GROUP 2: Vinoya, Shaina V.
Acosta, Ryan Jay M.
Sagloria, Jonny

Title: Development of TAGA-TSEK: An Educational Grammar Checker Tool for Pokus ng Pandiwa Explored using DistilBERT Algorithm

FINALS
STATSPRO/ELECTIVE4

I. Model Training

- Introduction

Grammar checking serves as a foundational application in the realm of natural language processing (NLP) due to its pivotal role in verifying the accuracy of input sentences. The correctness of sentences significantly impacts various other NLP applications. To ensure the validity of sentences, an underlying grammar of the natural language is utilized. This grammar comprises a set of rules that dictate the structure and combination of sentence components, such as clauses and phrases. A properly formed phrase is one where all individual words harmonize with each other, satisfying the necessary morphological and syntactic agreement features.

Within the broader domain of grammar checkers, the specific focus of this study revolves around the different kinds of *Pokus ng Pandiwa* in the Tagalog language. Previous research in grammar checking has primarily focused on detecting syntactical errors such as spelling, punctuation, word form and word usage. This niche research area addresses the need for an educational tool grammar checker for the Tagalog language by utilizing a machine learning model to facilitate the training processes to recognize and apply these grammatical rules, specifically *Pokus ng Pandiwa*.

The core problem that this research endeavors to tackle is the development educational grammar checker tool that focuses on the area of *Pokus ng Pandiwa*. Increasing the comprehension of *Pokus ng Pandiwa* can improve the writing capabilities of the students. This endeavour seeks not only to improve foundational knowledge of these grammatical elements but also to inspire innovative solutions that simplify the development of writing skills, ultimately contributing to more proficient and effective communication.

- Dataset Collection:

The researchers used an artificial intelligence website to generate Tagalog sentences and ChatGPT as the primary source. Then, the researchers cleaned the data before proceeding to the pre-processing technique. The initial data cleansing includes removing duplicated sentences and rephrasing non-Tagalog words. The researchers gathered a total of 139, 974 sentences. The sentences undergo validation and correction by professionals in the field of Tagalog language.

Generate a dataset that contains Tagalog sentences for Pokus ng Pandiwa and Panlapi. The researchers need to generate a 100,000 dataset that contains Tagalog sentences for Pokus ng Pandiwa. The dataset includes different kinds of Pokus ng Pandiwa (Pokus sa Tagaganap, Pokus sa Layon, Pokus sa Ganapan, Pokus sa Tagatanggap, Pokus sa Kagamitan, Pokus sa Sanhi at Pokus sa Direksyon).

GENERATED TAGALOG SENTENCES			
POKUS	RAW DATASET	VALIDATED AND CORRECTED SENTENCES	CLEANSED DATASETS
Pokus sa Tagatanggap	24,315	14302	14219
Pokus sa Layon	20,179	14374	13961
Pokus sa Ganapan	15,014	14302	12647
Pokus sa Tagaganap	15,077	14303	13804
Pokus sa Kagamitan	20,513	14508	11528
Pokus sa Sanhi	21,452	14528	14112
Pokus sa Direksyon	23,424	14389	14255
TOTAL	139,974	100706	94526

Table 1. AI-Generated Tagalog Sentences (Dataset)

The researchers are currently validating and correcting the sentences generated by AI with the assistance of professionals in the field of the Filipino language to ensure that the gathered sentences adhere to the grammatical rules of "Pokus ng Pandiwa".

