Evaluation of ChatGPT as a classification model to detect prostitution sentences in SNS

1st Keisuke Yoneda
Department of Computer and Information Sciences, Sojo University
4-22-1 Ikeda, Nishi-ku, Kumamoto, 860-0082, Japan
yoneda@cis.sojo-u.ac.jp

2nd Daisaburo Yoshioka
Department of Computer and Information Sciences, Sojo University
4-22-1 Ikeda, Nishi-ku, Kumamoto, 860-0082, Japan
yoshioka@cis.sojo-u.ac.jp

***Abstract*—ChatGPT has attracted great attention because it can generate high-quality responses to human inquiries. However, it is still unclear whether ChatGPT works well on the language understanding task of SNS such as detection of illicit messages in Japanese SNS because such SNS's data mainly contain spoken languages, slangs, secrete languages, and Emoji, which are quite different with the characteristics of the ChatGPT's training data consisting of web articles and books in English. In this report, we evaluate classification tasks for illicit or inappropriate sentences related to prostitution in SNS using ChatGPT. Our resuls show that ChatGPT with only prompts including the well-known zero-shot CoT and few-shot prompting performs poorly, but a fine-tuned ChatGPT achieves an accuracy of over 90%, which is comparable in performance to the fine-tuned BERT model. We demonstrated that the performance of the model can be adjusted by how examples for few-shot prompting are presented.**

***Index Terms*—ChatGPT, text classification, SNS**

I. Introduction

Social network services (SNS) such as Facebook, Instagram, and Twitter which was branded as X are broadly spread all over the world thanks to its short and quick approach to posting. SNS allows the user to communicate each other and to became an active individual. Unfortunately, a door has also been opened for illegal businesses such as prostitution, drag, and illegal part-time jobs. The National Police Agency in Japan said that the number of minors in Japan who became victims of crimes through the use of SNS in 2022 is 1,732 [1]. Violations of the law banning child prostitution and pornography and juvenile protection ordinances that ban sexual acts accounted for some 90 percent of the total cases. Twitter is the most commonly used SNS for such illicit affairs due to its anonymity.

To inhibit such illicit affairs occurring on SNS, National Police Agency in Japan recruits collaborators called *cyber volunteers*. There are 281 organizations of the cyber volunteers all across Japan in 2022 [2]. The one of main activities of cyber volunteers is *cyber patrol*, where volunteers seek illicit or inappropriate sentences related to prostitution and drag in Twitter and report the users posting such sentences to a Prefectural Police. However, there is an issue that such manual works are a kind of burden.

On the other hand, the rise of large language models (LLMs) has a significant impact on the research field of natural language processing (NLP) and artificial intelligence (AI). LLMs have gained significant popularity due to their remarkable performance across a wide range of tasks. Among them, ChatGPT developed by OpenAI has consistently achieved state-of-the-art results, demonstrating its ability to generate coherent and contextually relevant responses in conversational settings. In [4], it is empirically demonstrated that ChatGPT attains a comparable understanding ability of natural language understanding tasks compared with fine-tuned BERT (Bidirectional Encoder Representations from Transformers) models. However, it seems that there are some risks of hallucinations in languages other than English because ChatGPT has been trained on vast quantities of data which mostly consist of web data and books in English. Moreover, there are special characteristics of SNS's data such that short messages and spoken languages are commonly used and there are many slangs or secrete languages and Emoji, which are quite different with the training data for ChatGPT. Therefore, it is still unclear whether ChatGPT works well on the language understanding task in SNS such as detection of illicit sentences in Japanese SNS.

In this paper, we evaluate the classification task of ChatGPT to detect personal prostitution sentences in japanese SNS. First, we design prompts for our desired task, then the performance of ChatGPT is compared with fine-tuned BERT classification model. And we evaluate the performance changes caused by the examples presented in few-shot prompting.

II. Classification task for illicit or inappropriate sentences related to prostitution in SNS

This study aims to detect tweets intended for prostitution by individuals in Japan. In recent years, a person who wants prostitution advertises his/her "sexual services" using Twitter with seemingly innocuous terms such as "papa-katsu" or "enjo kosai" shortened to "enko" which are often translated as "compensated dating" or "subsidized dating". The most common connotation of the term "enko" in Japan is that it is a form of minor prostitution whereby participating girls sell their bodies in exchange for money. The term "papa katsu" has an almost

