Progress Report 2:
Parallelizing BERT-based Detection of Fake News in the Filipino Language

Mary Nathalie Dela Cruz

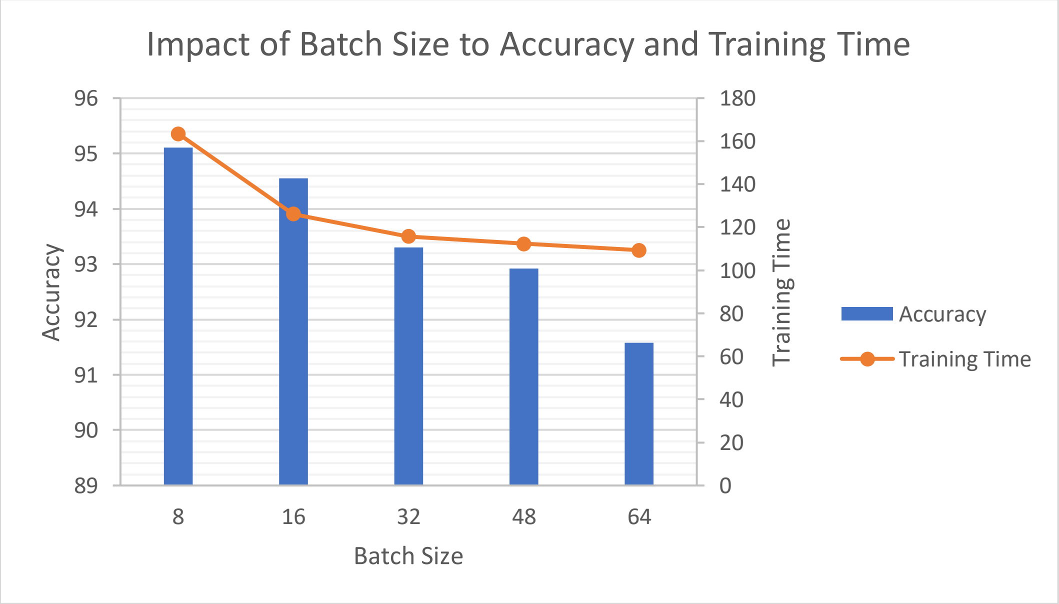
The research objective is to evaluate the impact of running GPUs for a fake news classification task in Filipino context. The programming language is Python.

