iTANONG-DS : A Collection of Benchmark Datasets for Downstream Natural Language Processing Tasks on Select Philippine Languages

Moses Visperas and **Christalline Joie Borjal** and
Aunhel John Adoptante and **Elmer Peramo**
Computer Software Division, Advanced Science and Technology Institute
Department of Science and Technology, Diliman, Quezon City, Philippines
{moses.visperas,christallinejoie.borjal,aunheljohn.adoptante,elmer}@asti.dost.gov.ph

Danielle Shine Abacial
Mindanao State University
- Iligan Institute of Technology
Tibanga, Iligan City, Philippines
danielleshine.abacial@g.msuiit.edu.ph

Ma. Miciella Decano
Far Eastern University - Alabang
Alabang, Muntinlupa City, Philippines
201910804@feualabang.edu.ph

Abstract

Benchmark datasets are crucial for evaluating algorithms and models objectively. They provide a standardized basis for comparisons, promote reproducibility, and drive innovation by establishing baselines and encouraging advancements in the field. Limited benchmark datasets exist for various natural language processing tasks in low-resource languages, including most Philippine languages. As part of iTANONG's 10 billion token dataset initiative, the authors release the first iteration of iTANONG-DS[1], a collection of unlabeled and labeled datasets for different NLP tasks such as sentiment analysis, part-of-speech tagging, named entity recognition for Tagalog, and language modeling for Cebuano.

1 Introduction

Recent years have seen an exponential expansion in the field of natural language processing (NLP), revolutionizing a number of applications including information retrieval, sentiment analysis, and machine translation. Despite these developments, the lack of structured benchmark datasets, particularly for low-resource languages, continues to be a problem in NLP research (Cruz and Cheng, 2020).

While the Philippines present a plethora of languages across all of its islands, Tagalog and Cebuano come out as two of the most prominent and widely-used languages in the countries. Both languages exhibit unique linguistic intricacies that reflect the culture of their respective native speakers. Tagalog is a highly inflected language with a complex system of noun cases, verb conjugations, and prepositions. It also has a rich morphology, with many affixes that can be used to modify nouns, verbs, and adjectives. On the other hand, Cebuano is an agglutinative language, which means that words are formed by adding affixes to a root word. This can make the language seem complex and difficult to learn for speakers of other languages. Cebuano also has a rich system of noun cases, which can be used to indicate the role of a noun in a sentence.

Despite the fact that both languages are widely used by a lot of people, they are still considered to be low-resource languages in the research community. This is mainly because there are not any extensive datasets for them that would be useful for the creation and testing of NLP models and algorithms.

While formal datasets like Wiki-text (Merity et al., 2016) and OSCAR (Ortiz Suárez et al., 2019; Abadji et al., 2022) offer a sizable amount of textual data for Tagalog and Cebuano, they fall short of capturing the subtle nuances of these languages as they are utilized in everyday interactions and social media posts. The ability to comprehend these languages in their natural environments necessitates datasets that faithfully capture the dynamic essence of the language, including its informal expressions, geographical differences, and linguistic patterns found in everyday usage. The current datasets fall short of accurately portraying this heterogeneous landscape, impeding the advancement of NLP research for Philippine languages.

The authors of this work provide a thorough account of their efforts to develop task-built datasets

[1]Datasets are publicly available here: https://huggingface.co/dost-asti

for Tagalog and Cebuano, two widely used Philippine languages, primarily for various NLP applications. Recognizing the necessity for targeted and application-specific datasets, they have meticulously collected resources for sentiment analysis, part-of-speech tagging, named entity identification, and language modeling.

This work intends to encourage and promote NLP research for Philippine languages by offering these task-specific datasets and comprehensive insights into the data collection and processing methods. By enabling academics and practitioners to explore and develop in the context of various languages, these datasets significantly close a gap in the NLP research environment. The authors aspire to facilitate the creation of reliable NLP models and algorithms that can successfully manage the distinct language characteristics and difficulties of Tagalog and Cebuano by making high-quality and contextually rich datasets available.

After presenting an introduction to the existing datasets in Section 2, Section 3 proceeds to outline the dataset curation process employed in the study. The initial two subsections of Section 3 detail the data collection process, data sources, and the preprocessing steps applied to the raw data before creating specialized subsets for various downstream NLP tasks. Subsequent subsections provide a comprehensive explanation of the steps involved in curating distinct labeled datasets. Additionally, they offer a comparative analysis between these newly proposed datasets and the currently available datasets for each individual downstream NLP task, allowing for an evaluation of their quality and suitability. Finally, Section 4 gives an insight to the labeling process done on the data in this study.