TABLE I
THE RESULT OF THE FINE-TUNED BERT MODEL

The num. of data	Accuracy	Precision	Recall	F-measure
100	0.883	0.989	0.883	0.939
500	0.906	0.990	0.906	0.943
1000	0.921	0.989	0.921	0.952
2000	0.925	0.993	0.926	0.957

similar meaning to "enko", but it can also include simply spending time together for compensation. Whose slangs or secret words in Japanese are *papa-katsu* and *enko*. In Japanese which means searching a man who wants to meet a girl or woman for sexual reasons. First, we collect tweets by using Twitter API with search words *papa-katsu* and *enko*, which are a kind of slang or secret word meaning personal prostitution. then, the tweets are classified as either "suspicious" or "not suspicious" of being related to personal prostitution or not. The decision is manually made by a majority of three persons. The tweets classified in "not suspicious" include some that are truly unrelated to prostitution, as well as many are posted from dating and matchmaking services. Japanese law does not allow minors to register on dating and matchmaking sites, so tweets from such sites shall not be detected.

A fine-tuned BERT is used for this text classification task. BERT is a powerful natural language processing model that uses a deep neural network architecture based on the transformer model. A pre-trained BERT is trained on massive amounts of text data, such as books, articles, and websites, before it's fine-tuned for specific tasks, including text classification. In this study, we use the pre-trained Japanese BERT model named *bert-base-japanese-char-word-masking* which is trained on the Japanese portion of CC-100 dataset and the Japanese version of Wikipedia [7]. Then, the fine-tuned BERT model is developed using the training data from 100 to 2,000 samples. The number of tweets labeled as positive and negative in the trained data are the same in the training data. The performance of the fine-tuned BERT classifier was evaluated using 1,000 test data based on average accuracy, average precision, average recall, and average F-Measure. No tweets in the test data are duplicated in the training data and the number of tweets labeled as positive and negative in the test data are the same as well as the training data. Table I summarizes average accuracy, precision, recall, and F-measure for each number of training data. It is confirmed from this table that the accuracy of 88% is achieved even with 100 test data, and all values are exceed 90% with more than 500 test data. The precision is also close to 99% even with a small number of training data. As a result, our classification task can be solved by the fine-tuned BERT model.

III. CLASSIFICATION OF INAPPROPRIATE SENTENCES RELATED TO PROSTITUTION IN SNS USING CHATGPT

In LLM, one can use a text or template known as a prompt to strongly guide the generation to output answers for desired tasks. We call such prompts with explicit conditioning on few task examples as Few-shot prompting, and other template-only prompts as Zero-shot prompts. First we create the basic template for the Zero-shot prompt, which is shown in Prompt 1.

Prompt 1

> Analyze the given sentence, then, determine whether the sentence means personal prostitution or not based on the analysis. The output should be "T" or "F" if the decision is positive, or negative, respectively. If you cannot make a decision, output "N".

Hereafter, in all of experiments, Propmpt 1 is always given to ChatGPT.

Furthermore, we also explore the effect of some advanced prompt strategies, i.e., the few-shot prompting known as *in-context learning* [3] and zero-shot CoT prompting [6]. In Few-shot prompting, each three examples of tweets labeled as positive and negative, totally 6-shot examples, are added in the last of Prompt 1. Zero-shot CoT has been shown to be effective in improving performance on a variety of tasks, particularly such as arithmetic and symbolic reasoning. It differs from the original CoT prompting as it does not require step-by-step few-shot examples and simply requires the one sentence *"Let's think step by step"*. Our basic prompt with Zero-shot CoT prompting is also evaluated in this experiments, where only the sentence *"Let's think step by step."* is included in the last sentence in Prompt 1.

We evaluate the performance of ChatGPT with three types of prompts using as the same test data as that used for the fine-tuned BERT evaluation. Note that ChatGPT API GPT3.5 Turbo (gpt-3.5-turbo-0613) is utilized in our experiments. The GPT 3.5 Turbo is a language model which has about 350 billion parameters. Table II summarizes average accuracy, precision, recall, and F-measure for each prompt. The numbers of "N" meaning "hard to decide" are 64, 67, and 44 for Zero-shot, Zero-shot CoT, and Few-shot prompting, respectively. The answers of "N" are excluded in all of the result of Table II. One can see from Table II that the accuracies are about 50 % for all of three kinds of prompts. Since half of the test data is "positive", the result implies that ChatGPT performs poorly on our classification task even with the advanced prompting strategies. Only the recall of Zero-shot CoT achieves about 98% with high probability. This is due to the fact that almost all answers of ChatGPT with Zero-shot CoT are "positive". These results may imply that the well-known prompting methods of the few-shot and zero-shot CoT cannot facilitate the understanding ability of ChatGPT for a task where ChatGPT does not have enough knowledge related to it.

Next, we add an answer of ChatGPT after giving the basic prompt as shown "Answer" to confirm the desired task correctly.

