through various methodologies, primarily focusing on rule-based and template-based systems. The work by Tolentino and Borra (2018) introduced an exhaustive rule-based affix extraction method for stemming in Tagalog. This approach generates a tree structure where each node represents a word form derived from the input, addressing issues of understemming and unstemmed errors by exhaustively showing all stemming possibilities.

Another significant contribution is the morphological and template-based approach by Ong and Ballera (2023), which leverages predefined templates to handle the complex affixation in Filipino. While effective in capturing common morphological patterns, this method may struggle with exceptions and less frequent word forms. Enhancing this system with a hybrid model that incorporates statistical learning could improve its adaptability and accuracy.

The Tagalog Stemming Algorithm (TagSA) (Bonus, 2003) is another notable effort, focusing on extracting stems from Tagalog words through a series of linguistic rules. While TagSA provides a solid foundation for Tagalog stemming, its rule-based nature limits its scalability and adaptability to new linguistic data. Future improvements could involve the integration of neural network-based models to dynamically learn and update stemming rules.

2.3 Transformer Models and Morphological Languages

While transformer models like BERT and RoBERTa have been adapted for multilingual settings, their performance on morphologically rich languages is still an area of ongoing research. Studies have shown that these models tend to underperform on languages with complex morphology compared to analytic languages like English (Soulos et al., 2021). One reason for this underperformance is that these models are typically pretrained on large corpora where morphologically rich languages are underrepresented, leading to suboptimal contextual embeddings for these languages (Pires et al., 2019).

2.4 Hybrid Models in NLP

Hybrid models offer an effective solution for addressing the limitations of transformer models in handling morphological complexity, especially in morphologically rich languages. By combining the precision of rule-based systems with the contextual depth of neural networks, hybrid models achieve enhanced performance. For instance, Dwivedi et al. (2024) showed that integrating rule-based morphological analysis with neural machine translation (NMT) significantly improved translation quality for low-resource languages like Hindi, Marathi, and Bengali, effectively capturing grammatical rules alongside contextual fluency.

Similarly, Tong (2020) demonstrated that hybrid models in multilingual automatic speech recognition (ASR) outperformed purely neural approaches by better managing inflectional variations. This reinforces hybrid models' utility in low-resource settings, where data scarcity challenges data-driven models. Zhu et al. (2023) also emphasized that hybrid models excel at incorporating external knowledge sources, such as linguistic rules or knowledge bases, into neural architectures, making them more interpretable and effective for low-resource languages.

2.5 Filipino-Specific Transformer Models

A significant advancement in Filipino NLP emerged with RoBERTa-Tagalog, a specialized variant of the RoBERTa architecture pre-trained on large-scale Filipino corpora (Cruz and Cheng, 2022). This model demonstrates substantial improvements over previous transformer-based models across multiple benchmarks, achieving consistent performance gains of 4-5% over baseline BERT models in tasks ranging from hate speech detection to natural language inference. These improvements suggest enhanced capability in capturing Filipino's linguistic nuances and contextual relationships.

3 Methodology

3.1 Rule-Based Stemmer

Our rule-based stemmer draws from and extends established methodologies in Filipino linguistic studies, most notably the works of Bonus (Bonus, 2003), Roxas and Mula (Roxas and Mula, 2008), Rafael (Rafael, 2018), Tolentino and Borra (Tolentino and Borra, 2018), and Ong and Ballera (Ong and Ballera, 2023). These foundational studies offer robust strategies for managing affixation, infixation, circumfixation, reduplication, and morphophonemic variations, all of which are essential in accurately processing Filipino words.

3.1.1 Influences from Existing Literature

Inspired by the aforementioned works, our rule-based stemmer systematically addresses the following Filipino morphological phenomena:

- **Prefixes**: The handling of common prefixes such as *mag-*, *pag-*, and *ka-* is influenced by the strategies proposed by Bonus (Bonus, 2003), who emphasized the importance of recognizing morphophonemic changes that these prefixes can induce in root words.

- **Infixes**: Building on the framework of TagSA, our stemmer identifies and removes infixes like *-um-*, *-in-*, ensuring their correct interpretation within the context of the word (Bonus, 2003).

