Emoji Prediction of Japanese X Posts by LLMs

Yijie Hua, Takehito Utsuro

Proceedings of the 38th Pacific Asia Conference on
Language, Information and Computation (PACLIC 38)

Nathaniel Oco, Shirley N. Dita, Ariane Macalinga Borlongan, Jong-Bok Kim (eds.)

2024

Emoji Prediction of Japanese X Posts by LLMs

Yijie Hua and **Takehito Utsuro**
Deg. Prog. Sys.&Inf. Eng., Grad. Sch. Sci.&Tech., University of Tsukuba
1-1-1 Tennoudai, Tsukuba, Ibaraki, Japan
{ s2420830, utsuro.takehito.ge }_@_u.tsukuba.ac.jp

Abstract

Social media users often enhance and complement the emotions underlying in their posts by using emojis. This leads to an increase in research on sentiment analysis using text accompanied by emojis. While there are plenty of previous works for predicting emojis in English posts based on understanding the meaning of posts and classifying them into appropriate emoji categories, research on emoji prediction in Japanese is scarce. Additionally, all of those previous works utilize classification models like BERT instead of large language models. Therefore, in this paper, we utilize three large language models, ChatGPT[1], Claude[2] and Gemini[3] to predict emojis for Japanese X posts, and compare the results to pre-trained models such as XLM (Conneau and Lample, 2019), Japanese BERT[4] and Japanese RoBERTa[5]. The results show that Claude with 8 shots provided performs the best.

1 Introduction

Social media users often enhance and complement the emotions underlying in their posts by using emojis. Emojis have become an indispensable element in NLP. Many studies have attempted to understand the meaning of text using emojis. For example, the performance of irony detection can be improved by utilizing an emoji prediction model as a transfer learning approach (Golazizian et al., 2020).

There also exists the task of emoji prediction, where the most suitable emojis are predicted from text-only posts, or those text-only posts are classified into appropriate emoji categories. The emoji prediction task is crucial for understanding and analyzing the meaning of posts on social media (Barbieri et al., 2017, 2018a,b; Cappallo et al., 2015; Felbo et al., 2017; Lee et al., 2022; Ma et al., 2020; Singh et al., 2022; Tomihira et al., 2018, 2020). Particularly in classification tasks using large language models (LLMs) like ChatGPT, it is necessary to not only understand the meaning of the text but also comprehend the meanings and usages of emojis, and correlate them with the text.

However, studies on emoji prediction have focused mainly on English and the models studied so far are classification models (Barbieri et al., 2017, 2018a,b; Cappallo et al., 2015; Lee et al., 2022; Ma et al., 2020; Singh et al., 2022; Tomihira et al., 2018), where research on emoji prediction in Japanese using LLMs is scarce. Furthermore, most of these studies have not considered the validity of emojis annotated to the text by the users and have not studied whether the emoji annotated to each post is predictable or not by humans. Also, emojis with similar usages and meanings exist, so it is necessary to categorize emojis into appropriate emoji groups before prediction. Plus, not every post on social media is emoji-predictable since usages of emojis are determined by an individual human. Therefore, in this paper, we first propose to group emoji labels considering the emotion each emoji label represents, where similar emojis are grouped together so that not each individual emoji but each emoji group should be predictable. We also develop datasets consisting of emoji-predictable Japanese X posts to evaluate emoji prediction models such as large language models (ChatGPT, Claude and Gemini) and compare the results with pre-trained models such as XLM, Japanese BERT and Japanese RoBERTa.

[1] https://platform.openai.com/docs/models
[2] https://docs.anthropic.com/en/docs/about-claude/models
[3] https://ai.google.dev/gemini-api/docs/models/gemini
[4] https://huggingface.co/tohoku-nlp/bert-base-japanese-v3
[5] https://huggingface.co/rinna/japanese-roberta-base

step	# posts	ratio (%)	step	# posts	ratio (%)	step	# posts	ratio (%)
1	5,568,951	100.00	3	1,234,431	22.17	5	592,546	10.64
2	5,568,951	100.00	4	1,192,752	21.42	6	315,746	5.67

Table 1: Numbers of X posts after each preprocessing step

Our contributions are as follows:[6]

1. We demonstrate that Claude with few shots provided is useful in emoji prediction task. This is also the first study to conduct emoji prediction task using LLMs in Japaneese.

2. We create datasets that consist only of posts that are emoji-predictable by humans and ones that consist of both posts that are emoji-predictable by humans and posts that are not. We show that better performance can be achieved when predicting the former rather than the latter.