TABLE II
THE RESULT OF CHATGPT WITH THREE TYPES OF PROMPTS

	Accuracy	Precision	Recall	F-measure
Zero-shot	0.4690	0.4907	0.8236	0.6150
Zero-shot CoT	0.5059	0.5028	0.9849	0.6657
Few-shot	0.5052	0.5037	0.7080	0.5886

TABLE III
THE RESULT OF CHATGPT WITH THREE TYPES OF PROMPTS AND ANSWER

	Accuracy	Precision	Recall	F-measure
Zero-shot	0.5688	0.5476	0.6463	0.5929
Zero-shot CoT	0.5246	0.5100	0.9160	0.6550
Few-shot	0.6927	0.6767	0.7423	0.7080

Answer

We analyze the given sentence, then, determine whether the sentence means personal prostitution or not based on the analysis. We output "result:○○". If examples are given, we also judge while referencing them. Please give me sentences.

The results of evaluation under the same conditions as Table II are summarized in Table III. In this experiment, the numbers of "N" meaning "hard to decide" are 117, 26, and 141 for Zero-shot, Zero-shot CoT, and Few-shot prompts, respectively. A tendency to refrain from judgments is observed when zero-shot and few-shot prompts are used. Compared with Table II, accuracy and precision are improved by giving Answer. Some performance improvements of all results are accomplished especially for few-shot prompts. It can be said from this results that giving an answer to reconfirm a desired This result may indicate that few-shot prompting with an answer to reconfirm a desired task could enhance ChatGPT's capabilities.

IV. EVALUATION ON FINE-TURNING AND FEW-SHOT PROMPTING

The results were evaluated on a trained model with fine-tuned ChatGPT. Note that ChatGPT API GPT 4.0o mini (gpt-4o-mini) is utilized in our experiments. The fine-tuned gpt-4o-mini model is developed using the training data from 2,000 samples.

The results are summarized in Table IV. We also evaluated the accuracy of Few-shot prompting, comparing three patterns: 5 examples statements each of true and false, and only 10 True example sentences and 10 False example sentences. The model with Few-shot prompting showed a slight improvement compared to the model with fine-tuning only. There was no difference in the percentage of correct answers depending on how the example sentences were given to Few-shot prompting.

We confirmed that the performance of the models changed depending on how the Few-shot was given. The fit rate was higher and the recall rate was lower when only True example sentences were used than when both True and False example sentences were used. When only the false example sentences were used, the fit rate was lower than when both the true and false example sentences were used, and the reproduction rate was higher. When there is bias in the output of the fine-tuning model, it is expected that the method of providing Few-shot example sentences will be effective in calibrating the bias in the responses.

TABLE IV
THE RESULT OF CHATGPT WITH FINE-TRUNING AND FEW-SHOT PROMPTING

	Accuracy	Precision	Recall	F-measure
FT only	0.938	0.919	0.960	0.938
FT + Few-shot	0.950	0.941	0.96	0.950
FT + Few-shot(True only)	0.943	**0.949**	0.935	0.942
FT + Few-shot(False only)	**0.953**	0.928	**0.980**	**0.953**

V. CONCLUSION

In this study, we evaluated classification task of ChatGPT to detect sentences related to personal prostitution in Japanese SNS. Although only prompts techniques such as Zero-shot, Few-shot, and Zero-shot CoT do not achieve enough good accuracy, the fine-tuned models of ChatGPT with Few-shot examples result in comparable performance to the fine-tuned BERT.

We demonstrated that the precision and recall can be adjusted by the types of example tasks presented in few-shot prompting.

ACKNOWLEDGMENT

We would like to express our gratitude to Mr. Shota Yanachi, who graduated from Sojo University in 2024, for his significant contributions to all aspects of this research.

REFERENCES

[1] The National Police Agency, Community Safety Bureau, Juvenile Division Web site https://www.npa.go.jp/bureau/safetylife/syonen/pdf-r4-syonenhikoujyokyo-kakutei.pdf
[2] The National Police Agency, The number of cyber volunteers in 2022, https://www.npa.go.jp/bureau/cyber/pdf/r4-volunteer.pdf
[3] T. B. Brown, et al., "Language models are few-shot learners," airXiv, 2020.
[4] Q. Zhonn, J. Liu, B. Du, and D. Tao, "Can ChatGPT understand too? A comparative study on ChatGPT and Fine-tuned BERT," airXiv, 2023.
[5] T. Koide, et al., "Detecting Phishing Sites Using ChatGPT," airXiv, 2023.
[6] T. Kojima, et al., "Large Language Models are Zero-Shot Reasoners," airXiv, 2023.
[7] Tohoku University, Japanese BERT model, https://github.com/cl-tohoku/bert-japanese