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 15, 2024

Contents

List of Tables

Chapter 1

Introduction

1.1 Overview

Language is how humans communicate and express themselves (Crystal & Robins, 2024). It is dynamic because there are endless structural possibilities, changes in word meanings, and new words created (Libretexts, 2021). Slang is a great example of the dynamic nature of language. Slang is an informal language used by people in the same social group (Fernández-Toro, 2016). It serves social purposes: to identify a group's members, communicate informally, and oppose established authority (McArthur, 2003). 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' (J. Liu, Zhang, & Li, 2023). Internet slang is a collection of everyday language forms used by diverse groups online (Barseghyan, 2014). Ujang et al. (2018, as cited in (binti Sabri, bin Hamdan, Nadarajan, & Shing, 2020)) 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 (Maulidiya, Wijaya, Mauren, Adha, & Pandin, 2021), 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 (McCrindle & Fell, 2020). Generation Alpha is also called the

first true digital natives (Jukić & Škojo, 2021). 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 (Prensky, 2001). 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 (Jha, 2020).

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 (Ghazali & Abdullah, 2021)). 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 (Vacalares, Salas, Babac, Cagalawan, & Calimpong, 2023). 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 age groups, enhancing relationships and reducing miscommunication.

1.2 Problem Statement

Internet slang fosters informal, relatable communication within the younger generation (Ghazali & Abdullah, 2021), 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 (Vacalares et al., 2023). This study investigates the communication barriers internet slang creates, particularly between Generation Alpha and older generations, and explores possible solutions to bridge this gap.

1.3 Research Objectives

1.3.1 General Objectives

This study aims to modify an existing 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 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 (S. Liu, Gui, Zuo, & Dai, 2019). In the study, *The Use of Online Slang for Independent Learning in English Vocabulary* (Ambarsari, Amrullah, & Nawawi, 2020), 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 (Jeresano & Carretero, 2022).

This miscommunication is prominent between generations. Suslak (Suslak, 2009) argues that age influences language use, noting that language evolves across generations. Supporting this, a study by Teng and Joo (Teng & Joo, 2023) 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 (Heydari, Albadvi, & Khazeni, 2024). They were able to create a model to convert texts from social media into sentiments for classification. Nocon et al. (Nocon, Kho, & Arroyo, 2018) 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 (Abdulstar Ibrahim & Shareef Mustafa, 2023) used models obtained from Hugging Face, a repository of pre-trained models, and retrained 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 (Hu et al., 2021). 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. (Zhao et al., 2024), 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. (Nguyen, Wilson, & Dalins, 2023) 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. (Heydari et al., 2024) used deep learning to translate Persian slang, while Nocon et al. (Nocon et al., 2018) created a Filipino slang translator using statistical models. Moreover, Ibrahim and Mustafa (Abdulstar Ibrahim & Shareef Mustafa, 2023) 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. (Zhao et al., 2024) and Nguyen et al. (Nguyen et al., 2023) 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. It will serve as both the training and testing dataset for the fine-tuning of the LLM.

3.1.2 Identification of potential LLM to be used.

A report on potential LLMs to use for this study will be created using existing studies about LoRA finetuning and slang translation. This report will include each LLMs strengths and weaknesses as well as existing studies supporting each evidence.

3.1.3 Lookup on available GPU on demand services

A research on available GPU rental services will be done to obtain the necessary computing power to conduct the LLM finetuning. These services will be compared with each other to obtain the service fitting for this study.

3.1.4 Study on LoRA implementation for LLM

LoRA implementation on LLMs will be studied upon. It will require reading various guides, primarily one created by HuggingFace as they are one of the largest repositories for prebuilt LLMs. They also have several in-depth guides on fine-tuning models for specific purposes..

3.1.5 Preprocessing of data

The dataset will be verified and cleaned before use for the fine-tuning of the model. It is to ensure that all sentences contain at least one slang and their formal translations are grammatically and semantically correct. As LoRA does not tamper with existing knowledge of the model (Hu et al. 2021), we are free to focus on teaching the model the slang while leveraging its original knowledge to provide proper sentences. In addition, after cleaning up the dataset, it will be split into a training and testing set. A dataset for fine-tuning is ready by the end.

3.1.6 Prototype implementation of LoRA

The implementation of LoRA on the selected model will require a prototype implementation to make the full implementation easier and simpler. An option to use qLoRA for the smaller memory requirements in exchange of longer runtime (Raschka, 2023) can be used instead to allow the use of lower end hardware for this study. This prototype will serve as the foundation for the complete implementation of the algorithm and thus, requires it to use the selected computing service to prevent future alterations to adjust to the platform.

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

Evaluation metrics will be implemented to compare the base model with the fine-tuned one. These metrics will be used to determine if the fine-tuned model will perform better than the base model.

3.1.9 Testing and Analysis of Results

The fine-tuned model will be tested using the testing set of the dataset and will use the evaluation metrics to determine its performance. This would include descriptive information regarding the model and comparison with the original model.

3.1.10 Documentation

All members are tasked to provide accurate and detailed logs of their activities. 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			••				
Testing and Analysis of Results				••••			
Documentation	••	••••	••••	••••	••••		

References

Abdulstar Ibrahim, A., & Shareef Mustafa, B. (2023, Oct). Intelligent system to transformer slang words into formal words. *NTU Journal of Engineering and Technology*, *2*(2). doi: 10.56286/ntujet.v2i2.689

Ambarsari, S., Amrullah, A., & Nawawi, N. (2020, Aug). The use of online slang for independent learning in english vocabulary. *Proceedings of the 1st Annual Conference on Education and Social Sciences (ACCESS 2019)*, *465*, 295–297. doi: 10.2991/assehr.k.200827.074

Barseghyan, L. (2014). *On some aspects of internet slang.* Retrieved from https://api.semanticscholar.org/CorpusID:51730779

binti Sabri, N. A., bin Hamdan, S., Nadarajan, N.-T. M., & Shing, S. R. (2020, Jun). The usage of english internet slang among malaysians in social media. *Selangor Humaniora Review*, *4*(1), 16-17.