- **Suffixes**: The rules for removing suffixes such as *-an*, *-in*, and their allomorphic variants are guided by Tolentino and Borra's methods (Tolentino and Borra, 2018), enabling precise suffix removal without altering the meaning of the root word.

- **Circumfixes**: A layered approach to circumfixes (e.g., *ka-...-an*, *pag-...-an*) is adopted, ensuring simultaneous consideration of both prefix and suffix components, as discussed by Rafael in the context of Tagalog morphology (Rafael, 2018).

- **Reduplication**: Our stemmer adeptly handles both partial and full reduplication, a crucial feature in Filipino morphology, by applying the comprehensive analysis techniques described by Tolentino and Borra (Tolentino and Borra, 2018).

The rule-based stemmer applies these processes systematically, as illustrated in Algorithm 1, ensuring a high degree of accuracy in handling the morphological complexity of the Filipino language.

3.2 Neural Component: HybridBERTStemmer

The HybridBERTStemmer, our proposed neural component, integrates the rule-based stemmer with

Algorithm 1 Rule-Based Filipino Stemmer

```
Require: word
Ensure: stem
stem $\leftarrow$ remove_particles(word)
stem $\leftarrow$ remove_reduplication(stem)
stem $\leftarrow$ remove_circumfix(stem)
while stem changes do
    stem $\leftarrow$ remove_prefix(stem)
    stem $\leftarrow$ remove_infix(stem)
    stem $\leftarrow$ remove_suffix(stem)
end while
if stem $\in$ valid_words then return stem
elsereturn word
end if
```

a fine-tuned BERT model, creating a hybrid architecture that benefits from both linguistic rules and deep learning. This approach is grounded in recent advancements in Natural Language Processing (NLP) that demonstrate the effectiveness of combining rule-based systems with neural networks to enhance performance on complex linguistic tasks (Yambao, 2021).

3.2.1 Model Architecture

The HybridBERTStemmer architecture is designed to combine the strengths of both the rule-based and neural approaches. The architecture utilizes BERT to generate contextual embeddings for both the original word and its rule-based stem. These embeddings are then combined and passed through a classification layer to predict the most likely stem. The architecture is formally described as follows:

$$\begin{aligned} H_w &= \text{BERT}(w) \\ H_r &= \text{BERT}(r) \\ H_c &= \frac{H_w + H_r}{2} \\ y &= \text{softmax}(WH_c + b) \end{aligned} \tag{1}$$

Here, w represents the input word, r is the rule-based stem, H_w and H_r are the hidden representations from BERT, and H_c is the combined representation. The output y is a probability distribution over the possible stems.

3.3 Data and Preprocessing

The dataset used in this research comprises 16,055 Filipino words paired with their corresponding stems, sourced from the *Komisyon sa Wikang Filipino (KWF) Diksiyonaryong Filipino* (Komisyon sa Wikang Filipino, 2021). This dataset is invaluable due to its comprehensiveness and its authoritative status as a linguistic resource in the Philippines. The KWF, as the official linguistic body of the country, ensures that the dictionary encapsulates a broad spectrum of lexical variations, regional dialects, and complex morphological structures (Lee, 2010). This makes it an ideal resource for developing and rigorously evaluating stemming algorithms in Filipino.

To ensure a balanced representation of different morphological patterns, the dataset was stratified into training (70%), validation (15%), and test (15%) sets.

3.4 Training Procedure and Optimization

The training of the HybridBERTStemmer involved fine-tuning three BERT variants—BERT Multilingual, RoBERTa Tagalog, and XLM-RoBERTa—with specific optimizations to balance computational efficiency and model performance. Our implementation incorporated several key technical components:

3.4.1 Model Configuration

- **Optimizer**: AdamW with a learning rate of (2×10^{-5})

- **Batch Size**: 32, with gradient accumulation for memory efficiency

- **Epochs**: Maximum of 10, with early stopping based on validation loss

- **Loss Function**: Cross-entropy loss with mixed-precision optimization

- **Hardware**: NVIDIA L4 GPU (22.5 GB memory) with 53 GB system RAM

3.4.2 Optimization Techniques

We implemented several optimization strategies to enhance training efficiency while maintaining model accuracy and ensuring practical deployability of the system:

Mixed-Precision Training. We employed FP16 arithmetic for computation while maintaining FP32 for weight updates, reducing memory usage and training time by up to 3x while preserving numerical stability (Micikevicius et al., 2018). This dual-precision approach enabled efficient resource utilization without compromising model performance.

