CEBBERT: A Lightweight Data-Transparent DistilBERT Model for Cebuano Language Processing Tasks

Gian Carlos Tan[Σ*] **Jhan Kyle Canlas**[Σ*] **Ren Joseph Ayangco**[Σ*]
Daeschan Blane Gador[Σ*] **Mico Magtira**[Ω] **Jean Malolos**[Ω]
Ramon Rodriguez[Ω] **Joseph Marvin Imperial**[Ω,Γ] and **Mideth Abisado**[Ω]
[Σ]Silliman University, Philippines [Ω]National University, Philippines
[Γ]University of Bath, UK
mbabisado@national-u.edu.ph

Abstract

One of the many reasons why low-resource Philippine languages struggle with research visibility can be attributed to the lack of language-optimized accessible resources, including computational models such as BERT and GPT. In this work, we make a push aligned to this initiative of democratizing resources for low-resource languages by introducing **CEBBERT**, a lightweight, data-transparent DistilBERT model for the Cebuano language processing tasks. Compared to other models, CEBBERT uses a compilation of diverse, multi-domain data sources ranging from Cebuano literary works, religious texts, news articles, translations, and speech transcripts, among others. Our results upon evaluating CEBBERT with challenging multiclass and multilabel tasks, including figures-of-speech identification and online symptom classification in Cebuano, show promising results and even outperform comparable Cebuano-based models such as MBERT and DOST-BERT.[1]

1 Introduction

In recent years, research in natural language processing (NLP) models has rapidly advanced due to the development of the Transformer architecture (Bahdanau, 2014; Vaswani et al., 2017). This led to more efficient processing of text data and a substantial increase in model performance, especially for machine translation. Deriving from this major contribution, the Bidirectional Encoder Representations from Transformers (BERT) architecture (Devlin et al., 2019) was released. This architecture used multiple encoder layers of the original Transformers, bidirectional processing, specific next-sentence prediction, and masked language modeling objectives for improved context representations of natural language understanding (NLU) tasks. With BERT, researchers were able to *finetune* more models to cater to several downstream tasks which set state-of-the-art performances in named entity recognition, language inference, text classification, and question answering (Howard and Ruder, 2018; Merchant et al., 2020; Mosbach et al., 2021).

In the context of low-resource languages, the rise of modern language models like BERT and its derivations have lagged due to the lack of the required amount of publicly available language-specific data for pre-training (Lovenia et al., 2024). For instance, Cebuano (CEB), a language spoken by roughly 20 million people primarily in the Southern and Central regions of the Philippines, boasts great cultural and linguistic diversity (Wolff, 2001). Due to the limited availability of resources such as diverse machine-readable corpora, there have not been many NLP applications being developed for the Cebuano language (Imperial et al., 2022; Aji et al., 2023).

Building on this motivation, we introduce CEBBERT, a new Cebuano-based encoder model based on the DistilBERT architecture (Sanh et al., 2019). DistilBERT is a lighter, faster, and more efficient version of BERT and uses a special knowledge distillation method to reduce the original size of BERT by 40% but preserves comparable performance across downstream NLP tasks and runs 60% faster. By creating a Cebuano-based adaptation of the DistilBERT model, we aim to expand the accessibility and usability of NLP tasks for the language. In constructing CEBBERT, we compiled diverse open-source Cebuano corpora from the web ranging from news articles, translations, transcripts, literary texts such as stories and poems, and many more.

To specify, our main contributions to this work on expanding NLP initiatives for Cebuano are two-fold:

[*]Work done during internship for the HealthPH Project at National University Philippines.

[1]Code and data: https://github.com/gctanuser/CebuanoDistilBERT

1. We introduce CEBBERT, a new lightweight DistilBERT model trained from a collection of purely open-source diverse multi-domain datasets for Cebuano language processing tasks.

2. We present an empirical evaluation of CEBBERT and showcase the efficiency and high performance of CEBBERT across two challenging unseen NLP tasks of online symptom report classification and figures-of-speech identification in Cebuano.

2 Related Works

2.1 Multilingual Language Models

Multilingual models, particularly derivations from Transformer and BERT architectures, have been studied by Wu and Dredze (2019) and Pires et al. (2019), showing that these models can perform cross-lingual generalization surprisingly well. These models also create multilingual representations, but these representations exhibit systematic deficiencies affecting certain language pairs. Their research demonstrated that a single model could effectively learn from various languages, establishing robust baselines for tasks in non-English languages. There are already existing mulitilingual models of BERT that exist, but single-language models have shown better performance in their respective languages. Examples of these models include CamemBERT (Martin et al., 2020) and FlauBERT (Le et al., 2020) for French, BERTje (de Vries et al., 2019) and RobBERT (Delobelle et al., 2020) for Dutch, FinBERT (Virtanen et al., 2019) for Finish and Spanish BERT (Cañete et al., 2023).

