Lost in Translation: Translating Generation Alpha Internet Slang using Machine Learning

A Special Problem Proposal
Presented to
the Faculty of the Division of Physical Sciences and Mathematics
College of Arts and Sciences
University of the Philippines Visayas
Miag-ao, Iloilo

In Partial Fulfillment
of the Requirements for the Degree of
Bachelor of Science in Computer Science by

FLAUTA, Neil Bryan
GIMENO, Ashley Joy
GIMENO, Carl Jorenz

Francis DIMZON
Adviser

November 27, 2024

Contents

1 Introduction **1**

1.1 Overview 1

1.2 Problem Statement 2

1.3 Research Objectives 2

1.3.1 General Objectives 2

1.4 Specific Objectives 3

1.5 Scope and Limitations of the Research 3

1.6 Significance of the Research 3

2 Review of Related Literature **4**

2.1 Communication Gap between Generations 4

2.2 Existing Studies 4

2.3 LoRA for Fine Tuning 5

2.4 Chapter Summary 5

3 Research Methodology **7**

3.1 Research Activities 7

3.1.1 Creation of the dataset 7

3.1.2 Identification of potential LLM to be used 7
3.1.3 Lookup on available GPU on demand services 8
3.1.4 Study on LoRA implementation for LLM 8
3.1.5 Preprocessing of data 8
3.1.6 Prototype implementation of LoRA 8
3.1.7 Implementation of LoRA on selected model 9
3.1.8 Implementation on LLM Evaluation Metrics 9
3.1.9 Model Evaluation and Analysis of Results 9
3.1.10 Documentation 9
3.2 Calendar of Activities 9

List of Tables

3.1 Timetable of Activities 10

Chapter 1

Introduction

1.1 Overview

Language is how humans communicate and express themselves (?, ?). It is dynamic because there are endless structural possibilities, changes in word meanings, and new words created (?, ?). Slang is a great example of the dynamic nature of language. Slang is an informal language used by people in the same social group (?, ?). It serves social purposes: to identify a group's members, communicate informally, and oppose established authority (?, ?). Slang is highly contextual and pervasive, even in non-standard English. Its figurative nature and how it twists the definitions of the words used in it make it hard for outsiders to understand.

In recent years, the internet has become a significant medium for the evolution and spread of language, giving rise to 'internet slang' (?, ?). Internet slang is a collection of everyday language forms used by diverse groups online (?, ?). Ujang et al. (2018, as cited in (?, ?)) state that Internet slang is not easily understood by people outside the social group or people who are not fluent in the language where slang is used. This phenomenon is particularly prominent among the younger generation (?, ?), where they use it to communicate and interact with friends.

Today, Generation Alpha is the youngest generation. Generation Alpha refers to people born between 2010 and 2025. They were born into an era of rapid technological advancement, where digital devices and the internet are integral to their daily lives (?, ?). Generation Alpha is also called the first true digital natives (?, ?). They are expected to be the most "technologically" skilled and most educated generation as they are the native speakers of the language of the Internet (?, ?). According to the study *Understanding Generation Alpha*, Generation Alpha

is socially driven, which may let them grow up to be creative and unconventional, potentially shaping them to be assets in the future (?, ?).

Since Generation Alpha was born with technology, the usage of Internet slang has been prominent in this generation. However, it can create communication barriers between older and younger generations (Venter, 2017 as cited in (?, ?)). The communication barriers caused by the usage of Internet slang also affect people from the younger generation, especially individuals who are less active on social media and have less exposure to them (?, ?). This gap highlights the need for a tool that can bridge the generational divide, making it easier for individuals to understand the language of Generation Alpha. By fostering a mutual understanding, such a tool can promote more effective and harmonious interactions across generations, enhancing relationships and reducing miscommunication.

1.2 Problem Statement

Internet slang fosters informal, relatable communication within the younger generation (?, ?), especially Generation Alpha, but it presents challenges in understanding for people outside this demographic. The gap in comprehension with older generations widens as internet slang evolves, often leading to miscommunication affecting social relationships that contribute to the generational divide (?, ?). A more specific translation tool developed using language models use in many digital platforms can be used to bridge this divide.