Gradient Accumulation. To simulate larger batch sizes while managing memory constraints, we implemented gradient accumulation (Ott et al., 2018). This technique accumulated gradients over multiple forward and backward passes, enabling effective training with larger effective batch sizes without exceeding hardware limitations.

Dynamic Learning Rate. We employed an adaptive learning rate schedule with warmup steps as described in the original transformer architecture (Vaswani et al., 2017), complemented by early stopping based on validation loss to prevent overfitting (Prechelt, 1998). Additionally, we used dynamic batching to handle variable-length inputs more efficiently.

3.4.3 Evaluation Metrics

To assess the performance of our hybrid model across various dimensions of Filipino morphological analysis, we employed a multifaceted evaluation framework. This framework encompasses both standard metrics and specialized measures tailored to the unique challenges of agglutinative languages.

Our primary metric, Exact Accuracy (A_e), quantifies the model's precision in stem generation:

$$A_e = \frac{\text{Correct Stems}}{\text{Total Predictions}} \tag{2}$$

To capture the nuanced performance in a multi-class setting, we utilized the following metrics:

- **Precision** (P): Measures the model's ability to avoid false positives, crucial for maintaining linguistic fidelity:

$$P = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \tag{3}$$

- **Recall** (R): Evaluates the model's capacity to identify all correct stems, essential for comprehensive morphological coverage:

$$R = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \tag{4}$$

- **F1-score** (F_1): Provides a balanced measure of precision and recall, particularly valuable for imbalanced datasets common in morphologically rich languages:

$$F_1 = 2 \cdot \frac{P \cdot R}{P + R} \tag{5}$$

To account for the diverse morphological patterns in Filipino, we employed two variants of the F1-score:

- **Macro F1** (F_1^M): An unweighted mean of F1-scores across all morphological classes, providing equal emphasis to rare and common patterns:

$$F_1^M = \frac{1}{|C|} \sum_{c \in C} F_1^c \tag{6}$$

where C is the set of all classes and F_1^c is the F1-score for class c.

- **Weighted F1** (F_1^W): Adjusts for class imbalance by weighting each class's F1-score by its support:

$$F_1^W = \frac{\sum_{c \in C} w_c F_1^c}{\sum_{c \in C} w_c} \tag{7}$$

where w_c is the support for class c.

By analyzing these metrics in conjunction, we can assess the model's effectiveness across various linguistic phenomena, from common affixation patterns to rare morphological constructs.

3.5 Cross-Validation and Statistical Significance

To ensure the robustness and generalizability of our results, we conducted a 5-fold stratified cross-validation. This method involves dividing the dataset into five subsets, each serving as a test set once while the remaining four subsets are used for training. This approach helps mitigate the risk of overfitting and provides a more reliable estimate of model performance across different splits of the data (Arlot and Celisse, 2010; Kohavi, 1995).

Additionally, we employed McNemar's test to evaluate the statistical significance of performance differences between the BERT variants. McNemar's test is particularly well-suited for paired comparisons of models on the same dataset, allowing us to determine whether the observed differences in accuracy between models are statistically significant or likely due to chance (McNemar, 1947; Dietterich, 1998).

3.6 Model Interpretability and Error Analysis

To further understand the model's behavior, we conducted a detailed error analysis, identifying common sources of error such as overstemming and understemming. We analyzed errors across different word lengths, affix types, and morphological complexities, using confusion matrices and other visualizations to pinpoint areas where the model struggled. This analysis provided insights into the strengths and limitations of both the rule-based and neural components, guiding further refinements of the hybrid model.

4 Results and Discussion

4.1 Model Performance

The results in Table 1 show that hybrid models outperform both untrained models and the rule-based stemmer in exact accuracy. While untrained models achieved accuracy scores between 11.11% (XLM-RoBERTa) and 11.76% (BERT Multilingual), and the rule-based stemmer reached 59.21%, the hybrid RoBERTa Tagalog attained the highest accuracy at 98.62%, a substantial improvement over both baselines. This result underscores the effectiveness of combining rule-based and neural methods for Filipino NLP.