With the raw data of 139, 974, the researcher got only 100,100 unique sentences. The researchers validated the initial dataset used which will be fed into the model for initial results. The data cleaning included removing duplicate sentences, correcting non-Tagalog words and removing incoherent words/sentences. Still, since the researchers are not professionals in the Filipino language, the researchers looked for a professional in the field to correct, validate and double-check the dataset that the researchers made. Ensuring that the validated and corrected dataset will enhance the prediction of the model. The dataset undergoes pre-processing such as:

1. Transform all letters to lowercase
2. Tokenize
3. Padding the sentences

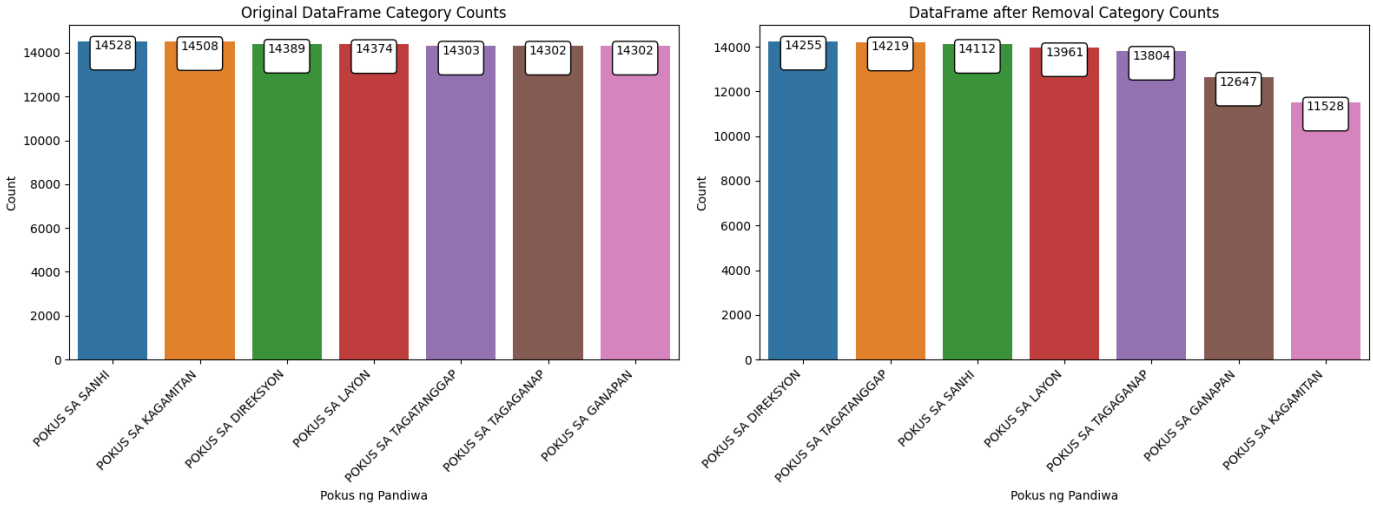


Figure 1. Bar Graph Distribution of Number of Sentences

As a result, the researchers were able to gather the dataset. The bar graph was renamed according to the Pokus ng Pandiwa sentences. Table 1 shows the raw and cleansed sentences and Figure 1 shows the demographics of each Pokus with the respective lengths of sentences generated.