2.2 NLP Initiatives for Southeast Asian Languages

Research initiatives on open corpora building for low-resource languages drive the growth and development of the future of NLP. The biggest and most notable work for Southeast Asia was the SEACrowd Project[2] (Lovenia et al., 2024) led by AI Singapore and around 60+ researchers all over the world. The SEACrowd Project contains the largest multimodal catalog of online available datasets in Southeast Asian languages as well as benchmark experiments on recent open and commercial models for SEA language understanding and generation. Through the years, researchers from specific SEA member countries have also pushed their own contributions for releasing open-source SEA corpora. The works of Cahyawijaya et al. (2021, 2023) and Winata et al. (2023) covered works on local Indonesian languages for language generation, crowdsourcing, and sentiment analysis. The works of Dita et al. (2009); Dita and Roxas (2011), Oco et al. (2016), Cruz and Cheng (2022), and Visperas et al. (2023) for Philippine languages have developed from releasing small compiled resources in Filipino to releasing language models trained from modern deep learning architectures like BERT. A similar observation from Thai is seen with works from Kruengkrai et al. (2020); Noraset et al. (2021); Lowphansirikul et al. (2021) focusing on question-answering and NER systems. Overall, these initiatives, no matter how big or small, ensure the research survivability of Southeast Asian languages in the NLP scene.

2.3 Current Research in NLP for Cebuano

Focusing on Cebuano, most of the NLP works on this language have developed only very recently, which supports the need for more open-sourced and publicly available low-resource languages. For named-entity recognition (NER), the earliest work was done by Maynard et al. (2003) using software originally made for the English but was only continued after 19 years with the works of Gonzales et al. (2022) and Pilar et al. (2023) developing Cebuano-specific models for the task. In machine translation, the works of Adlaon and Marcos (2019) and Fernandez and Adlaon (2022) have focused on alleviating the alignment problem and using Filipino as the anchor language. In readability analysis and text complexity prediction, extensive works by Imperial et al. (2022); Imperial and Kochmar (2023b,a) evolved from developing Cebuano-specific models using traditional features to bigger models capturing closely similar languages such as Kinaray-a, Minasbate, and Hiligaynon which collectively improved model performances.

3 CEBBERT: A Lightweight Data-Transparent LLM for Cebuano

In this section, we discuss the main recipe for developing CEBBERT. We cover information on corpus collection and processing, pre-training and architecture details, and model configurations.

[2] https://seacrowd.github.io/seacrowd-catalogue/

Dataset	Domain	Format	Instances	Paper / Source	License
Bible Verses	Religion	phrase-level	23,296	Sermon Online	CC BY 4.0†
News Articles	News	document-level	4,250	Pilar et al. (2023)	CC BY NC 4.0
Sentences	General	sentence-level	103,378	Huggingface	CC0 1.0
Instruction Pairs	General	sentence-level	62,076	Upadhayay and Behzadan (2023)	CC BY 4.0†
Speech Transcripts	General	paragraph-level	1,933	Huggingface	CC BY 4.0†
Translations	General	phrase-level	82,752	Huggingface	CC BY 4.0†
Literary Texts	Literature	paragraph-level	348	Katitikan	CC BY 4.0†
Children's Books	Literature	paragraph-level	3,094	Imperial and Kochmar (2023a)	CC BY NC 4.0
Wikipedia	General	document-level	584	Wikipedia	CC BY SA

Table 1: Breakdown and related information of compiled diverse publicly available Cebuano datasets used for pertaining CEBBERT. We provide characteristics of each dataset, including domain, format, instances, downloadable links, source published works, and associated licenses. As a disclaimer, datasets with † have no identified specific licenses but can be accessed and used for non-commercial research. Thus, we identify a default license of CC BY 4.0 based on the nature of these datasets.

3.1 Cebuano Pre-training Data Information

To pre-train CEBBERT, we compiled all publicly available text data in the Cebuano language from the web, including resources from repositories such as Huggingface, Github, and artifacts from published papers. From this, we were able to build a diverse Cebuano corpus covering biblical texts, news articles, literary texts, Wikipedia pages, instructions, and speech transcripts. Table 1 reports the distribution of the compiled dataset from various sources with corresponding information on domain, format, instance counts, website links, paper sources, and licenses. Overall, the compiled Cebuano corpus to pretrain CEBBERT contains 253,539 unique rows of texts and a vocabulary of approximately 30,000 tokens.

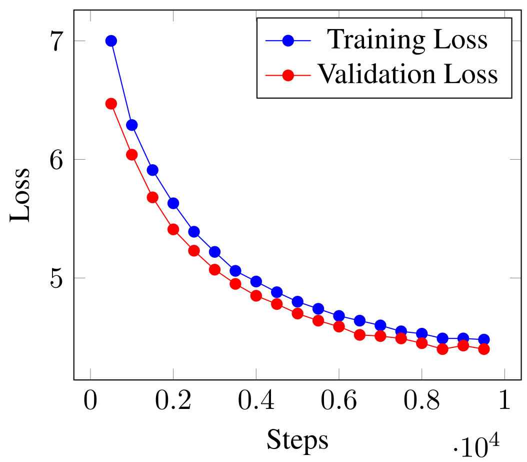

Figure 1: Loss values of pre-training the CEBBERT model using masked language modeling (MLM) and distillation training objective as done in the DistilBERT (Sanh et al., 2019) architecture.