1.3 Research Objectives

1.3.1 General Objectives

This study aims to modify an existing Large Language Model (LLM) for use in the translation of Generation Alpha internet slang used by Filipino children in social media.

1.4 Specific Objectives

- To create a dataset of sentences containing gen alpha slang and its formal translation
- To create a Low Rank Adaptation (LoRA) implementation for fine-tuning an existing model
- To fine-tune an existing LLM to translate sentences containing gen alpha slang into formal sentences
- To evaluate the performance of the trained model and compare it to the based model using several performance metrics

1.5 Scope and Limitations of the Research

This study will focus on the usage of internet slang by Filipino Generation Alpha, with an emphasis on English language since it is widely use on different digital platforms such as social media.

1.6 Significance of the Research

The study contributes to understanding the evolving linguistic landscape shaped by internet slang, especially as used by Generation Alpha. Insights gained from this study may aid educators, parents, and communication professionals in bridging intergenerational communication gaps and fostering better understanding across age groups.

Chapter 2

Review of Related Literature

2.1 Communication Gap between Generations

Internet slang is a result of language variation and is often regarded as informal (?, ?). In the study, *The Use of Online Slang for Independent Learning in English Vocabulary* (?, ?), students used internet slang to express their feelings and emotions and because their friends also use it, However, it suggests that younger generation should use slang to communicate with each instead of older generations because it might cause confusion between them (?, ?).

This miscommunication is prominent between generations. Suslak (?, ?) argues that age influences language use, noting that language evolves across generations. Supporting this, a study by Teng and Joo (?, ?) found that the older a person is, the less likely they are to understand internet language.

2.2 Existing Studies

Khazeni et al. used deep learning to create a model for translating Persian slang text into formal ones (?, ?). They were able to create a model to convert texts from social media into sentiments for classification. Nocon et al. (?, ?) created a Filipino colloquialism translator using Tensorflow's sequence-to-sequence model and Moses' phrase-based statistical machine translation. They found that the Moses model was able to create a natural sounding translation, while the Tensorflow model often produced bad sentences.

A slang translation system developed by Ibrahim and Mustafa (?, ?) used models obtained from Hugging Face, a repository of pre-trained models, and re-trained it using a dataset containing slang and their corresponding definition and example. They determined that these models can be tweaked into learning the relationship between the slang and its meaning.

2.3 LoRA for Fine Tuning

Low Rank Adaptation (LoRA) is an efficient Parameter Efficient Fine Tuning (PEFT) method proposed by Hu et al (?, ?). It can significantly decrease the required storage for training while producing comparable results and in some cases, even outperforming other adaptation methods. In addition, it has minimal chance of catastrophic forgetting as the original weights are not being tampered with, unlike other finetuning methods. These factors make it a suitable option for slang translation as a quick yet accurate solution. In a study conducted by Zhao et al. (?, ?), they determined that some LLMs using Low Rank Adaptation (LoRA) for fine tuning can outperform GPT-4, one of the most advanced LLM models currently. A study by Nguyen et al. (?, ?) used LoRA in fine tuning a pre-trained Llama 2 7B model for text classification of a dataset that contains slang. They were able to create a more accurate model compared to models by existing studies at that time.

2.4 Chapter Summary

This chapter shows how generational differences create communication gaps, especially due to internet slang. Younger people tend to use slang to express emotions and connect with friends, but this can confuse older generations who aren't as familiar with these terms. Research shows that as language changes over time, older people are generally less likely to understand the newest internet language. To bridge this gap, some recent studies have utilized machine learning to translate slang into more standard language. For instance, Khazeni et al. (?, ?) used deep learning to translate Persian slang, while Nocon et al. (?, ?) created a Filipino slang translator using statistical models. Moreover, Ibrahim and Mustafa (?, ?) fine-tuned pre-trained models to learn slang meanings. One of the promising techniques for this is Low Rank Adaptation (LoRA), which is a fine-tuning method that keeps the original model stable while using less storage. Studies by Zhao et al. (?, ?) and Nguyen et al. (?, ?) show that LoRA models are not only efficient but can even outperform advanced models like GPT-4 when it comes to slang

translation and text classification.