2 Related Works

2.1 Monolingual Open-Source Data

Although monolingual data is readily available and widely accessible, there are still limitations with existing datasets such as WikiText-TL (Cruz and Cheng, 2019), NewsPH-NLI (Cruz et al., 2021), and the extensive parallel dataset MT560 (Gowda et al., 2021). While the first two datasets are valuable for creating language models as they are sourced from formal entities, they may not adequately capture the colloquial terms and complex Tagalog expressions used in social media and everyday contexts. On the other hand, the MT560 dataset covers a wide range of Philippine languages but predominantly consists of religious content, which may not be suitable for certain NLP tasks. These limitations underscore the need for more diverse and comprehensive datasets that encompass the intricacies of colloquial language usage and address the specific requirements of various NLP tasks.

2.2 Labeled Task Specific Data

While there are existing task-specific datasets available for certain Philippine languages, such as benchmark datasets for sentiment analysis like the Fake News dataset (Cruz et al., 2020) and the Hate Speech dataset (Neil Vicente Cabasag and Cheng, 2019), the availability of benchmark datasets for other NLP tasks remains limited. In particular, there is a scarcity of benchmark datasets for essential tasks like part-of-speech tagging and named entity recognition (NER). While WikiAnn (Rahimi et al., 2019) offers a considerable NER dataset, its main emphasis is on monolingual Tagalog and may not effectively capture the intricacies of informal language usage where code-switching between languages is prevalent.

3 Methodology

3.1 Data Gathering

To create a comprehensive text corpus, a methodical data collection technique was used, which included a wide range of sources such as government and media websites, social media platforms, and online forums. This multifaceted approach made it possible to collect a wide range of textual information, taking into account different genres, styles, and linguistic nuances.

A variety of scripting tools were used to collect the data effectively, utilizing their many features and functionalities. Notably, the data collection process was automated using tools such Selenium, BeautifulSoup, and snscrape, among others. With the help of these tools, it was possible to browse through many websites, gather pertinent data, and put together a sizable dataset.

Language	Formal	Informal
Tagalog	5,159,917	3,057,180
Cebuano	194,001	1,816,735

Table 1: Total Amount of Lines Gathered Per Language

Following the completion of the data gathering phase, the obtained text data were meticulously

Token	Regex Code	Replacement
emojis	[\U00010000 \U0010ffff]	XX_EMOJI
line beaks, feeds, etc.	([\r\n\t\f\v]+( )*)+	”.”
URLs that start with http/https	https?:\/\/([\w\-]+\.)+ ([\w \-]+)+(\/[^\s]+)*	XX_URL
<word>@<word>.<word>	[\SÑñ]+@ ([\SÑñ] +\.)+[\SÑñ]+	XX_EMAIL
URLs that end with com, net, org, co, us, ph	([\w\-]+\.)+(com| net| org | co | us | ph | io) (\/[^]+)*	XX_URL
Starts with @	@[^\s.,!?]+	XX_USERNAME
Starts with #	#[a-zA-ZÑñ0-9_]+	XX_HASHTAG

Table 2: Pre-processing done on the corpus, patterned from the work of Velasco et al. (2022)

divided into two major categories: formal and informal texts. The source of the data and its innate qualities served as the foundation for this categorization.

Social media posts from prominent personalities and the government were taken into account as articles for formal writing. These posts retained a formal tone appropriate for official communication channels because they came from reliable sources. Incorporating this subset of social media content served the purpose of capturing the subtleties of formal language usage in a digital setting.

On the other hand, information found through keyword searches of social media data and online forums was included in the category of informal writing. Online communities are renowned for enabling casual dialogue and debate, frequently displaying user-generated content and colloquial language usage. With the help of these sources, it was possible to accurately represent the informality and variety of language idioms that characterize typical online conversations.

By meticulously categorizing the data into formal and informal subsets, it was ensured that the dataset had a vast range of text types ranging from official correspondence to casual internet interactions. Table 1 shows the amount of lines of text gathered per language.

3.2 Pre-processing

A preprocessing methodology inspired by the work of Velasco et al. (2022) was used in this study. By addressing particular components that frequently appear in online textual information, this strategy intended to improve the quality and consistency of the text data. In the preparation phases, special tokens were used to substitute emojis, emails, URLs, and usernames. Sentences containing tokens that had less than three tokens were also removed. These special tokens are highlighted in Table 2.

