A Multilingual NLG Model for Legal Assistance Inquiry in the Philippines using Conversational Language Understanding Model

***Abstract*—This paper presents a comprehensive approach to developing and deploying an AI model specialized in Philippine law, leveraging Azure environment tools. The process begins with data collection, where a Python script extracts legal texts from official websites, categorizes them, and prepares them for training. Microsoft's Azure AI Language Studio is then utilized to train the model, dividing the data into training and testing sets. The trained model achieves notable performance metrics, demonstrating clear entity distinction and robustness. Deployment involves authentication and authorization via Azure AD and Key Vault, followed by request processing through bot logic and UX. A logging mechanism captures conversations and feedback for continuous quality assurance and enhancement. This integrated approach ensures the model's efficiency, security, and reliability in addressing legal inquiries, contributing to advancements in AI application within the legal domain.**

***Keywords*—AI model, Philippine law, Azure environment, legal inquiry**

I. INTRODUCTION

Artificial intelligence has reached unprecedented levels of prominence since the widespread adoption following the launch of GPT-4. Its applications span across various domains, including education, professional endeavors, and leisure, making it ubiquitous in modern life. But there are scarce to nonexistent applications of AI in the field of criminal justice specially in the Philippines.

The objective of this project is to construct an innovative Natural Language Processing model meticulously trained to adhere to the national laws of the Philippines and to create a comprehensive dataset encompassing Philippine legislation. This model transcends traditional applications of Natural Language Processing (NLP) in legal contexts by enabling the retrieval of pertinent case laws, statutes, and regulations. It undergoes training across diverse legal contexts, and the curated dataset includes texts from various legal sources such as Constitutions, Treaties, Signed resolutions, General Orders, Presidential Proclamations, Rules of Court, Letters of Instruction, Memorandum Orders, Administrative Acts, Commonwealth Acts, Executive Orders, Court Decisions, Letters of Implementation, Memorandum Circulars, Batas Pambansa, National Administrative Register, Presidential Decrees, and Republic Acts. The collection of laws spans from 1900 to the present day, ensuring a comprehensive repository of legal information throughout Philippine history.

Additionally, this model offers multilingual legal guidance, supporting the most spoken languages in the Philippines, including English, Tagalog, or Taglish. It gathers information regarding the user's legal inquiries to offer preliminary guidance on potential solutions or courses of action. Importantly, it enables users to engage in real-time text-based conversations with the chatbot, facilitating clarification and follow-up questions as needed.

II. MATERIALS AND METHODS

A. Dataset Gathering

A Python script named webscraping.py is employed to systematically retrieve a diverse array of file formats, such as HTML, PDFs, DOCs, and other relevant materials, from a multitude of official law websites across the Philippines. Since a significant portion of the retrieved files are in HTML format, necessitating the conversion of the required PDF files for NLP training, an additional Python script called htmltotext.py is implemented. This script seamlessly converts the HTML files into text format, ensuring compatibility with the subsequent NLP training process.

B. Text Preprocessing

The collected text was systematically categorized based on relevant criteria. To prepare for robust NLP training, an essential step involved thorough data cleansing. This process aimed to eliminate any extraneous elements that could potentially impact the accuracy and efficacy of the training. Specifically, punctuation marks, special characters, and watermarks were meticulously removed from the collected texts to optimize their suitability for training purposes.

C. Text Sorting

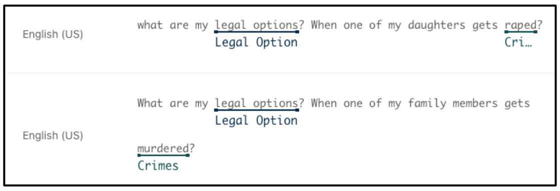
After extraction, the gathered text is organized into distinct classes to streamline the training process. These classes include Legal Option, Crimes, Government Official, Legal Process, Human Rights, Occupation, Health, Constitutional Law, Road Laws, Administrative Law, National Security, and Environmental Laws. This categorization ensures that the data is structured and readily accessible for the subsequent stages of analysis and model development.