Notably, the hybrid BERT Multilingual model performed below the rule-based baseline, highlighting the advantage of language-specific pre-training.

4.2 Computational Efficiency Across Model Variants

All models were evaluated for runtime performance covering the full inference pipeline—including input preprocessing, model inference, and postprocessing. For hybrid models, this evaluation incorporated both rule-based preprocessing and neural computation phases.

- **Hybrid BERT Multilingual:** The fastest among hybrid models, completing in 134.05s (55.67 ms/word) and achieving a 20.61% reduction in runtime compared to its untrained counterpart (168.86s).
- **Hybrid RoBERTa Tagalog:** Processed in 150.04s (62.31 ms/word), showing a 23.18% improvement over the untrained model (195.31s).
- **Hybrid XLM-RoBERTa:** Displayed the longest runtime at 230.04s (95.53 ms/word), with a slight increase of 1.36% over the untrained version (226.96s).

4.3 Statistical Significance and Ablation Study

To quantify the impact of each component in our hybrid architecture, we conducted a comprehensive ablation study and statistical significance testing using McNemar's test. The results, presented in Table 2, clearly demonstrate the necessity of integrating both rule-based and neural components.

The ablation study results show that:

1. Removing the rule-based component from the hybrid models results in a performance drop, especially for BERT Multilingual, which relies more heavily on the rule-based preprocessing.
2. The BERT-only variants further degrade in performance, emphasizing the importance of rule-based preprocessing in handling Filipino's complex morphology.
3. RoBERTa Tagalog and XLM-RoBERTa demonstrate more resilience, though their performance also benefits significantly from the hybrid approach.

4.4 Error Case Analysis

The hybrid RoBERTa Tagalog model demonstrates a trade-off in morphological processing, reducing affixation errors to 20% (compared to 45% in other models) but increasing reduplication errors to 65%, as shown in Figure 1. An in-depth error analysis, summarized in Table 3, highlights three critical challenges in Filipino morphological processing:

1. **Context-Dependent Affixation**: The high error rate in handling words like 'kinakausap' → 'kausap' demonstrates that models struggle to distinguish between core morphemes and affixes when their role is context-dependent. This suggests that purely sequential approaches to affix stripping may be insufficient for Filipino, pointing to the potential benefit of tree-structured or graph-based morphological analysis approaches.
2. **Reduplication Complexity**: The significant increase in reduplication errors in the hybrid model (65% versus 30-35% in other models) indicates that neural approaches may oversimplify reduplication patterns. Cases like 'binabasa-basa' → 'babasa' show that the model fails to recognize the semantic significance of reduplication in indicating aspect or intensity.

Table 1: Performance Metrics for BERT Variants and Rule-Based Stemmer

Model	Exact Accuracy	Precision	Recall	F1-score	Macro F1
Untrained BERT Multilingual	11.76%	85.81%	11.76%	19.80%	3.68%
Untrained RoBERTa Tagalog	11.59%	84.67%	11.59%	19.51%	3.63%
Untrained XLM-RoBERTa	11.11%	82.34%	11.11%	18.74%	3.50%
Rule-Based Stemmer	59.21%	59.21%	59.21%	59.21%	17.47%
Hybrid BERT Multilingual	56.37%	47.79%	56.37%	45.95%	1.92%
Hybrid RoBERTa Tagalog	**98.62%**	**97.65%**	**98.62%**	**98.12%**	0.57%
Hybrid XLM-RoBERTa	98.37%	97.02%	98.37%	97.68%	0.14%

Table 2: Ablation Study Results (F1-scores)

Model Variant	Full Model	No Rule-Based	BERT Only
Untrained BERT Multilingual	19.80%	-	-
Rule-Based Stemmer	59.21%	-	-
Hybrid BERT Multilingual	45.95%	43.21% (-5.9%)	39.87% (-13.3%)
Hybrid RoBERTa Tagalog	**98.12%**	**96.54% (-1.6%)**	**95.32% (-2.9%)**
Hybrid XLM-RoBERTa	97.68%	95.89% (-1.8%)	94.76% (-3.0%)