Additionally, a filtering step was added to the preprocessing pipeline to get rid of phrases that had sentences with less than three tokens. The goal of this phase was to remove very short sentences from the dataset because they are likely to have little semantic relevance and might possibly contribute noise. This criterion was enforced to guarantee that the final dataset had a greater level of coherence and significance.

The goal was to improve the text data's quality, consistency, and interpretability by using this methodical and thorough preprocessing methodology. This would then allow for more reliable and accurate downstream NLP analyses and models.

3.3 Sentiment Analysis

The authors utilized Latent Dirichlet Allocation (LDA) (Blei et al., 2003) – a topic modeling technique – in constructing the sentiment analysis dataset. LDA helped identify relevant topics within the data, ensuring diversity. We randomly selected 9,000 sentences from three LDA-identified subjects.

To enhance dataset quality and granularity, the `GPT 3.5` model was used to classify 500 phrases as positive, negative, or neutral. Additionally, sentence embeddings were obtained with the help of sentence transformers in order to capture semantic nuances.

The extracted embeddings were used as a feature in a label propagation method – leveraging Scikit-learn and a radial basis function kernel – to propagate labeled sentiments to unlabeled data. This process generated a substantial collection of predicted sentiment labels which were then manually validated.

The entire labeling process resulted in a dataset, which contained two subsets: one with two unbalanced sentiment distributions and the other with a balanced distribution. Details of the labeled datasets are shown in Table 3.

Dataset	Positive	Negative	Neutral
Balanced	3000	3000	3000
Unbalanced-A	1203	2827	4970
Unbalanced-B	1354	2825	4821

Table 3: Sentiment Analysis Dataset Size

We benchmarked the datasets for sequence classification using an uncased Tagalog RoBERTa model fine-tuned over 45 epochs using the transformers library provide by Huggingface. In Table 4, we present the validation and test scores, and we also include results for a binary classification task using the HateSpeech dataset (Cruz and Cheng, 2022) for comparison. Despite the differences between our new dataset and the hate speech dataset in terms of content, number of labels, and the presence of code-switching, an interesting observation emerges. Even when we extend the training epochs for our dataset, it consistently yields lower scores compared to training with the hate speech dataset, which reaches early stopping at 15 epochs. This discrepancy in training outcomes underscores the unique challenges and nuances, particularly the code-switching aspect, inherent in our benchmark dataset.

Dataset	Validation Acc.	Test Acc.
Balanced	63.55%	65.33%
Unbalanced-A	75.44%	76.11%
Unbalanced-B	67.5%	68.78%
Hatespeech*	78.07%	99.03%

Table 4: Benchmark scores for sentiment classification using a fine-tuned RoBERTa Model

3.4 Part-of-Speech Tagging

In crafting the POS(Part-of-Speech) tagger dataset, the researchers selected initial data points from the pool of scraped news articles. Sentence tokenization follows this process which involves splitting the articles to individual sentences.

To efficiently handle the annotation process with limited time and minimal human effort, the researchers adapted by using a model called SMT-POST (Nocon and Borra, 2016) - a statistical machine translation approach for POS tagging on the collection of sentences. This model was selected as it achieved a higher score at 84.7% compared to earlier POS token classifiers. In order to select high-quality data points, the researchers applied a few filtering mechanisms such as the removal of five-word sequences or fewer. Overall, there were 3,919 tagged sequences with the MGNN tagset. Table 5 shows the word statistics of the tagged chosen data points.

Category	Count
Number of Words	71,444
Vocabulary Size	15,636
Min words per sequence	6
Max words per sequence	26

Table 5: Word Statistics of Part-of-Speech Text Data

The researchers also observed the POS tags' co-occurrence patterns to demonstrate diversity in terms of syntactical structure. The dataset exhibited a total count of 3,868 unique POS patterns, revealing a wide range of nuances in the collected sentences. To get a clearer observation, they extracted the trigrams in each sequence and calculated the frequency distribution of the corresponding POS tags.

Pattern	Count
FW FW FW	2371
NNP NNP NNP	993
NNC CCB NNC	473
DTP NNP NNP	429
CCT NNC CCB	359

Table 6: Frequency Distribution of Top 5 POS Co-occurrence Pattern

Table 6 shows the top 5 most common POS co-occurrence indicating prevalent code-switching as evidenced by the *FW* tag for foreign words.