Class	**Text**
Crimes	2029
Legal Process	1555
Legal Option	1371
Health	552
National Security	160

Administrative Law	121
Human Rights	256
Constitutional Law	305
Environmental Laws	137
Road Laws	440
Occupation	400
Government Official	440

D. Data Labelling

In addition to compiling laws, a separate file was created specifically focusing on a wide range of legal questions and corresponding answers. This file underwent manual labeling into various categories to ensure accuracy and relevance. This curated collection of legal Q&A serves as a valuable resource for training the AI model to effectively respond to diverse legal inquiries and scenarios.



English (US) — what are my legal options? When one of my daughters gets raped?
Legal Option — Cri…

English (US) — What are my legal options? When one of my family members gets murdered?
Legal Option — Crimes

Figure 1. Sample Labelling of Utterances

E. Model Traning

The model underwent training using Microsoft's Azure AI Language Studio, where the collected text was partitioned into two sets: training data and testing data. Approximately 80% of the text, totaling around 2400 entries, was allocated for training purposes, while the remaining 20%, or roughly 600 entries, were reserved for testing the model's performance.

F. Model Deployment

The deployment of the model involved leveraging various Azure environment tools. The process commences with a user's request, typically a legal inquiry, directed to the bot logic and UX interface. Subsequently, the request undergoes authentication and authorization procedures facilitated by Azure Active Directory (AD) and Key Vault, ensuring secure access to the system.

Once authenticated and provided with token keys, the request proceeds to the bot's cognition and intelligence layer, where it undergoes processing. The resulting output is then delivered back to the user, providing the desired information or assistance.

Additionally, a logging mechanism is implemented to capture and record conversations, user feedback, and relevant logs. These logs serve as valuable insights for further quality assurance and enhancements, facilitated by the DevOps team, ensuring continual improvement of the system's performance and user experience.

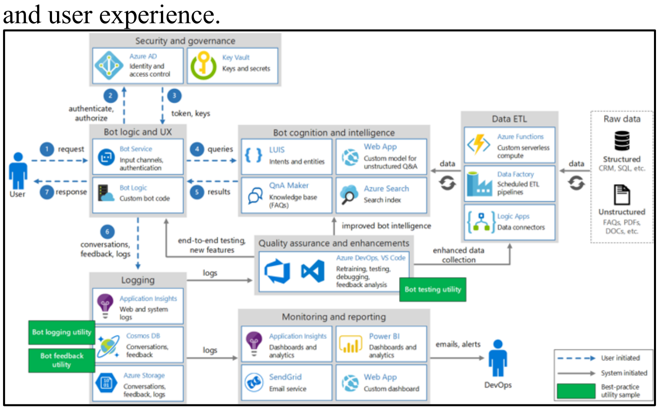
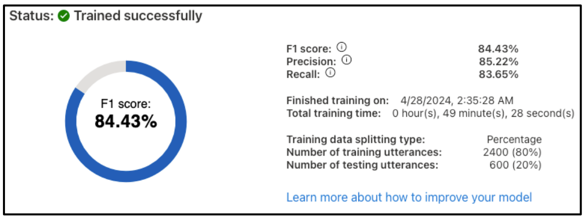


Figure 2. Deployment Maintenance

III. RESULTS AND DISCUSSION

Upon completion, the final model required 49 minutes for training and 28 minutes for testing. It achieved notable performance metrics, with an F1 score of 84.43%, Precision of 85.22%, and Recall of 83.65%.

These results indicate that the model demonstrates a clear distinction between entities within the training set, possesses sufficient data for effective training, and successfully recognizes all entities present in the test set.



Status: Trained successfully

F1 score: 84.43%

F1 score: 84.43%
Precision: 85.22%
Recall: 83.65%

Finished training on: 4/28/2024, 2:35:28 AM
Total training time: 0 hour(s), 49 minute(s), 28 second(s)

Training data splitting type: Percentage
Number of training utterances: 2400 (80%)
Number of testing utterances: 600 (20%)

Learn more about how to improve your model

Figure 3. Result of *Training*

IV. CONCLUSIONS

In conclusion, the development and deployment of an advanced AI model tailored to Philippine law entail meticulous steps and the utilization of sophisticated technologies. Leveraging Azure environment tools, the model is equipped to handle legal inquiries efficiently and securely. From data collection and training to deployment and logging mechanisms, each stage is carefully orchestrated to ensure accuracy, reliability, and user satisfaction. With a robust infrastructure in place, supported by continuous monitoring and improvement efforts, the deployed model stands ready to provide valuable legal insights and assistance while adhering to the highest standards of quality and security