Indonesian E-commerce Chatbot Project Using Merak and Translated Customer Services Datasets

Introduction

The rapid rise of Indonesia e-commerce has been an everyday news for us, Indonesian people. From buying and selling stuff from e-commerce platforms into many assorted services such as consumer logistics and finance, we can see how big the impact of the e-commerce adoptions especially on the consumer side through our everyday life.

We can see from the infographics by AJ marketing that these e-commerce activities have been on the rise with the exception of online price comparison services which have decreased y-o-y. The rise has been felt by people who have buy or sell a product or service through e-commerce platforms. This increasing trend also contributed a lot in easing the barrier of people who wanted open a small to medium businesses like shops and agency or even a company; however, it can be daunting tasks for people to open up their own business or running an existing one even with all of the support that has been brought by e-commerce platforms. The technology helped a lot of things in managing businesses like payment, algorithm, advertising and even integration of all of this. But I have noticed that there is one area where it can be helped with technology a lot, but the growth of the technology is not as advanced as other areas in the business or maybe even people considered it as not important: Customer Support.

Background

Brought by personal experience of opening up a small shop while working full-time, one of the unpleasant experiences is that you have to be standby almost all of the time waiting to check your shop chats for either people wanted to ask about the products or even complaining. If you take too long to answer, your shop score will go down and it will give unpleasant experiences to both the seller for missing potential sales and the customer for being patient. Not only that, during inactive hours where people need or want to communicate while they cannot get response, both customers and sellers will get another bad experience.

This is not only become a problem for small and medium enterprises, this is also made big enterprises to spend a lot of money to avoid giving bad experiences and protecting their brand image. Even big enterprises tried to reduce the costs of customer services by outsourcing them as shown by the reasons in the infographic made by pac-biz.com. in fact, for big enterprises, customer services also another form of public relations that needs to be maintained all of the time due to their operations scale. In conclusion, customer services has been a huge problem for all level of businesses with big enterprises needed substantial amount of funds to solve while for small and medium enterprises, they might have lack of funds or emphasize on the problem. However, there is actually a form of technology that can alleviate this problem: Chatbots.

Chatbots

Chatbots have been around for a long time starting from 1966 with the 1st chatbot is named ELIZA created at MIT by Joseph Weizenbaum.

```
  Welcome to
                 EEEEEE  LL      IIII   ZZZZZZ   AAAAA
                 EE      LL       II       ZZ   AA   AA
                 EEEEE   LL       II     ZZZ    AAAAAAA
                 EE      LL       II    ZZ      AA   AA
                 EEEEEE  LLLLLL  IIII ZZZZZZ    AA   AA

  Eliza is a mock Rogerian psychotherapist.
  The original program was described by Joseph Weizenbaum in 1966.
  This implementation by Norbert Landsteiner 2005.


ELIZA: Is something troubling you ?
YOU:   Men are all alike.
ELIZA: What is the connection, do you suppose ?
YOU:   They're always bugging us about something or other.
ELIZA: Can you think of a specific example ?
YOU:   Well, my boyfriend made me come here.
ELIZA: Is it important to you that your boyfriend made you come here ?
YOU:   He says I'm depressed much of the time.
ELIZA: I am sorry to hear that you are depressed.
YOU:   It's true. I am unhappy.
ELIZA: Can you explain what made you unhappy ?
YOU:
```

ELIZA is categorized as **Rule-based** chatbot where the chatbot is programmed to send scripted replies based on the question asked by users. When the chatbot is being asked questions/prompts outside programmed script, it will come back to a general response that are being scripted like “I’m sorry, I cannot understand that”, etc. or just cannot answer. In the e-commerce platform or apps, businesses often used this and will give customer/user a general reply of telling the customer to wait for a representative to reply manually (this also doesn’t solve the inactive hours problem).

With the dawn of machine learning and AI, now chatbot can also self-learn and be able to answer some questions that are not programmed. AI chatbot have been classified as **Discriminative** model and **Generative** model. Discriminative actually followed the same logic as rule-based; however, Discriminative chatbot can try to answer the questions that are not programmed or trained by trying to determine which area or nearest topic that the unprogrammed question fell into. While Generative chatbot can create their own human-like answer by generating their own response based on the training data and Natural Language Processing (NLP) at the cost of more resources such as data and hardware. By deploying Generative AI chatbot, we can get a lot more advantage such as:

1. Generative AI chatbots use NLP to interpret various sentence structures, slang, and idioms, allowing them to understand a wide range of user inputs. This capability means they can handle queries phrased in numerous ways, providing relevant responses even when the user's language is informal or complex.

2. Generative AI chatbots can track and remember the context of a conversation. This ability allows them to provide responses that are coherent over the course of an interaction, leading to a more natural conversation flow.