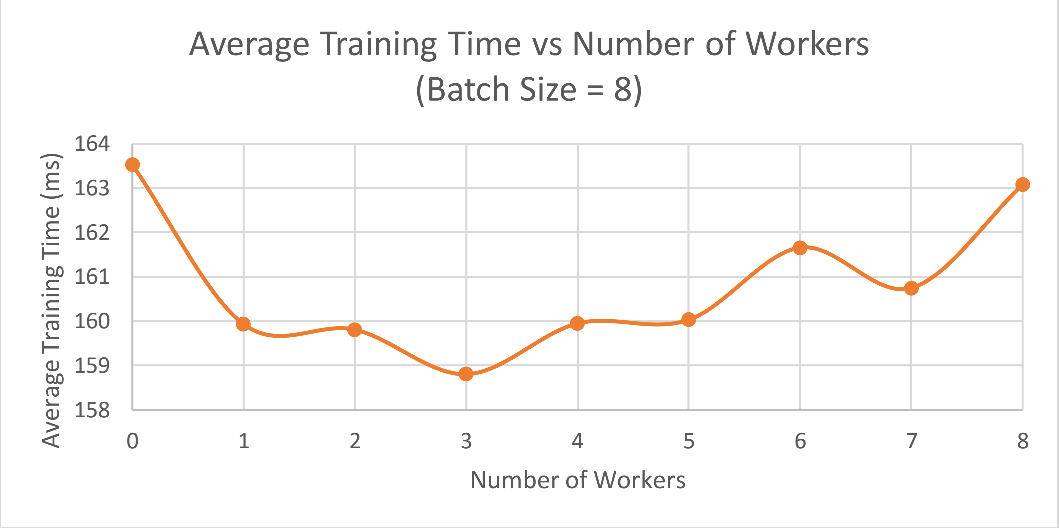
Workflow

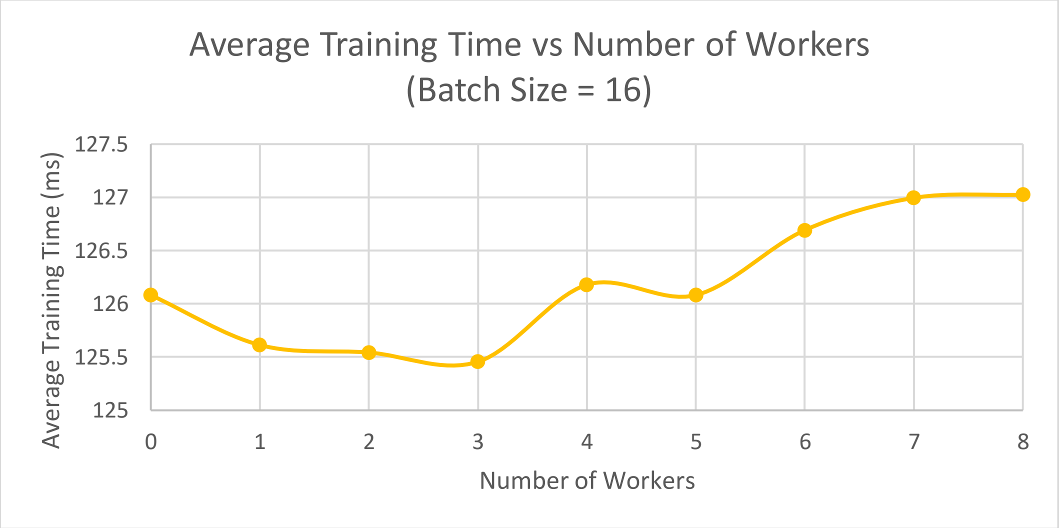
1. Split the dataset by extracting the 'article' column as features (X) and the 'label' column as labels (y) from the dataset downloaded from HuggingFace.
2. Load the BERT tokenizer outside of for loops to avoid redundancy. Tokenize and encode the dataset using the BERT tokenizer.
3. Define a function that computes evaluation metrics such as accuracy, precision, recall, and F1 score.
4. Create a custom dataset class to handle encodings and labels. This class can initialize with encodings and labels, and define methods to get items by index and to get the length of the dataset.
5. Set up hyperparameters like batch sizes, the number of workers, learning rates, and the number of epochs.
6. Perform k-fold cross-validation and training by iterating through each combination of learning rate, batch size, and epoch, where for each fold, split the data and create custom datasets made in Step 4 for training and validation.
7. Load the BERT model for sequence classification and define training arguments. Initialize the Trainer with the model, training arguments, datasets, and metrics.
8. Train the model and evaluate it using the function made in Step 3.
9. Collect the evaluation results and training details, then append them to a results list. Save the results to a CSV file.