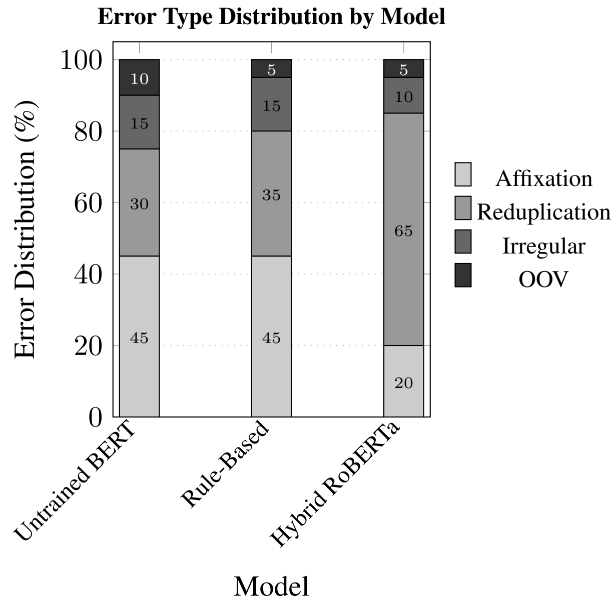

Figure 1: Comparison of error types across stemming models.

3. **Morphological Ambiguity**: Superlative forms like ‘pinakamahusay’ → ‘mahusay’ reveal a systematic failure to handle cases where multiple valid stemming options exist, depending on the intended meaning and grammatical role. This suggests the need for more sophisticated disambiguation strategies that consider broader syntactic context.

4.5 Practical Applications and Cross-Linguistic Generalizability

Our hybrid architecture demonstrates significant potential for practical applications in Filipino NLP systems, with the model achieving 98.61% accuracy on stemming tasks while maintaining reasonable processing times (ranging from 134.05s to 230.04s across different variants). This performance level makes it particularly valuable for downstream tasks such as information retrieval and text classification, where accurate morphological analysis is crucial (Tsarfaty et al., 2013). The importance of such accuracy is heightened for Filipino, where morphological complexity significantly impacts task performance (Roxas and Mula, 2008).

The success of our approach suggests broader applicability to other morphologically rich, low-resource languages through its adaptable architecture. The neural component can be extended to new languages by modifying the rule set and fine-tuning on target language data, while the modular separation of rule-based and neural components enables systematic adaptation across languages without architectural changes. Recent advances in cross-lingual transfer learning demonstrate that fine-tuning multilingual models on small language-specific datasets can significantly improve performance on previously underrepresented languages (Pires et al., 2019; Lample and Conneau, 2019).

Table 3: Representative Error Cases Across Models

Word	Correct Stem	Predicted Stem	Error Type
pinagkakaisahan	isa	kaisahan	Overly Conservative (BERT-M)
nagpapakain	kain	pakain	Partial Affixation (RB)
binabasa-basa	basa	babasa	Reduplication (RT)
kinakausap	usap	kausap	Infix Handling (BERT-M)
pinakamahusay	husay	mahusay	Superlative Form (RT)

BERT-M: Untrained Multilingual BERT, RB: Rule-Based Stemmer, RT: Hybrid RoBERTa Tagalog

This approach shows particular promise for other Austronesian languages that share morphological characteristics with Filipino, where the underlying architectural principles could be effectively leveraged to address comparable morphological challenges (Blust, 2009; Roxas et al., 2009).

5 Conclusion

This study introduces a hybrid neural-rule based architecture tailored to the morphological intricacies of the Filipino language, demonstrating the power of combining linguistic knowledge with advanced neural models. The integration of a robust rule-based stemmer with pre-trained BERT variants provides a comprehensive solution for Filipino stemming, yielding several important findings.

The RoBERTa Tagalog model emerged as the most effective, consistently outperforming both multilingual and rule-based approaches. Achieving an Exact Accuracy of 98.61% and an F1-score of 98.11%, RoBERTa Tagalog underscores the critical importance of language-specific pre-training.

The rule-based component of our architecture significantly enhanced performance, particularly in scenarios where models lacked extensive Filipino-specific pre-training. The hybrid approach consistently outperformed standalone neural models and the rule-based stemmer alone, highlighting the value of combining traditional linguistic rules with the contextual understanding provided by neural networks. This synergy is particularly evident in the model's ability to handle the rich morphological structure of the Filipino language, where complex affixation patterns and infixes challenge purely neural approaches.