3.2 The DistilBERT Architecture

To build a more efficient and lightweight Cebuano model, we use DistilBERT (Sanh et al., 2019) as its main architecture. DistilBERT focuses on reducing the size of the original BERT model (Devlin et al., 2019) by pretraining smaller general-purpose language representation models with knowledge distillation. Knowledge distillation is a technique wherein it extracts knowledge from the teacher and utilizes that knowledge for the student to learn and adapt (Gou et al., 2021) where the student is the compact model that is trained to reproduce the behavior of the teacher, the larger model. This concept was used for DistilBERT, which was able to retain 97% of the original BERT model's performance across downstream NLP tasks while being 40% smaller and 60% faster than BERT. To our knowledge, this work is the first publicly available Cebuano DistilBERT model trained from a diverse collection of Cebuano datasets and evaluated on unseen Cebuano language tasks.

3.3 Pretraining Configurations

The CEBBERT model was trained on a single NVIDIA Tesla L4 GPU using PyTorch and Huggingface. For hyperparameter configurations, CEBBERT model was configured with a GELU activation function, a hidden size of 768, an attention dropout rate of 0.1, and a feed-forward network hidden size of 3072 while preserving the cased function. The model used 12 attention heads across its 6 layers, with a maximum sequence length of 256 tokens. An initializer range of 0.02 was used, and dropout rates of 0.1 and 0.2 were applied to attention and classifier layers, respectively. The training process involved 3 epochs with a learning rate of

5e-05 using the compiled Cebuano corpus previously discussed. No warmup steps were employed during training. We show the trend of training and validation loss curves in Figure 1.

4 Evaluating CEBBERT for Unseen Cebuano Tasks

In this section, we describe the two NLP tasks we consider for evaluating the quality of embeddings and predictions from our CEBBERT model. We consider the datasets as *unseen* as they are newly collected and have never been published, thus making these datasets fit for evaluating Cebuano-based language models.

4.1 Task 1 - Multilabel Classification of Online Cebuano Symptom Reports

The first task we considered for evaluation is a multilabel classification task of identifying potential ailments from online symptom reports written in Cebuano. The dataset for this task was obtained from the National University Philippines's HEALTHPH: Intelligent Disease Surveillance for Public Health using Social Media Project[3] funded by the Department of Science and Technology (DOST). This dataset contains a total of 1,028 rows of social media posts across multiple platforms describing the user's expression with mentions of symptoms. Each post has been annotated by two medical professionals based on their potential to be classified in one or more possible ailments covering AURI for acute upper respiratory infection, COVID for coronavirus disease, PN for pneumonia, and TB for tuberculosis. As a multiclass classification task, one post can have more than one label from these potential ailments.

Label	Example (+ EN Translation)	Count
AURI	*Ataya ani nga hilanat oy!* *(This fever is so annoying!)*	484
COVID	*Grabe ang ubot sipon huhu* *(My cough and cold are really bad)*	403
PN	*Imbes magmayad ang ubo naglala pa* *(My cough has gotten worse)*	297
TB	*Di ako nilulubayan ng ubo ha* *(The cough won't leave me alone)*	238

Table 2: Breakdown of counts and examples for the multilabel online symptom data for Task 1. For brevity and visualization constraints, we selected shorter examples for each class.

[3] https://healthphproject.org/

4.2 Task 2 - Cebuano Figures of Speech Identification

The second task we considered for evaluation is a multiclass classification task of identifying figures of speech in Cebuano. Similar to Task 1, we also obtained this dataset from the HEALTHPH Project, specifically from the NLP Working Group. This acquired dataset was scraped from Wiktionary[4] and contains 943 rows of Cebuano figures of speech texts divided across four categories covering LITERAL or language which convey widely-accepted meaning, CATCHPHRASES or phrases that have been popularized, IDIOMS or phrases which convey subjective meaning in contrast to literals, and EUPHEMISMS or language that indirectly refer to something controversial. We use these categories as gold-standard labels for model training.

Label	Example (+ EN Interpretation)	Count
CATCH	*Klaro kaayo sa pattern* *(Clear as day)*	65
EUPH	*Anak sa hulaw* *(A short person)*	112
IDIOM	*Abot sa dunggan ang ngisi* *(To be overjoyed, extremely happy)*	619
LITERAL	*Manggihatagon* *(Generous)*	147

Table 3: Breakdown of counts and examples for the multiclass figures of speech identification for Task 2. For brevity and visualization constraints, we selected shorter examples for each class.