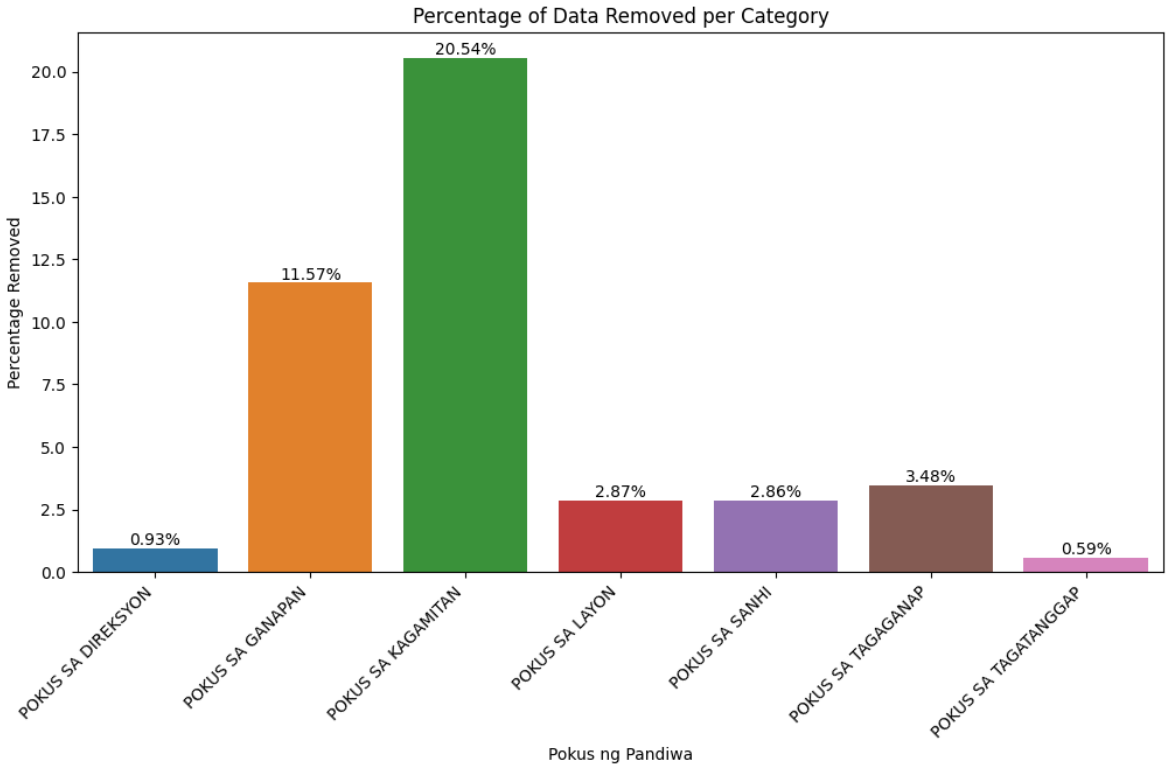


Figure 2. Percentage of Data Removed per Pokus

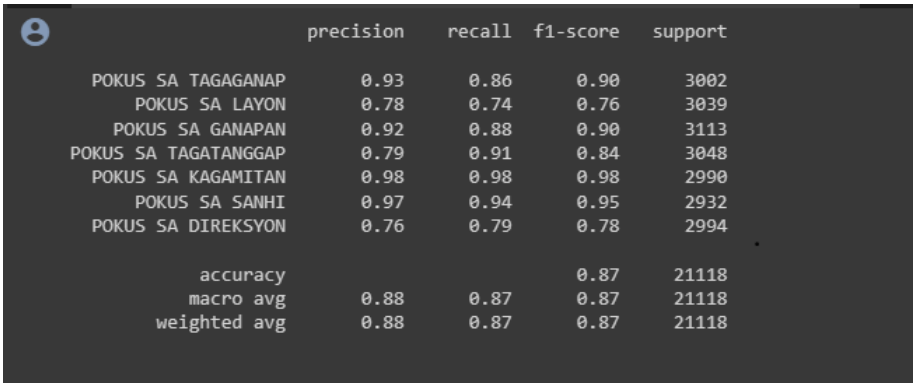
- Model Selection and Training:
    - The researchers will use (DistilBERT) as an NLP Model architecture for grammar checkers.
    - Compared to RoBERTa, DistilBERT has several benefits. Being a condensed form of BERT, it is a lighter, smaller model with fewer parameters that nonetheless maintains most of BERT's functionality. DistilBERT's memory efficiency is improved by this size decrease, which also speeds up training and inference. DistilBERT is more computationally efficient than RoBERTa, as evidenced by its lower computing power and resource requirement. This makes it especially useful in situations when resources are scarce or for smaller-scale applications. DistilBERT's smaller design also enables quicker training times, which facilitates experimentation and model iteration. Moreover, DistilBERT learns from a more comprehensive pre-trained model such as BERT through knowledge distillation, a component of its training technique. This transfer of information enhances its ability to capture crucial linguistic elements, making up for its smaller architecture.