Despite the overall success of the hybrid architecture, challenges remain. Reduplication continues to present difficulties, even for the high-performing models. This persistent challenge suggests the need for further refinement, potentially through specialized data augmentation strategies or more sophisticated neural architectures capable of better capturing reduplication patterns.

In addition to accuracy, the study also examined computational efficiency, revealing that RoBERTa Tagalog, while requiring moderately higher processing time (150.04s) compared to BERT Multilingual (134.05s), offers the best balance between accuracy and processing speed. This balance is crucial for practical applications, where both performance and efficiency are paramount. Statistical significance testing through McNemar's test confirms the robustness of these findings, particularly the superior performance of language-specific models over multilingual variants, reinforcing the importance of specialized architectural adaptations for morphologically rich languages.

Future research should explore advanced techniques for integrating rule-based and neural components, such as attention mechanisms or gating networks, to further enhance model performance. Targeted data augmentation could address specific challenges like reduplication, improving model robustness in handling complex morphological phenomena. Additionally, extending this hybrid architecture to other Filipino NLP tasks, such as part-of-speech tagging or named entity recognition, could demonstrate its versatility and effectiveness in various linguistic contexts.

Moreover, benchmarking this approach against emerging multilingual models and investigating transfer learning strategies across other Austronesian languages could provide further insights and broaden the applicability of this research. Such efforts would also contribute to the development of NLP tools for other low-resource languages facing similar challenges.

References

Mirna Adriani, Jelita Asian, Bobby Nazief, S M M Tahaghoghi, and Hugh E Williams. 2007. Stemming Indonesian: A confix-stripping approach. *ACM Transactions on Asian Language Information Processing*, 6(4):1–33.

Sylvain Arlot and Alain Celisse. 2010. A survey of cross-validation procedures for model selection. *Statistics Surveys*, 4:40–79.

Frank R. Blake. 1917. Reduplication in tagalog. *The American Journal of Philology*, 38(4):425–431.

Leonard Bloomfield. 1917. *Tagalog Texts with Grammatical Analysis*, volume 3 of *Illinois Studies in Language and Literature*. University of Illinois, Urbana.

Robert Blust. 2009. *The Austronesian Languages*. Pacific Linguistics, Research School of Pacific and Asian Studies, Australian National University, Canberra.

Don Erick J. Bonus. 2003. The tagalog stemming algorithm (TagSA). In *Proceedings of the Natural Language Processing Research Symposium*, Manila. De La Salle University.

Aaron John V. Boquiren, Raymond A. Garcia, Chrisrenee Jerard D. Hungria, and Joel C. de Goma. 2022. Tagalog sentiment analysis using deep learning approach with backward slang inclusion. In *Proceedings of the International Conference on Industrial Engineering and Operations Management (IEOM)*, Nsukka, Nigeria.

Charibeth Ko Cheng and Solomon See. 2006. The revised wordframe model for the Filipino language. *Journal of Research in Science, Computing and Engineering*, 3:1–1.

Jan Christian Blaise Cruz and Charibeth Cheng. 2022. Improving large-scale language models and resources for Filipino. In *Proceedings of the 13th International Conference on Language Resources and Evaluation*, Marseille, France. European Language Resources Association.

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.

Thomas G. Dietterich. 1998. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, 10(7):1895–1923.

Pankaj Kumar Dwivedi et al. 2024. Hybrid nmt model and comparison with existing machine translation systems. *Multidisciplinary Science Journal*, 7(e2025146).

Erencan Erkaya. 2022. A comprehensive analysis of subword tokenizers for morphologically rich languages. Master's thesis, Boğaziçi University.

Albert Gatt and Emiel Krahmer. 2018. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, 61:65–170.

Francis Katamba. 1993. *Morphology*. Modern Linguistics Series. Macmillan Press Ltd, London.

Ron Kohavi. 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI)*, volume 2, pages 1137–1143. Morgan Kaufmann Publishers Inc.