4.3 Finetuning and Embedding Extraction Configurations

For the finetuning setup, we set hyperparameters epoch to 5, learning rate α to 2e-05, and batch size to 32. We initially explored other values for these hyperparameters, but the aforementioned values resulted in the best performances for both multiclass and multilabel tasks. For the extraction of embeddings, from CEBBERT, MBERT, and DOST-BERT, we obtained the mean layer representations with a dimension of 768 for each instance from the task datasets. These embeddings will be used directly as features for the Random Forest model to evaluate the quality of word representations given by each model. Lastly, we use a 90-10 train-test split for each task for evaluation.

[4] Data from Wiktionary is covered by the CC BY-SA 3.0 license, which allows use and sharing in research.

4.4 Baseline Models and Metrics

As a point of performance and quality comparison, we perform the same finetuning and embedding extraction to two adjacent Cebuano-based BERT models available online: mBERT or multilingual BERT by Devlin et al. (2019) and DOST-BERT by Visperas et al. (2023). In terms of the quality of data used, mBERT was pretained using a compilation of Wikipedia dumps which includes Cebuano while DOST-BERT was pretrained with internet-scraped data from formal and informal resources. For evaluation metrics, we compute the Accuracy, F1 score, and Hamming Loss for each task. Accuracy and F1 show insight on the correctly classified labels for the models, while the Hamming loss shows how much fraction of labels were incorrectly predicted.

5 Results

In this section, we discuss the results from the experimentation procedures using the two unseen tasks in evaluating CebBERT and its adjacent Cebuano-based language models.

5.1 Model Performances for Tasks

First, we focus on the results from Task 1 on the multilabel classification of online symptoms as reported in Table 4. From the Table, we see that using embedding representations as features from the DOST-BERT model obtained the highest accuracy score of 0.367 and Hamming loss of 0.337. This is followed by embeddings from CebBERT with 0.349 and mBERT last with 0.339. The high accuracy score denotes a possibility that the embeddings from DOST-BERT were able to correctly predict the labels from the majority class. However, since the data is imbalanced for this task, we emphasize the importance of the F1 score, which CebBERT takes the lead with 0.747. A high F1 score means the model was able to balance precision and recall predictions of correct labels, especially for minority classes in the task. On the other hand, for the finetuning setup, we see a change in model performance where mBERT now takes the lead across all metrics with 0.694 in accuracy, 0.762 in F1, and 0.165 for Hamming loss. We posit that this effectiveness from mBERT for finetuning may have from the generalizability of multiple language data where the model was trained which has also been observed in previous works (Conneau and Lample, 2019). The second best-performing model comes from CebBERT with 0.664 in accuracy, 0.722 in F1, and 0.182 for Hamming loss. Interestingly, the mBERT and CebBERT were models not pretrained from Cebuano social media posts, which is the domain of the dataset in Task 1, but were the top models for this Task. From this, we believe that the quality of pretraining data is more contributive to the performance than quantity.

Overall, as a model trained from a distilled version of BERT using fewer parameters, the performance from CebBERT for Task 1 shows its efficiency and effectiveness as a qualified Cebuano-based model.

Setup	**Acc**	**F1**	**HLoss**
RF + mBERT$_{emb}$	0.339	0.721	0.340
RF + DOST-BERT$_{emb}$	**0.367**	0.708	**0.337**
RF + CebBERT$_{emb}$	0.349	**0.747**	0.339
mBERT$_{FT}$	**0.694**	**0.762**	**0.165**
DOST-BERT$_{FT}$	0.619	0.736	0.175
CebBERT$_{FT}$	0.664	0.722	0.182

Table 4: Performance of finetuned ($_{FT}$) and embedding-based Random Forest model ($_{emb}$) for Task 1 - Online Symptom Reports Multilabel Classification.

Next, we look at model performances for Task 2 on the identification of Cebuano figures of speech as reported in Table 5. From the Table, we now see even more favorable performance for CebBERT. For both the Random Forest model trained from the models' embedding features and the finetuned versions, we observe CebBERT taking the lead in terms of performance across all metrics with 0.600 and 0.879 accuracy scores, 0.588 and 0.894 F1 scores, and 0.400 and 0.054 Hamming losses for embedding and finetuned setup accordingly. These are followed by performances from DOST-BERT and mBERT. Looking at the nature of Task 2, which is in the domain of literary knowledge, the advantage of CebBERT being trained with literary datasets in the form of children's books, poems, and short stories has been instrumental in boosting its performance for identification of figures of speech.

5.2 Error Analysis

Aside from looking at model performances, we also analyze errors through misclassifications by CebBERT on specific cases by visualizing confusion matrices for each task.