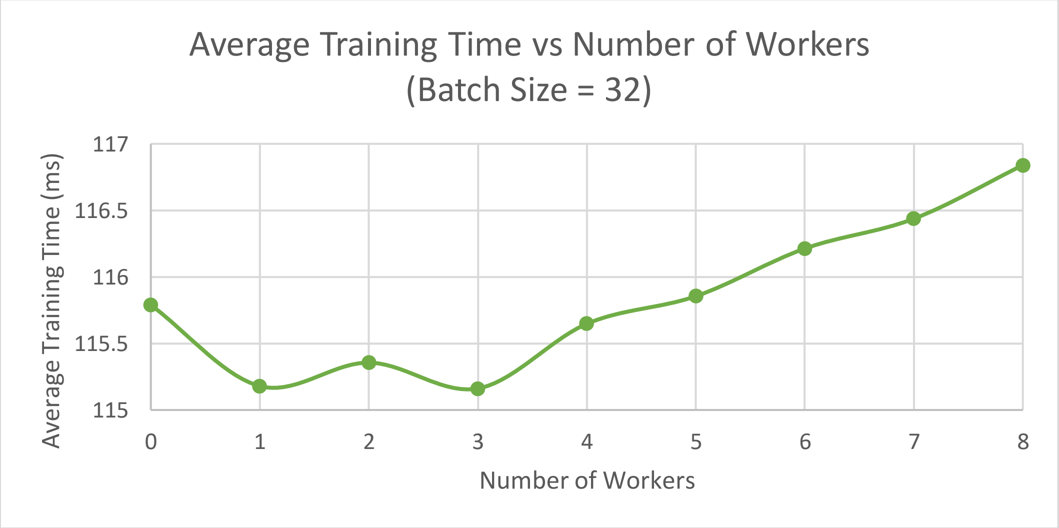
Sample Figures

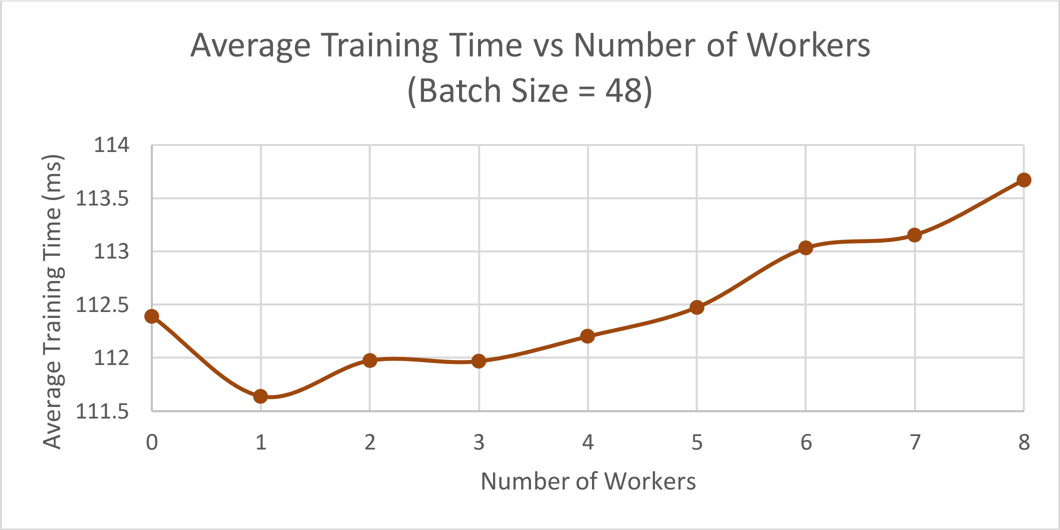
Parameters Investigated: Number of Workers and Batch Size