3. Generative AI chatbots can analyze past interactions and user data, AI chatbots can tailor their responses to individual users. For instance, they might offer product recommendations based on a user's previous inquiries or purchases, or they might adjust their communication style to match the user's preferences.

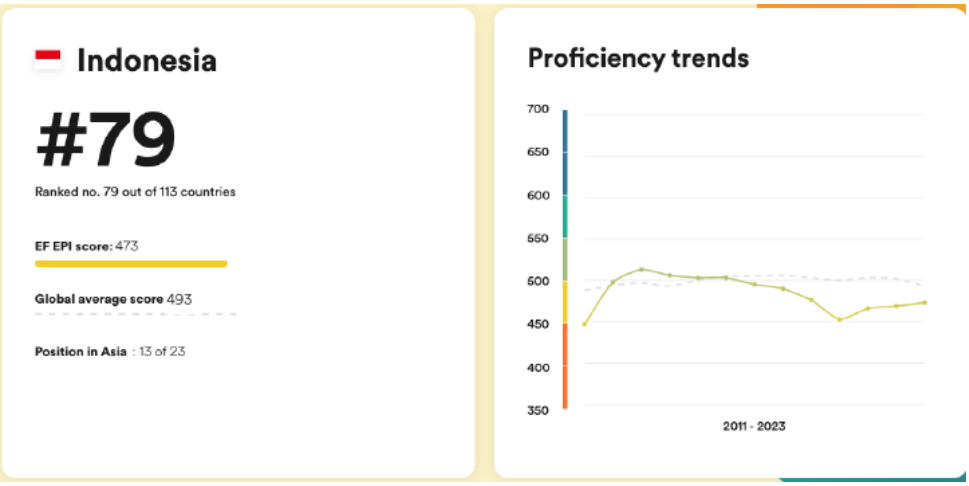
4. As Generative AI chatbots learn and adapt, they become better equipped to handle a larger variety of queries, including complex ones, without the need for constantly expanding the rule set

5. While Generative AI chatbots require a more substantial initial investment in terms of development and training, they offer long-term cost savings due to their adaptability, reducing the need for frequent manual updates.

The advantages of Generative AI chatbots lie in their sophisticated capabilities for natural language processing, adaptability, contextual and personalized interactions, and scalability. These features make them more effective in providing a user-friendly, efficient, and satisfying customer service experience compared to traditional rule-based systems; however, we will have our biggest barrier in implementing these efficiently: Indonesian native language.

Indonesia Native Chatbot

In order to be able provide an effective support toward customer, Indonesian language or Bahasa Indonesia needs to be supported by the chatbot as Indonesia people English literacy is ranked as 79 out of 113 countries and fared below the average of Global English proficiency by the EF international education institute.



As right now, there are not a lot of advancement in Indonesian NLP language compared to the English one and even then we have no data that are needed to train our Indonesian Generative AI Chatbot. Most of the models that can generate Indonesian text are mostly trained on machine translated datasets or sites that can provide Indonesian language such as Wikipedia Indonesia. Hence, we will try to find a proper data that can supported our customer service chatbot which most likely to be translated one.

Data Preprocessing

So the plan for this research is to find a model that is capable in generating text in Indonesian and the dataset that are related with customer service question and answer pairing. We find a model called Merak-7B which can generate Indonesian language. This model is based on Mistral-7B-OpenOrca and fine tuned by some of Indonesia Wikipedia articles and English conversational data translate into Indonesian with Marian NMT and pretrained model from Helsinki-NLP/opus-mt-en-id.

After researching Huggingface datasets list, we found out that there is a English customer service conversational data called Bitext which we translated by using Helsinki-NLP/opus-mt-en-id. This dataset encompasses:

- **27 intents** across **10 categories**, reflecting common customer service scenarios.
- A total of **26,872 question/answer pairs**, providing a rich training ground for the chatbot.
- **30 entity/slot types** and **12 different types of language generation tags**, offering a nuanced understanding of customer interactions.

Along with 5 columns: flags for generation tags, intents, categories, instructions and response. Later, the response column will be the dependent variable while the rest will be combined as independent variables depending on the method of the training.

Here are the summary of translation and preprocessing steps performed:

1. Standardize the entity tag value in both **response** and **instruction** columns (value in double brace format)
2. Replace contractions with their expansion like “I’m” become “I am”
3. Performing spellcheck with Spellcheck() package for both **response** and **instruction**
4. Replace the entity tag with unique identifier so that it would not be translated (later will be reverted back after translated)
5. Translate both the **response** and **instruction** columns
6. Replacing additional mistranslation and typo
7. Replacing all of the translated column entity tags into the original ones