The researchers fine-tuned a Tagalog RoBERTa model as a token classifier using the curated POS dataset for 30 epochs. Table 7 displays the validation and test accuracy that achieved a high score of

```
You are tasked to label a Tagalog sentence with a named entity tag using
the IB02 format.

Use contextual clues to identify if the word is indeed a named entity.

You will be given an input sentence and you will return the named entity
labels like in the following example:

Input: Sinabi ni Hindun Angsa...

Labels: Sinabi[[O]] ni[[B-PER]] Hindun[[I-PER]] Angsa,[[I-PER]]...

DO NOT explain the labels
```

Figure 1: Prompt used for Named-Entity Recognition Task

Dataset	Validation Acc.	Test Acc.
iTANONG-POS	93.10%	92.84%

Table 7: Benchmark scores for Part-of-Speech using a fine-tuned RoBERTa Model

more than 90%. This findings indicate that the prior automated labeling process by SMTPOST resulted to tags with utmost syntactic consistency despite the nuances caused by code-switching. Additionally, iTanong used real-world Taglish texts from media compared to existing researches like CRF-POST(Olivo et al., 2020) that utilized manually translated Wikipedia texts.

3.5 Named Entity Recognition

The researchers gathered 6,230 data points and employed the `GPT 3.5` model to label named entity for each word. To ensure consistent tags, IOB2 tagging format was adopted. In this format, the identified named entity tags are prefixed with *B-* where it begins and *I-* for the subsequent words that are part of the entity. A word that is not a named entity is tagged *O*.

The researchers directed the model with the system payload shown in Figure 1.

The model was instructed with two explicit key points. Firstly, was making contextual inferences to recognize that word entity type may vary depending on the context. For instance, the phrases *Sinabi ng Malacañáng* and *Ginanap sa Malacañáng* uses a common word *Malacañáng* differently as an organizational representative body in one case and as a location of the Philippine President's office in the other.

However, formulating a prompt this way causes `GPT 3.5` to reason out after labeling which results in the excessive generation of tokens. In order to address this issue, another statement was added instructing `GPT 3.5` not to explain the labels. This way of prompting helped the researchers circumvent their problem and ensured the desired result from the model.

Classification	Count
O	124,623
B-PER	4,350
I-PER	4,518
I-ORG	2,773
B-ORG	2,171
I-LOC	1,654

Table 8: Frequency Distribution of 7 Named Entity Tags used in iTanong

The researchers ran the model at a low *temperature* (0.1) to attain a realistic and predictable set of labels. A total of 143,012 named entities were identified by the model with 7 unique classifications. Table 8 shows the distribution of the most frequent tags produced by `GPT 3.5`.

The Tagalog RoBERTa model was fine tuned for 30 epochs to investigate whether the iTANONG NER outperforms WikiAnn in the named entity classification task. It's important to note that the respective test sets for each model were employed, with the iTANONG model being evaluated on the iTANONG test set and the WikiAnn model being assessed on the WikiAnn test set. As depicted in Table 9, WikiAnn performed better at the task by a significant amount.

3.6 Pre-Trained Word Embedding Models

Word embeddings have been a game-changing and groundbreaking force in the field of natural language processing (NLP) ever since they were first used, revolutionizing a variety of tasks within the

Dataset	Validation Acc.	Test Acc.
iTANONG-NER	92.63%	91.10%
WikiAnn*	97.25%	97.53%

Table 9: Benchmark scores for Part-of-Speech using a fine-tuned RoBERTa Model

discipline (Si et al., 2019). These embeddings function as effective numerical representations of words, utilizing the power of neural networks to precisely capture the semantic and grammatical subtleties of words while encoding their contextual essence into multi-dimensional vectors (Cheng, 2022). This outstanding capability has resulted in their undeniable superiority in improving performance across a wide range of downstream NLP activities, establishing their position as a remarkably dependable method of word representation (Ravindran and Murthy, 2021).

In this section, the proponents are pleased to present a set of word embeddings that were developed using the Formal text dataset from the renowned corpus which has been thoroughly detailed in previous sections. These unique embeddings were created utilizing two well-known and distinct techniques, namely Word2Vec (Mikolov et al., 2013) and FastText (Bojanowski et al., 2016). The proponents systematically constructed both the continuous bag of words (CBOW) and skip-gram variants for every technique. Furthermore, each model is provided in a variety of vector dimensions, ranging from the simple 20 to the intricate 300, allowing for a wide range of representation options to choose from. Specifically, there are six different vector sizes to choose from, namely 20, 30, 50, 100, 200, and 300.