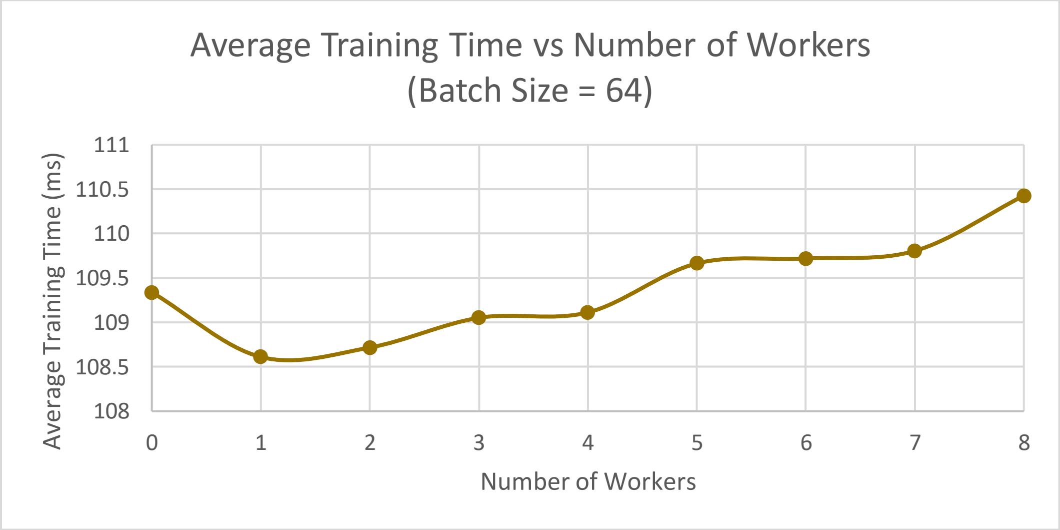


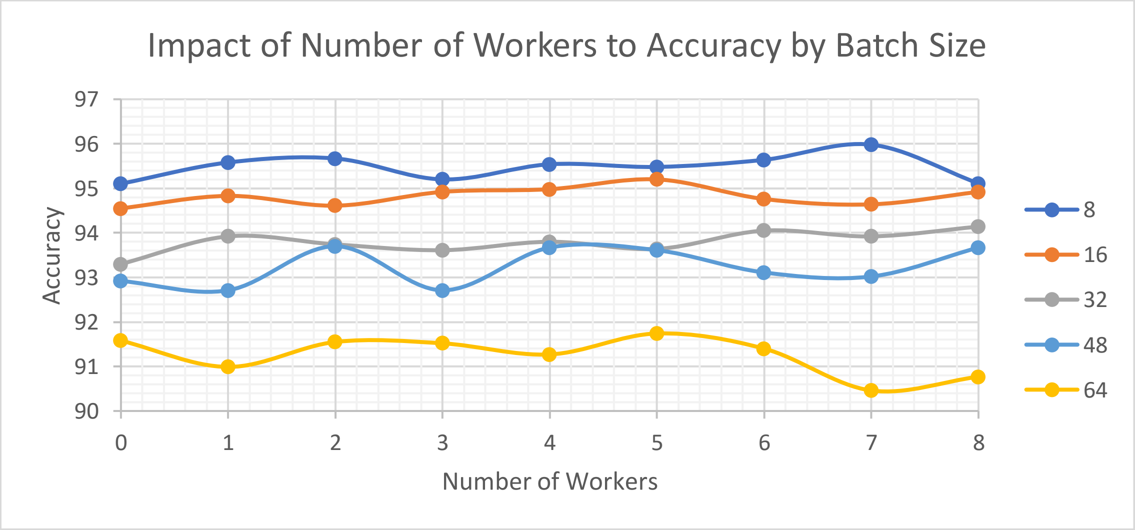

Average Training Time vs Number of Workers
(Batch Size = 8)
Average Training Time (ms)
164
163
162
161
160
159
158
0
1
2
3
4
5
6
7
8
Number of Workers

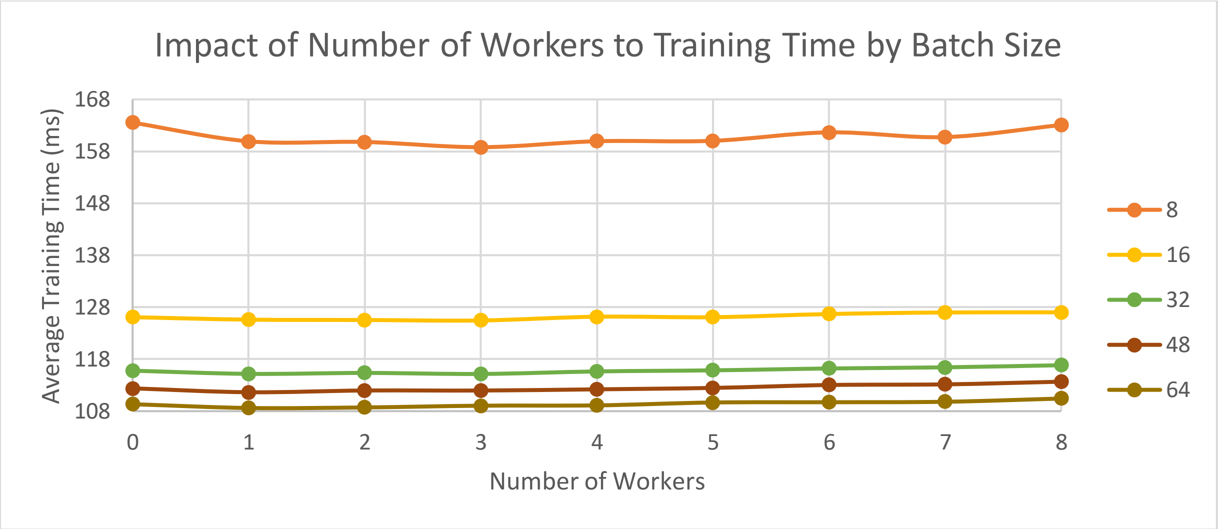

Average Training Time vs Number of Workers
(Batch Size = 16)
Average Training Time (ms)
127.5
127
126.5
126
125.5
125
0
1
2
3
4
5
6
7
8
Number of Workers