Komisyon sa Wikang Filipino. 2021. KWF Diksiyonáryo ng Wíkang Filipíno. An online adaptation of the 1989 Diksiyonaryo ng Wikang Filipíno, updated to reflect current linguistic and orthographic standards in Filipino.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*.

Aldrin P. Lee. 2010. The filipino monolingual dictionaries and the development of filipino lexicography. *Philippine Social Sciences Review*, 62(2):370–397.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Shervin Malmasi and Mark Dras. 2014. Language transfer hypotheses with linear SVM weights. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1385–1390, Doha, Qatar. Association for Computational Linguistics.

Paul McNamee and James Mayfield. 2004. Character n-gram tokenization for European language text retrieval. *Information Retrieval*, 7(1-2):73–97.

Quinn McNemar. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157.

Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory Diamos, Erich Elsen, David Garcia, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, and Hao Wu. 2018. Mixed precision training. In *Proceedings of the 6th International Conference on Learning Representations (ICLR 2018)*.

Hans J Nelson. 2004. A two-level engine for Tagalog morphology and a structured XML output for PC-Kimmo. Master's thesis, Brigham Young University, Provo, Utah, USA.

Great Allan M Ong and Melvin A Ballera. 2023. From a Filipino morphological and template-based stemming: A text based analyzer and design. In *2023 4th International Informatics and Software Engineering Conference*, pages 1–6. IEEE.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2018. Scaling neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 1–9.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual bert? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Lutz Prechelt. 1998. Automatic early stopping using cross validation: Quantifying the criteria. *Neural Networks*, 11(4):761–767.

Ria P Rafael. 2018. Revisiting word structure in Tagalog. *Diliman Review*, 62(2):43–63.

Neil Christian R. Riego, Danny Bell Villarba, Ariel Antwaun Rolando C. Sison, Fernandez C. Pineda, and Herminiño C. Lagunzad. 2023. Enhancement to low-resource text classification via sequential transfer learning. *United International Journal for Research & Technology*, 4(8):72–80.

Rachel Edita Roxas, Charibeth Cheng, and Nathalie Rose Lim. 2009. Philippine language resources: Trends and directions. In *Proceedings of the ACL Workshop for Asian Language Resources*, pages 131–138, Suntec, Singapore. ACL and AFNLP.

Robert Roxas and Gersam Mula. 2008. A morphological analyzer for Filipino verbs. In *Proceedings of the 22nd Pacific Asia Conference on Language, Information and Computation*, pages 467–473.

Carl R Galvez Rubino. 2002. *Tagalog-English, English-Tagalog Dictionary*. Hippocrene Books, New York.

Paul Soulos, Sudha Rao, Caitlin Smith, Eric Rosen, Asli Celikyilmaz, R Thomas McCoy, Yichen Jiang, Coleman Haley, Roland Fernandez, Hamid Palangi, et al. 2021. Structural biases for improving transformers on translation into morphologically rich languages. *Machine Translation Summit*, pages 6–15. 4th Workshop on Technologies for MT of Low Resource Languages.

Laurenz Adriel Tolentino and Allan Borra. 2018. An exhaustive rule-based affix extraction for stemming in Tagalog. In *Proceedings of the Philippine Computing Science Congress*. Computing Society of the Philippines.

Sibo Tong. 2020. *Multilingual Training and Adaptation in Speech Recognition*. Ph.D. thesis, École Polytechnique Fédérale de Lausanne.

Reut Tsarfaty, Djamé Seddah, Sandra Kübler, and Joakim Nivre. 2013. Parsing morphologically rich languages: Introduction to the special issue. *Computational Linguistics*, 39(1):15–22.

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

Moses L Visperas, Christalline Joie Borjal, Aunhel John M Adoptante, Danielle Shine R Abacial, Ma. Miciella Decano, and Elmer C Peramo. 2023. iTANONG-DS: A collection of benchmark datasets for downstream natural language processing tasks on select Philippine languages. In *Proceedings of the 6th International Conference on Natural Language and Speech Processing*, pages 316–323. Association for Computational Linguistics.

Arian N Yambao. 2021. A hybrid approach in analyzing Filipino morphology. Master's thesis, De La Salle University.

Yiming Zhu et al. 2023. Synergizing machine learning & symbolic methods: A survey on hybrid approaches to natural language processing. *Expert Systems with Applications*.