The generated models have been saved in both `.bin` and `.txt` formats to enable maximum ease and accessibility while supporting a variety of application scenarios. The combined result of these efforts is an astonishing ensemble of 24 unique models, each of which is available in two different file formats.

3.7 Unlabeled Data & Pre-Trained Language Models

The authors conducted the pretraining of multiple BERT-based language models on the remaining unlabeled data in both Tagalog and Cebuano. Recognizing the necessity of training models from scratch to effectively capture the linguistic subtleties and intricacies inherent to these languages, the research team embarked on this endeavor.

They implemented an 80-20 data split, allocating the available dataset between training and validation sets, thereby enabling meticulous model evaluation and ensuring a robust training process. Despite the existence of Tagalog models, the authors opted for a rigorous training approach on a comprehensive dataset that encompassed both formal and informal language usage. This deliberate incorporation of informal language in the training data was aimed at enhancing the model's ability to adeptly address the diverse linguistic variations encountered in real-world contexts.

In a noteworthy development, the authors pretrained a BERT-Cebuano model, which was a ground-breaking feat. This innovative initiative is, to the best of their knowledge, the first-ever attempt to train a BERT-based language model exclusively for Cebuano. The lack of NLP models designed specifically for Cebuano, a language with few resources and little research attention, is addressed by this study's concentration on Cebuano.

The goal of the work was to build reliable language models that could accurately capture the nuances of Tagalog and Cebuano, hence facilitating subsequent downstream NLP tasks. Results of the model training can be seen on Table 10.

Model	Language	Validation Perplexity
BERT	Tagalog	8.3493
	Cebuano	52.3625
RoBERTa	Tagalog	9.4295
	Cebuano	52.3799

Table 10: Released Pre-Trained Language Models

4 Analysis of Labeling Process

In terms of data labeling, particularly in the context of sentiment analysis where a substantial portion of our labeling process was automated, it's worth noting a limitation. On average, approximately 86% of the entire dataset was correctly labeled through automation – from ChatGPT labeling up to the label propagation. However, there is room for improvement in terms of accuracy, especially if there are plans to expand the dataset with additional labels in the future. Finding more accurate labeling methods or refining the existing automation process could enhance the overall quality of the dataset.

A few challenges were also observed in labelling

named entities. Firstly, ChatGPT tend to generate explanations crafted like a user feedback even though it was explicitly prompted to avoid additional details. This impediment resulted to some parsing errors. Then, it produced a lot of redundant tags like *B-PER* and *B-PERSON*. In our earlier attempts, the model produced quite a number of peculiar tags like *B-profanity*, *B-COLOR*, *B-RNA* among others. Finally, the researchers decided to limit the annotation for this iteration into 7 tags, comprising of three broad entity classes *person (PER)*, *location (LOC)* and *organization (ORG)*. The researchers are dedicated to refine the dataset in the next iterations of iTanong-DS from manually scrutinizing the labels to effectively utilizing emerging tools for a semi-automated annotation process.

5 Conclusion

In this paper, the authors present iTANONG-DS, an extensive collection of unlabeled and labeled datasets that have been carefully curated to cater to a wide range of Natural Language Processing (NLP) applications. Specifically, these datasets are designed to facilitate tasks such as sentiment analysis, part-of-speech tagging, named entity recognition, and language modeling for Tagalog and Cebuano languages. Alongside the datasets, they have developed pretrained embeddings and language models specifically tailored to these languages, thereby establishing a strong foundation for NLP research and enabling advancements in Philippine language processing.

In conclusion, iTANONG-DS, along with its accompanying pretrained embeddings and language models, serves as a valuable resource for researchers and practitioners working in the field of Philippine language processing. By providing comprehensive datasets, robust machine learning techniques, and specialized models, this work aims to foster advancements in sentiment analysis, part-of-speech tagging, named entity recognition, and language modeling for Tagalog and Cebuano. It is the hope that the availability of iTANONG-DS will stimulate further research and innovation in NLP for Philippine languages, contributing to the development of sophisticated language technologies and applications tailored to the unique linguistic characteristics of these languages.