Figures 2 and 3 show the disjointed per-class confusion matrices of CebBERT for setups us-

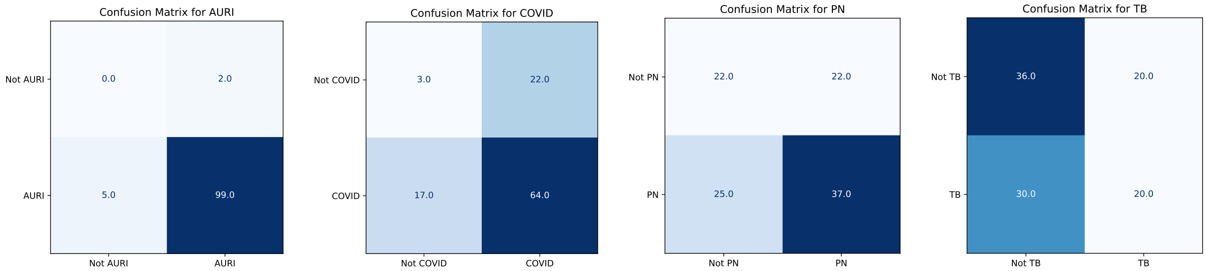

Figure 2: Confusion matrices from performance CEBBERT using Random Forest with extracted embeddings as features for Task 1 - Online Symptom Reports Multilabel Classification.

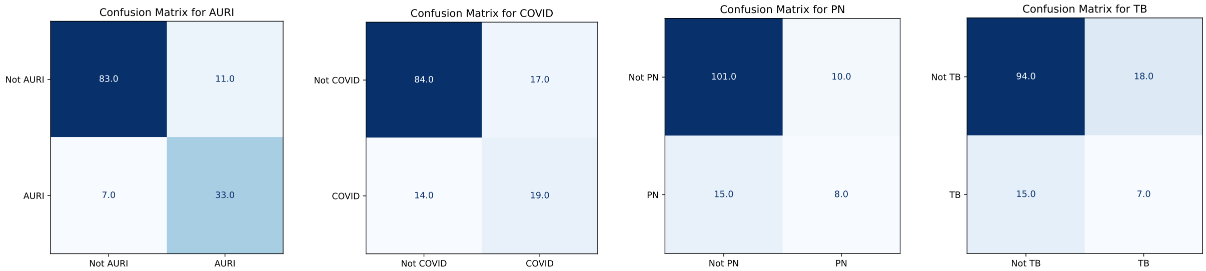

Figure 3: Confusion matrices from performance CEBBERT using finetuning for Task 1 - Online Symptom Reports Multilabel Classification.

Setup	**Acc**	**F1**	**HLoss**
RF + MBERT$_{emb}$	0.565	0.537	0.415
RF + DOST-BERT$_{emb}$	0.592	0.576	0.407
RF + CEBBERT$_{emb}$	**0.600**	**0.588**	**0.400**
MBERT$_{FT}$	0.811	0.830	0.081
DOST-BERT$_{FT}$	0.864	0.873	0.076
CEBBERT$_{FT}$	**0.879**	**0.894**	**0.053**

Table 5: Performance of finetuned ($_{FT}$) and embedding-based Random Forest model ($_{emb}$) for Task 2 - Cebuano Figures of Speech Identification.

ing Random Forest with extracted embeddings and finetuning, respectively for Task 1. From the visualizations, we see that using embeddings as features has caused some confusion to the Random Forest model trained with embeddings from CEBBERT specifically for texts with COVID and PN labels. However, this is alleviated if we move to the use of finetuning of CEBBERT itself. This change in misclassifications can be traced back to Table 4 where we see CEBBERT gaining almost double in performance in the finetuning setup (0.664 in accuracy and 0.772 in F1) compared to the embeddings approach (0.349 in accuracy and 0.747 in F1).

Figures 4 and 5 show the combined per-class confusion matrices of CEBBERT for setups using Random Forest with extracted embeddings and finetuning, respectively for Task 2. From the visualizations, we see the same trend where finetuning CEBBERT provides more stable and accurate predictions over using Random Forest and extracted embeddings as features. Likewise, this can also be traced in Table 5 where an increase in performance is observed with CEBBERT compared to other Cebuano-based BERT models. These findings from the error analysis of our work strengthen the practicality of using CEBBERT for both NLP tasks requiring extraction of representations for Cebuano texts as well as for finetuning activities with the model.

6 Discussion

Following the insights obtained from the experimental results, we put forward two main points of discussion covering the importance of diverse dataset quality for low-resource language models as well as the need for setting standards to ensure continuous growth of NLP research for the Cebuano language.