HISTORICAL TRAINING DATA

- The researcher split the training, validation and test into 60%, 20% and 20% respectively.
- Train the selected model using the error-containing dataset—Configure model parameters, including batch size, learning rate, and the number of training epochs. The researchers try different parameters and environment settings to check if the initial results fed in the model are the most appropriate parameters to be used.

II. Model Evaluation

- Performance Metrics:
    - Use relevant performance metrics to evaluate the grammar checker's correction capabilities. Options include:
        - Precision: Measures the proportion of correct grammar corrections.
        - Recall: Assesses the ability of the model to identify and correct errors.
        - F1-score: Balances precision and recall, providing a comprehensive measure of correction quality.



	precision	recall	f1-score	support
POKUS SA TAGAGANAP	0.93	0.86	0.90	3002
POKUS SA LAYON	0.78	0.74	0.76	3039
POKUS SA GANAPAN	0.92	0.88	0.90	3113
POKUS SA TAGATANGGAP	0.79	0.91	0.84	3048
POKUS SA KAGAMITAN	0.98	0.98	0.98	2990
POKUS SA SANHI	0.97	0.94	0.95	2932
POKUS SA DIREKSYON	0.76	0.79	0.78	2994
accuracy			0.87	21118
macro avg	0.88	0.87	0.87	21118
weighted avg	0.88	0.87	0.87	21118

This image represents the initial result of performance metrics in every Pokus ng Pandiwa, based on the given result the initial accuracy is 0.87 or 87%. As per the initial result, the difficulties of the model are on the “Pokus sa Layon”, “Pokus sa Tagatanggap”, and “Pokus sa Direksyon” they gathered the least precision, recall and f1-score.

III. Interpretation of Results

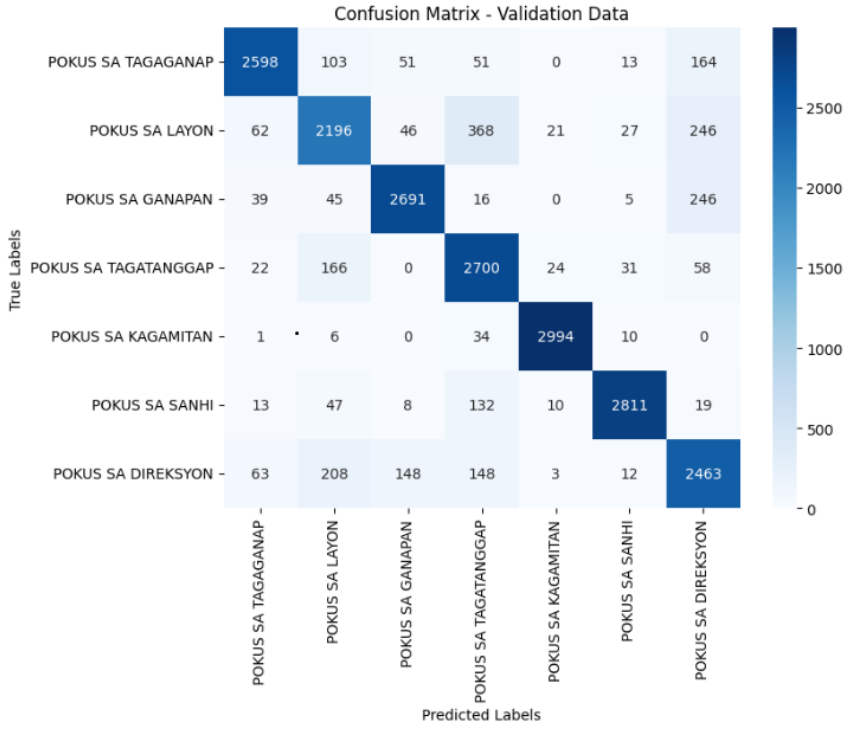
- Grammar Checker Analysis:
  - Analyze the model's ability to check grammar errors. Assess the performance based on the chosen metrics and discuss the overall correction quality.

```
new_sentence = "umuulan ng malakas dahil sa bagyo"
encoded_sentence = tokenizer(new_sentence, return_tensors='pt', padding=True, truncat
input_ids = encoded_sentence['input_ids'].to(device)  # Send the input to the device

with torch.no_grad():
    outputs = model(input_ids)
    predicted_label_id = torch.argmax(outputs.logits, dim=1).item()

# Convert the predicted label ID back to the label string
predicted_label = [label for label, index in label_to_index.items() if index == predi

print(f"Predicted Label: {predicted_label}")
```

```
Predicted Label: POKUS SA SANHI
```