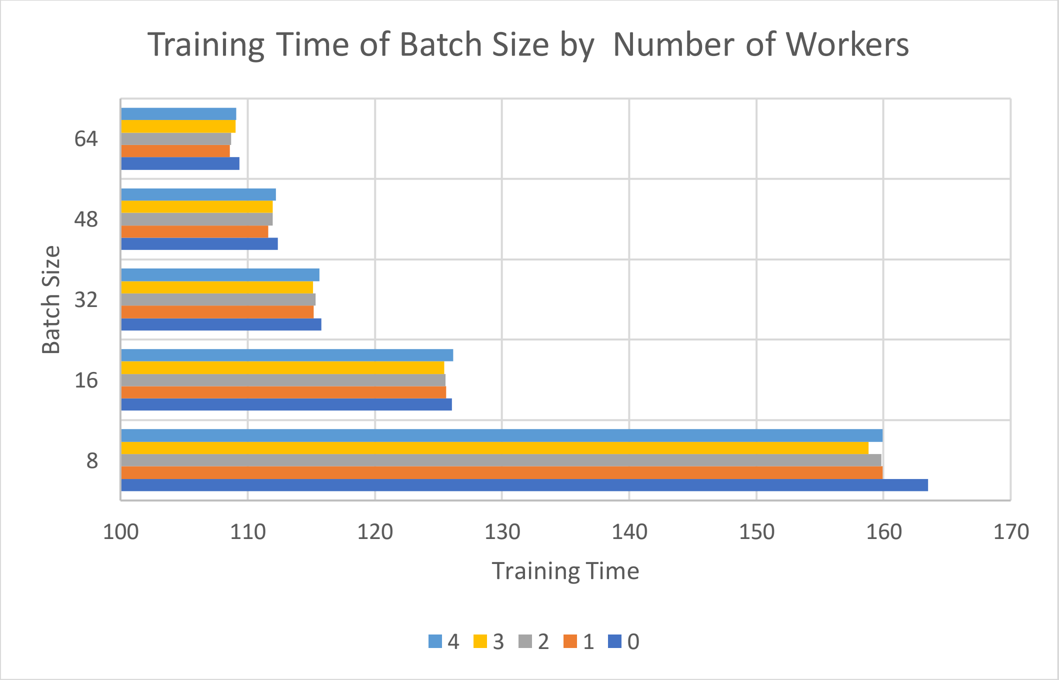

Average Training Time vs Number of Workers
(Batch Size = 32)
Average Training Time (ms)
117
116.5
116
115.5
115
0
1
2
3
4
5
6
7
8
Number of Workers


Average Training Time vs Number of Workers
(Batch Size = 48)
Average Training Time (ms)
114
113.5
113
112.5
112
111.5
0
1
2
3
4
5
6
7
8
Number of Workers

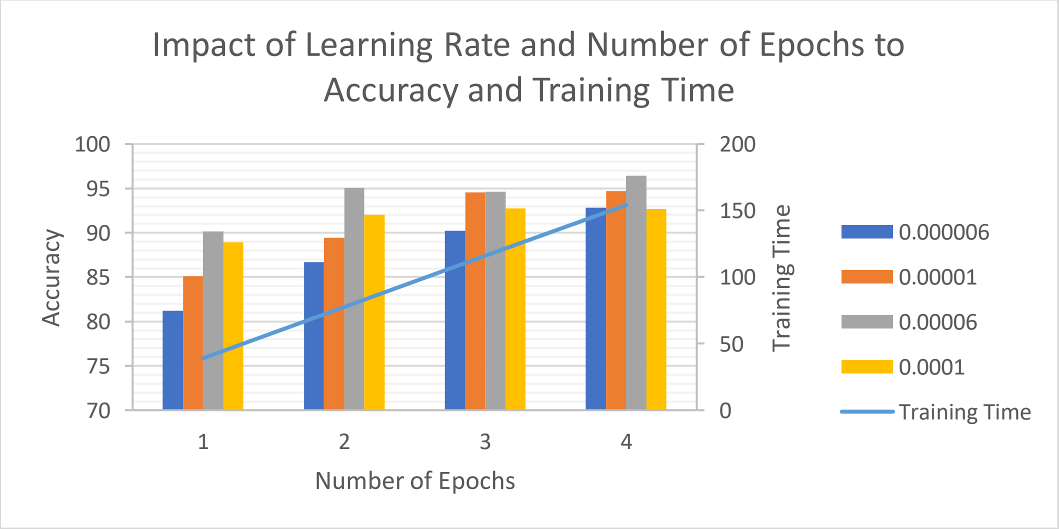

Average Training Time vs Number of Workers
(Batch Size = 64)
Average Training Time (ms)
111
110.5
110
109.5
109
108.5
108
0
1
2
3
4
5
6
7
8
Number of Workers


Impact of Number of Workers to Accuracy by Batch Size
Accuracy
97
96
95
94
93
92
91
90
0
1
2
3
4
5
6
7
8
Number of Workers
8
16
32
48
64





Parameters Investigated: Number of Workers and Batch Size


Impact of Learning Rate and Number of Epochs to Accuracy and Training Time
Accuracy
Training Time
Number of Epochs
Training Time