Importance of Diverse Datasets for Low-Resource Language Models Synthesizing the results and evidences found in Section 5, it is clear that the reason CEBBERT was able to obtain very comparable performance and even surpassing the only two available Cebuano-based BERT models, MBERT and DOST-BERT, is due to its data diversity and transparency. Our experience with collecting and aggregating the Cebuano datasets for pretraining CEBBERT is that sources from published papers and websites may come with small contributions but, if compiled all together, may produce a sizeable amount sufficient for exploring resource-efficient architectures such as DistilBERT. Using diverse, high-quality datasets from different domains such as news, literature, and religion enables the multipurpose usage of language models trained from these datasets. We echo the findings from Ibañez et al. (2022), where they tested a Tagalog BERT model trained purely from Tagalog Wikipedia dumps and found it impractical and low-performing for Tagalog NLP tasks such as storybook complexity classification where the input data are literary texts. Overall, we emphasize the notion of collecting diverse multi-domain datasets for pretraining language models, particularly for low-resource languages like Cebuano and other Philippine languages.

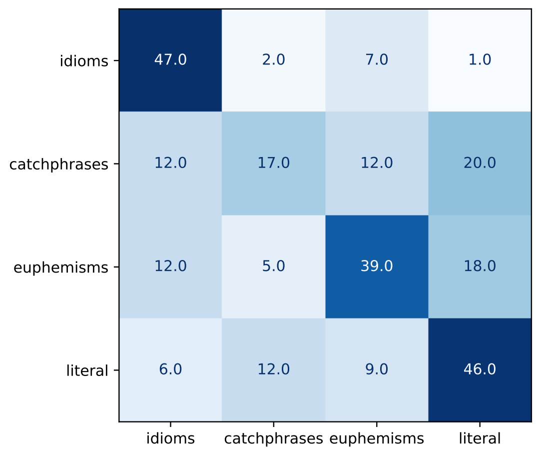

Figure 4: Confusion matrix from performance CEBBERT using Random Forest with extracted embeddings as features for Task 2 - Cebuano Figures of Speech Identification.

Setting Standards for Cebuano NLP Research The next point we want to discuss is the importance of setting good practices and following community-recognized standards for NLP research, particularly if the target languages are low-resource and the beneficial impact it will have on the community. In this work, our CEBBERT model has been trained from diverse opensource license-permitting datasets found in online repositories such as Huggingface and from published works in Cebuano (Imperial et al., 2022; Imperial and Kochmar, 2023b; Pilar et al., 2023) which future

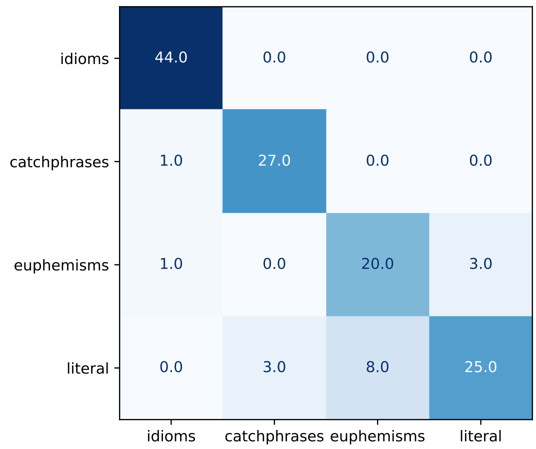

Figure 5: Confusion matrix from performance CEBBERT using finetuning for Task 2 - Cebuano Figures of Speech Identification.

research works can extend and improve. In the case of MBERT, as briefly mentioned in Section 4, it has only been trained purely with Wikipedia data which is not diverse, and issues have been raised regarding the Cebuano Wikipedia being machine-generated[5]. This is why we used only a small portion of the Cebuano Wikipedia as part of the pertaining set for CEBBERT. On the other hand, DOST-BERT (Visperas et al., 2023) was pretrained mostly with web-scraped data from formal and informal sources but have no clear or transparent breakdown of domain, data licenses, size or quantity, format, and source links or published papers, unlike what we showed in Table 1 for CEBBERT. Thus, we only consider DOST-BERT as an open weight and not an open source model due to the undisclosed nature of the research artifacts used. In summary, we consider CEBBERT as the first openly accessible language model for Cebuano following community-driven standards on dataset, artifact, and model transparency (McMillan-Major et al., 2021; Liu et al., 2024).

7 Conclusion

In this work, we introduced CEBBERT, a new lightweight and efficient model for Cebuano language processing tasks. Using the DistilBERT architecture, we pretrained CEBBERT with a diverse multi-domain collection of Cebuano data ranging from news articles, literary texts, speech transcripts, translations, and more. Through two unseen Ce-

[5] https://meta.wikimedia.org/wiki/Proposals_for_closing_projects/Closure_of_Cebuano_Wikipedia

buano NLP tasks covering figures of speech identification and online report classification, we show CebBERT effectiveness in achieving higher performance over previous larger BERT-based models in Cebuano. We envision CebBERT as the new go-to model for Cebuano NLP due to its full model and data transparency. Future works can explore using our compiled pretraining data and compare CebBERT to more advanced language model methods with Cebuano, including instruction-tuning and optimizing through feedback.