Crystal, D., & Robins, R. H. (2024, Oct). *Language.* Encyclopædia Britannica, inc. Retrieved from https://www.britannica.com/topic/language

Fernández-Toro, M. (2016, Jun). *Exploring languages and cultures.* Retrieved from https://www.open.edu/openlearn/languages/exploring-languages-and-cultures/content-section-3.2

Ghazali, N. M., & Abdullah, N. N. (2021, Dec). Slang language use in social media among malaysian youths: A sociolinguistic perspective. *International Young Scholars Journal of Languages*, *4*(2), 69. Retrieved from https://www.iium.edu.my/media/77652/Slang%20Language%20Use%20in%20Social%20Media%20Among%20Malaysian%20Youths_A%20Sociolinguistic%20Perspective.pdf

Heydari, M., Albadvi, A., & Khazeni, M. (2024). Persian slang text conversion to formal and deep learning of persian short texts on social media for sentiment classification. *Journal of Electrical and Computer Engineering Innovations (JECEI).* Retrieved from https://jecei.sru.ac.ir/article_2172.html doi: 10.22061/jecei.2024.10745.731

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., ... Chen, W. (2021). *Lora: Low-rank adaptation of large language models.* Retrieved from https://arxiv.org/abs/2106.09685

Jeresano, E., & Carretero, M. (2022, Feb). Digital culture and social media slang of gen z. *United International Journal for Research Technology*, *3*(4), 11–25. doi: http://dx.doi.org/10.1314/RG.2.2.36361.93285

Jha, A. (2020, Jun). *Understanding generation alpha.* doi: 10.31219/osf.io/d2e8g

Jukić, R., & Škojo, T. (2021). The educational needs of the alpha generation. In *2021 44th international convention on information, communication and electronic technology (mipro)* (p. 564-569). doi: 10.23919/MIPRO52101.2021.9597106

Libretexts. (2021, Jul). *3.1.2: Functions of language.* Author. Retrieved from `https://socialsci.libretexts.org/Courses/American_River_College/SPEECH_361%3A_The_Communication_Experience_(Coleman)/03%3A_Verbal_Codes/3.01%3A_Verbal_Communication/3.1.02%3A_Functions_of_Language`

Liu, J., Zhang, X., & Li, H. (2023, Aug). Analysis of language phenomena in internet slang: A case study of internet dirty language. *Open Access Library Journal*, *10*(08), 1–12. doi: 10.4236/oalib.1110484

Liu, S., Gui, D.-Y., Zuo, Y., & Dai, Y. (2019, Jun). Good slang or bad slang? embedding internet slang in persuasive advertising. *Frontiers in Psychology*, *10*. doi: 10.3389/fpsyg.2019.01251

Maulidiya, R., Wijaya, S. E., Mauren, C., Adha, T. P., & Pandin, M. G. R. (2021, Dec). *Language development of slang in the younger generation in the digital era.* OSF Preprints. Retrieved from `osf.io/xs7kd` doi: 10.31219/osf.io/xs7kd

McArthur, T. (2003). *Concise oxford companion to the english language* (1st ed.). Oxford University Press.

McCrindle, M., & Fell, A. (2020). *Understanding generation alpha.* McCrindle Research Pty Ltd.

Nguyen, T. T., Wilson, C., & Dalins, J. (2023). *Fine-tuning llama 2 large language models for detecting online sexual predatory chats and abusive texts.* Retrieved from `https://arxiv.org/abs/2308.14683`

Nocon, N., Kho, N. M., & Arroyo, J. (2018, Oct). Building a filipino colloquialism translator using sequence-to-sequence model. *TENCON 2018 - 2018 IEEE Region 10 Conference*, 2199–2204. doi: 10.1109/tencon.2018.8650118

Prensky, M. (2001, Oct). Digital natives, digital immigrants. *On the Horizon*, *9*(5). doi: https://doi.org/10.1108/10748120110424816

Suslak, D. F. (2009). The sociolinguistic problem of generations. *Language Communication*, *29*(3), 199-209. Retrieved from `https://www.sciencedirect.com/science/article/pii/S0271530909000196` (Reflecting on language and culture fieldwork in the early 21st century) doi: https://doi.org/10.1016/j.langcom.2009.02.003

Teng, C. E., & Joo, T. M. (2023). Is internet language a destroyer to communication? In X.-S. Yang, R. S. Sherratt, N. Dey, & A. Joshi (Eds.), *Proceedings of*

eighth international congress on information and communication technology (pp. 527–536). Singapore: Springer Nature Singapore.

Vacalares, S. T., Salas, A. F. R., Babac, B. J. S., Cagalawan, A. L., & Calimpong, C. D. (2023, Jun). The intelligibility of internet slangs between millennials and gen zers: A comparative study. *International Journal of Science and Research Archive*, *9*(1), 400–409. doi: 10.30574/ijsra.2023.9.1.0456

Vargas, A., & Marbella, F. (2023, Sep). Bokabularyong generation alpha sa pakikipagtalastasang filipino. *International Journal of Research Studies in Education*, *12*(8), 57–69. doi: http://dx.doi.org/10.5861/ijrse.2023.62

Zhao, J., Wang, T., Abid, W., Angus, G., Garg, A., Kinnison, J., ... Rishi, D. (2024). *Lora land: 310 fine-tuned llms that rival gpt-4, a technical report.* Retrieved from `https://arxiv.org/abs/2405.00732`