Chapter 3

Research Methodology

This chapter lists and discusses the specific steps and activities that will be performed to accomplish the project. The discussion covers the activities from pre-proposal to Final SP Writing.

3.1 Research Activities

3.1.1 Creation of the dataset

A dataset of sentences containing Generation Alpha slangs and its formal translation or an approximation of will be created. This will involve data scraping, use of existing datasets, or any other suitable methods of obtaining data. This dataset will be used for the training and evaluation of the model. To ensure it is a high quality dataset, it will be manually checked for accuracy and grammatically correctness. It will also be checked for any potential biases that may exist in the dataset or the data collection process..

3.1.2 Identification of potential LLM to be used

We will be reading upon existing LLM comparison studies to identify potential LLMs to be used for this study. We will be primarily using studies that used dataset containing slangs as they are the most similar to our required dataset.

3.1.3 Lookup on available GPU on demand services

Available computing power rental services will be looked up for this study. As LLM training are a resource-intensive process, it is important to ensure that the necessary computing power is available. However, this computing power requires expensive equipment that might not see usage after the project is completed. Thus, it has been decided that it is better to rent the computing power for the duration of the project. A report on available GPU on demand services will be created using market research and price to computing power ratio.

3.1.4 Study on LoRA implementation for LLM

A thorough study on the implementation of LoRA for fine-tuning will be done. This includes learning the necessary steps, logic behind the idea, and other necessary information necessary for implementation. For this step, reading upon guide materials regarding fine-tuning and LoRA as well as existing studies will be done. We will be primarily using the guide provided by HuggingFace as it is one of the largest repositories for prebuilt LLMs. In addition, they also provided guides for fine-tuning models for specific purposes and has model specific guides.

3.1.5 Preprocessing of data

The dataset used for the fine-tuning of the model will be cleaned up. This will require removal of non essential information such as email adresses, URLs, etc. This is to ensure that the model can focus on learning the patterns between the slang and its formal translation without being affected by noise.

3.1.6 Prototype implementation of LoRA

A prototype implementation of LoRA will be created using a less demanding model. This is to avoid incurring costs from constantly retraining the model due to bugs in the code. It will be also developed on the same platform as the final implementation to avoid any issues with the code running on different platforms. As it is a prototype, it will be used to create a foundation for the complete implementation of LoRA. It will ensure that during the final implementation, there will be no issues with the code and the model can be fairly evaluated.

3.1.7 Implementation of LoRA on selected model

A full implementation of LoRA will be done using the previously created prototype as a basis. Since it has been proven to work, this step will mostly involve fine-tuning the selected model and fixing any hidden bugs.

3.1.8 Implementation on LLM Evaluation Metrics

A set of evaluation metrics will be used to determine if the fine-tuned model will perform better than the base model. These metrics will be taken from existing studies on LoRA finetuning and slang translation. It will serve as the primary measure in which LLMs are compared with from each other.

3.1.9 Model Evaluation and Analysis of Results

The model obtained from previous steps will be evaluated using the evaluation metrics determined from the previous step. To do this, the testing set split of the dataset will be used as the basis of evaluation. In addition, descriptive information such as loss function per epoch, accuracy, precision, recall, and F1 score will be determined. This information will be used as supplement to evaluation metrics to determine if the fine-tuned model will perform better than the base model.

3.1.10 Documentation

All members are tasked to provide accurate and detailed logs of their activities. This includes steps on the task they are working on, the status of the work being done, and the time spent on the task. It will serve both as documentation and as a progress tracker to determine how far the project is from being done. It will be done every week at the member's leisure.

3.2 Calendar of Activities

Table 3.1 shows a Gantt chart of the activities. Each bullet represents approximately one week worth of activity.

Table 3.1: Timetable of Activities

Activities (2024-2025)	Nov	Dec	Jan	Feb	Mar	Apr	May
Creation of the dataset	•						
Identification of potential LLM to be used	•						
Lookup on available GPU on demand services	•						
Study on LoRA implementation for LLM	•						
Preprocessing of data	•••						
Prototype implementation of LoRA	•	••••					
Implementation of LoRA on selected model			••				
Implementation on LLM Evaluation Metrics			••				
Model Evaluation and Analysis of Results				••••			
Documentation	••	••••	••••	••••	••••		