Acknowledgment

We gratefully acknowledge the support provided by the Department of Science and Technology—Philippine Council for Health Research and Development (DOST-PCHRD) for the HealthPH: Intelligent Disease Surveillance using Social Media Project through the Grants-in-Aid (GIA) Program. We also acknowledge the creators and contributors of the datasets used in this paper for their valuable work in collecting and making this data publicly available. The acquired datasets were used for non-commercial research purposes only. JMI is supported by the National University Philippines and the UKRI Centre for Doctoral Training in Accountable, Responsible, and Transparent AI [EP/S023437/1] of the University of Bath.

References

Kristine Mae M Adlaon and Nelson Marcos. 2019. Building the language resource for a cebuano-filipino neural machine translation system. In *Proceedings of the 2019 3rd International Conference on Natural Language Processing and Information Retrieval*, pages 127–132.

Alham Fikri Aji, Jessica Zosa Forde, Alyssa Marie Loo, Lintang Sutawika, Skyler Wang, Genta Indra Winata, Zheng-Xin Yong, Ruochen Zhang, A. Seza Doğruöz, Yin Lin Tan, and Jan Christian Blaise Cruz. 2023. Current status of NLP in south East Asia with insights from multilingualism and language diversity. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics: Tutorial Abstract*, pages 8–13, Nusa Dua, Bali. Association for Computational Linguistics.

Dzmitry Bahdanau. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Samuel Cahyawijaya, Holy Lovenia, Alham Fikri Aji, Genta Winata, Bryan Wilie, Fajri Koto, Rahmad Mahendra, Christian Wibisono, Ade Romadhony, Karissa Vincentio, Jennifer Santoso, David Moeljadi, Cahya Wirawan, Frederikus Hudi, Muhammad Satrio Wicaksono, Ivan Parmonangan, Ika Alfina, Ilham Firdausi Putra, Samsul Rahmadani, Yulianti Oenang, Ali Septiandri, James Jaya, Kaustubh Dhole, Arie Suryani, Rifki Afina Putri, Dan Su, Keith Stevens, Made Nindyatama Nityasya, Muhammad Adilazuarda, Ryan Hadiwijaya, Ryandito Diandaru, Tiezheng Yu, Vito Ghifari, Wenliang Dai, Yan Xu, Dyah Damapuspita, Haryo Wibowo, Cuk Tho, Ichwanul Karo Karo, Tirana Fatyanosa, Ziwei Ji, Graham Neubig, Timothy Baldwin, Sebastian Ruder, Pascale Fung, Herry Sujaini, Sakriani Sakti, and Ayu Purwarianti. 2023. NusaCrowd: Open source initiative for Indonesian NLP resources. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13745–13818, Toronto, Canada. Association for Computational Linguistics.

Samuel Cahyawijaya, Genta Indra Winata, Bryan Wilie, Karissa Vincentio, Xiaohong Li, Adhiguna Kuncoro, Sebastian Ruder, Zhi Yuan Lim, Syafri Bahar, Masayu Khodra, Ayu Purwarianti, and Pascale Fung. 2021. IndoNLG: Benchmark and resources for evaluating Indonesian natural language generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8875–8898, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

José Cañete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. 2023. Spanish pre-trained bert model and evaluation data.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. *Advances in neural information processing systems*, 32.

Jan Christian Blaise Cruz and Charibeth Cheng. 2022. Improving large-scale language models and resources for Filipino. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6548–6555, Marseille, France. European Language Resources Association.

Wietse de Vries, Andreas van Cranenburgh, Arianna Bisazza, Tommaso Caselli, Gertjan van Noord, and Malvina Nissim. 2019. Bertje: A dutch BERT model. *CoRR*, abs/1912.09582.

Pieter Delobelle, Thomas Winters, and Bettina Berendt. 2020. RobBERT: a Dutch RoBERTa-based Language Model. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3255–3265, 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.

Shirley Dita and Rachel Edita Roxas. 2011. Philippine languages online corpora: Status, issues, and prospects. In *Proceedings of the 9th Workshop on Asian Language Resources*, pages 59–62, Chiang Mai, Thailand. Asian Federation of Natural Language Processing.

Shirley N. Dita, Rachel Edita O. Roxas, and Paul Inventado. 2009. Building online corpora of Philippine languages. In *Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation, Volume 2*, pages 646–653, Hong Kong. City University of Hong Kong.

Jenn Leana Fernandez and Kristine Mae M. Adlaon. 2022. Exploring word alignment towards an efficient sentence aligner for Filipino and Cebuano languages. In *Proceedings of the Fifth Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2022)*, pages 99–106, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Joshua Andre Huertas Gonzales, J-Adrielle Enriquez Gustilo, Glenn Michael Vequilla Nituda, and Kristine Mae Monteza Adlaon. 2022. Developing a hybrid neural network for part-of-speech tagging and named entity recognition. In *Proceedings of the 2022 5th Artificial Intelligence and Cloud Computing Conference*, pages 7–13.

Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. 2021. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129(6):1789–1819.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339, Melbourne, Australia. 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 and Ekaterina Kochmar. 2023a. Automatic readability assessment for closely related languages. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5371–5386, Toronto, Canada. Association for Computational Linguistics.

Joseph Marvin Imperial and Ekaterina Kochmar. 2023b. BasahaCorpus: An expanded linguistic resource for readability assessment in Central Philippine languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6302–6309, Singapore. Association for Computational Linguistics.

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.

Canasai Kruengkrai, Thien Hai Nguyen, Sharifah Mahani Aljunied, and Lidong Bing. 2020. Improving low-resource named entity recognition using joint sentence and token labeling. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5898–5905, Online. Association for Computational Linguistics.

Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoit Crabbé, Laurent Besacier, and Didier Schwab. 2020. FlauBERT: Unsupervised language model pre-training for French. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2479–2490, Marseille, France. European Language Resources Association.

Jiarui Liu, Wenkai Li, Zhijing Jin, and Mona Diab. 2024. Automatic generation of model and data cards: A step towards responsible AI. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1975–1997, Mexico City, Mexico. Association for Computational Linguistics.

Holy Lovenia, Rahmad Mahendra, Salsabil Maulana Akbar, Lester James V Miranda, Jennifer Santoso, Elyanah Aco, Akhdan Fadhilah, Jonibek Mansurov, Joseph Marvin Imperial, Onno P Kampman, et al. 2024. Seacrowd: A multilingual multimodal data hub and benchmark suite for southeast asian languages. *arXiv preprint arXiv:2406.10118*.

Lalita Lowphansirikul, Charin Polpanumas, Nawat Jantrakulchai, and Sarana Nutanong. 2021. Wangchanberta: Pretraining transformer-based thai language models. *arXiv preprint arXiv:2101.09635*.

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

Diana Maynard, Valentin Tablan, and Hamish Cunningham. 2003. NE recognition without training data on a language you don't speak. In *Proceedings of the ACL 2003 Workshop on Multilingual and Mixed-language Named Entity Recognition*, pages 33–40, Sapporo, Japan. Association for Computational Linguistics.

Angelina McMillan-Major, Salomey Osei, Juan Diego Rodriguez, Pawan Sasanka Ammanamanchi, Sebastian Gehrmann, and Yacine Jernite. 2021. Reusable templates and guides for documenting datasets and models for natural language processing and generation: A case study of the HuggingFace and GEM data and model cards. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 121–135, Online. Association for Computational Linguistics.

Amil Merchant, Elahe Rahimtoroghi, Ellie Pavlick, and Ian Tenney. 2020. What happens to BERT embeddings during fine-tuning? In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 33–44, Online. Association for Computational Linguistics.

Marius Mosbach, Maksym Andriushchenko, and Dietrich Klakow. 2021. On the stability of fine-tuning bert: Misconceptions, explanations, and strong baselines. In *International Conference on Learning Representations*.

Thanapon Noraset, Lalita Lowphansirikul, and Suppawong Tuarob. 2021. Wabiqa: A wikipedia-based thai question-answering system. *Information processing & management*, 58(1):102431.

Nathaniel Oco, Leif Romeritch Syliongka, Tod Allman, and Rachel Edita Roxas. 2016. Philippine language resources: Applications, issues, and directions. In *Proceedings of the 30th Pacific Asia Conference on Language, Information and Computation: Posters*, pages 433–438, Seoul, South Korea.

Ma. Beatrice Emanuela Pilar, Dane Dedoroy, Ellyza Mari Papas, Mary Loise Buenaventura, Myron Darrel Montefalcon, Jay Rhald Padilla, Joseph Marvin Imperial, Mideth Abisado, and Lany Maceda. 2023. CebuaNER: A new baseline Cebuano named entity recognition model. In *Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation*, pages 792–800, Hong Kong, China. Association for Computational Linguistics.

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.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Bibek Upadhayay and Vahid Behzadan. 2023. Taco: Enhancing cross-lingual transfer for low-resource languages in llms through translation-assisted chain-of-thought processes. *arXiv preprint arXiv:2311.10797*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. 30.

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

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 (ICNLSP 2023)*, pages 316–323, Online. Association for Computational Linguistics.

Genta Indra Winata, Alham Fikri Aji, Samuel Cahyawijaya, Rahmad Mahendra, Fajri Koto, Ade Romadhony, Kemal Kurniawan, David Moeljadi, Radityo Eko Prasojo, Pascale Fung, Timothy Baldwin, Jey Han Lau, Rico Sennrich, and Sebastian Ruder. 2023. NusaX: Multilingual parallel sentiment dataset for 10 Indonesian local languages. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 815–834, Dubrovnik, Croatia. Association for Computational Linguistics.

John U Wolff. 2001. Cebuano. *Facts about the world's languages: An encyclopedia of the world's major languages, past and present*, pages 121–26.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. 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 833–844, Hong Kong, China. Association for Computational Linguistics.