The image is set as a sample for the correct checking made by the model. The model tells what Pokus ng Pandiwa will be the output if the user puts any sentences. The researcher seeks a professional in the field to validate if the output is correct.

- Identification of Error Types:
  - Identify common types of errors that the model struggles with. Categorize errors by types of Pokus ng Pandiwa (e.g., Pokus ng Tagaganap, Pokus ng Layon) and discuss which types are particularly challenging for the model.



The image represents the model performance in grammar checking. The darker shade of blue (Pokus sa Kagamitan) performs better in checking while the light blue (Pokus sa Layon) performs less in checking. The Pokus ng Pandiwa with difficulty in checking in the model is Pokus sa Layon.

IV. Statistical Testing

- Hypothesis Testing:
    - Conduct hypothesis testing to compare the model's performance on different grammar errors.
    - The researchers will opt to use a One-Way Analysis of Variance (ANOVA) to determine significant differences between the developed models. However, since the study is a work in progress, no dummy tables were included as of the time the submission was made.
    - There are significant differences between at least two groups, which is why the researchers used post-hoc analysis after ANOVA.

Anova: Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
POKUS SA TAGAGANAP	13804	115180	8.343958273	3.840388765
POKUS SA LAYON	13804	112317	8.136554622	3.597014746
POKUS SA GANAPAN	12647	87737	6.937376453	2.163245465
POKUS SA TAGATANGGAP	13804	112422	8.144161113	2.169609461
POKUS SA KAGAMITAN	11528	121921	10.57607564	6.607901593
POKUS SA SANHI	13804	170869	12.3782237	10.05522669
POKUS SA DIREKSYON	13804	106696	7.72935381	2.57587958

The ANOVA analysis yielded a significant result with an F-statistic of 10865.37061, indicating that the means of the groups are more different from each other than expected by chance. The associated p-value of 0 provides strong evidence against the null hypothesis, suggesting that there are likely significant differences among the group means. In practical terms, this implies that there are meaningful distinctions between the groups being compared in the study or experiment.

Post-hoc test

t-Test: Two-Sample Assuming Equal Variances

	POKUS SA TAGAGANAP	*POKUS SA LAYON*
Mean	8.343958273	8.178783755
Variance	3.840388765	3.747117174
Observations	13804	13961
Pooled Variance	3.793489244	
Hypothesized Mean Difference	0	
df	27763	
t Stat	7.065379928	
P(T<=t) one-tail	8.19768E-13	
t Critical one-tail	1.644908514	
P(T<=t) two-tail	1.63954E-12	
t Critical two-tail	1.960049435	

V. Conclusion

- Summary of Findings:
    - Summarize the findings from the evaluation, including model performance and its correction capabilities across various error types.
    - Effectiveness Discussion:
        - Discuss the overall effectiveness of the grammar checker, highlighting its strengths and areas for improvement. Consider:
            - The model's performance on different types of errors.
            - It's the potential for real-world applications.
- Recommendations and Future Work:
    - Offer recommendations for enhancing the model's performance, such as increasing the diversity of the training data or fine-tuning the model for specific error types. Suggest areas for future research in the domain of grammar checking.
    - Future recommendations for this application will focus on other parts of speech that may be significant in checking the grammar.