Limitations

In the dataset presented in this paper, the tags for sentiment analysis are currently limited to three, and for Named Entity Recognition (NER), there are seven tags. However, there is room for expansion in terms of the number of possible labels. For sentiment analysis, this would involve adding more emotional categories beyond the current three, while for NER, it entails introducing more specific labels. Concurrently, the plan is to increase the number of labeled sentences within this dataset to enhance its comprehensiveness and applicability.

Additionally, it's worth noting that although there is a substantial amount of unlabeled Cebuano dataset, curation of task specific datasets was impossible due to the lack of native Cebuano speakers in the team. Also note that the pre-trained language models may lack capability when dealing with long sequences since the majority of the data used to train the models were taken from social media posts where sequence lengths are limited.

In this paper, the authors also released a collection of pre-trained word embeddings. However, these are only in the word2vec and fasttext formats. GloVE embeddings were not included in the collection.

References

Julien Abadji, Pedro Ortiz Suarez, Laurent Romary, and Benoît Sagot. 2022. Towards a cleaner document-oriented multilingual crawled corpus. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4344–4355, Marseille, France. European Language Resources Association.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomás Mikolov. 2016. Enriching word vectors with subword information. *Computing Research Repository*, arXiv:1607.04606.

Jerome Cheng. 2022. Neural network assisted pathology case identification. *Journal of Pathology Informatics*, 13:100008.

Jan Christian Blaise Cruz and Charibeth Cheng. 2019. Evaluating language model finetuning techniques for low-resource languages. *Computing Research Repository*, arXiv:1907.00409.

Jan Christian Blaise Cruz and Charibeth Cheng. 2020. Establishing baselines for text classification in low-resource languages. *Computing Research Repository*, arXiv:2005.02068.

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.

Jan Christian Blaise Cruz, Jose Kristian Resabal, James Lin, Dan John Velasco, and Charibeth Cheng. 2021. Exploiting news article structure for automatic corpus generation of entailment datasets. *Computing Research Repository*, arXiv:2010.11574. Version 3.

Jan Christian Blaise Cruz, Julianne Agatha Tan, and Charibeth Cheng. 2020. Localization of fake news detection via multitask transfer learning. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2596–2604, Marseille, France. European Language Resources Association.

Thamme Gowda, Zhao Zhang, Chris Mattmann, and Jonathan May. 2021. Many-to-English machine translation tools, data, and pretrained models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 306–316, Online. Association for Computational Linguistics.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2016. Pointer sentinel mixture models. *Computing Research Repository*, arXiv:1609.07843.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *Computing Research Repository*, arXiv:1301.3781.

Sean Christian Lim Mark Edward Gonzales Neil Vicente Cabasag, Vicente Raphael Chan and Charibeth Cheng. 2019. Hate speech in philippine election-related tweets: Automatic detection and classification using natural language processing. *Philippine Computing Journal*, XIV(1).

Nicco Nocon and Allan Borra. 2016. SMTPOST using statistical machine translation approach in Filipino part-of-speech tagging. In *Proceedings of the 30th Pacific Asia Conference on Language, Information and Computation: Posters*, pages 391–396, Seoul, South Korea.

John Francis T. Olivo, Prince Julius T. Hari, and Michael B. dela Fuente. 2020. Crfpost: Part-of-speech tagger for filipino texts using conditional random fields. In *Proceedings of the 2019 2nd International Conference on Algorithms, Computing and Artificial Intelligence*, ACAI '19, page 444–449, New York, NY, USA. Association for Computing Machinery.

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

Afshin Rahimi, Yuan Li, and Trevor Cohn. 2019. Massively multilingual transfer for NER. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 151–164, Florence, Italy. Association for Computational Linguistics.

Renjith Ravindran and Kavi Murthy. 2021. Syntactic coherence in word embedding spaces. *International Journal of Semantic Computing*, 15:263–290.

Yuqi Si, Jingqi Wang, Hua Xu, and Kirk Roberts. 2019. Enhancing clinical concept extraction with contextual embeddings. *Journal of the American Medical Informatics Association*, 26(11):1297–1304.

Dan John Velasco, Axel Alba, Trisha Gail Pelagio, Bryce Anthony Ramirez, Jan Christian Blaise Cruz, and Charibeth Cheng. 2022. Towards automatic construction of filipino wordnet: Word sense induction and synset induction using sentence embeddings. *Computing Research Repository*, arXiv:2204.03251.