2 Related Work

The emoji prediction task was introduced in 2015 or earlier (Cappallo et al., 2015) but started to receive attention from an NLP standpoint in 2017 (Barbieri et al., 2017), where it can be seen that the technologies studied in the task have contributed to various NLP tasks. State-of-the-art performance has been achieved in sentiment analysis, emotion recognition, and sarcasm detection benchmarks using an emoji prediction model DeepMoji (Felbo et al., 2017). The improved version of DeepMoji, a label-wise attention LSTM, is then utilized to predict emojis using data from SemEval 2018 Task 2 (Barbieri et al., 2018a,b). Although the label-wise attention LSTM fails to achieve state-of-the-art performance, it shows the strength of relationships between emojis and individual words, contributing to analyzing how the model predicts the emojis. Their method also improved performance of infrequently used emojis. Furthermore, a machine learning technique is utilized to predict emojis in multi-class and multi-label settings (Ma et al., 2020). Emoji prediction is also conducted as a multi-task learning task with emotion classification as an auxiliary task (Lee et al., 2022), or it is conducted with both sentiment analysis and emotion analysis (Singh et al., 2022). Here, it is important to note that those previous studies on emoji prediction have focused mainly on English.

In Tomihira et al. (2018, 2020), performance of emoji prediction in Japanese has been examined, and the comparison between emoji prediction for Japanese and English posts has been investigated using CNN, FastText, attention BiLSTM and BERT[7]. With the English dataset, it achieves higher accuracy than with the Japanese dataset in all models, whereas F_1 scores are lower in some models. In this previous work (Tomihira et al., 2018, 2020), the models employed are those other than LLMs such as CNN, FastText, BiLSTM and BERT and they have no discussion on the influence of emoji-predictability of posts. On the other hand, we utilize LLMs to predict emojis and discuss the influence of emoji-predictability of posts on the performance of the emoji prediction task. Specifically, we developed a separate dataset consisting only of posts whose emojis are predictable by humans, and evaluated proposed methods with the dataset.

3 X Posts Dataset

Emojis have multiple versions, and the OpenAI's model `gpt-3.5-turbo-0125` is trained with the oldest data among the three LLM models we use and can only accurately recognize emojis up to version 13.1. Therefore, in this paper, we adhere to `gpt-3.5-turbo-0125` and set emojis of version 13.1 as the recognition limit. In our preliminary evaluation, about 97% of emojis are up to version 13.1. We collect Japanese X posts and preprocess them as following, where the number of posts extracted at each step is shown in Table 1:

1. We randomly collect Japanese X posts that are available at time of January 2023 without any restriction on those post dates.

2. We remove URLs and user mentions since they can be noise for prediction.

3. In this paper, we consider the emoji prediction task as a multi-class classification task.

[6]The code and the data used in this study are available at the following URL: `https://anonymous.4open.science/r/emoji-prediction-3275`.

[7]`https://github.com/tommy19970714/EmojiPrediction`

emoji	# posts	emoji	# posts	emoji	# posts	emoji	# posts
😊	38,210	😂	19,120	💦	12,896	😃	8,210
✨	32,244	🙄	18,641	🥰	12,234	💕	7,841
😭	23,972	🤣	17,552	🤔	11,631	🥲	7,760
☺️	23,423	😅	13,461	👍	9,525	🎉	7,597
☀️	21,869	😆	13,180	😇	9,336	😄	7,044

Table 2: The top 20 most frequent emojis and their distribution in the M_{20} dataset

group ID	emoji group	representative emoji	group name	# posts	ratio (%)
1	😊☺️😃😆😄👍	😊	joy	99,592	31.54
2	🤣😂	😂	fun	36,672	11.61
3	😭🥲🙄	😭	sadness	50,373	15.95
4	🎉✨☀️	🎉	celebration	61,710	19.54
5	🥰💕	🥰	love	20,075	6.36
6	💦😅	💦	sweat	26,357	8.35
7	😇	😇	angel	9,336	2.96
8	🤔	🤔	question	11,631	3.68
total	—	—	—	315,746	100.00

Table 3: Representative emojis and group names after grouping emojis

Therefore, we extract posts each containing only one emoji.

4. We extract posts containing emojis up to version 13.1 as the recognition limit of the ChatGPT model mentioned earlier.

5. In this paper, we aim to understand the meaning of the entire text and predict emojis accordingly. Therefore, we extract posts where emojis are located at the end of the posts.

6. We extract posts with the top 20 most frequently occurring emojis to limit the variation of class labels.

It should be noted that in step 5, the number of the posts amounts to nearly half of the posts after the step 4. It is known that emoji prediction models behave differently according to the position of the emoji within a post (Kwon et al., 2022). The issue on how to handle the remaining half by considering the position of the emoji within a post is left as a future work. Thus, finally, the dataset obtained by the overall preprocessing above (i.e, 315,746 posts each with an emoji in it) is referred to as M_{20} for use in the following experiments.

4 Emojis for Evaluation

4.1 Selecting Emojis for Evaluation

Table 2 shows the top 20 most frequent emojis and their distribution in the M_{20} dataset. One problem here is that we cannot always predict emojis based on the text of each post because the emoji corresponding to the text is sometimes not uniquely determined, since emojis like 😊 and ☺️ have similar meanings and usages. Neither humans nor models of this paper can tell whether a text is more appropriate to 😊 or ☺️. On the other hand, in traditional multi-class classification tasks such as sentiment analysis, the class label corresponding to the text is uniquely determined, which typically means that a text can be usually classified into a single class (like either positive or negative). In order to address this issue, we group emojis so as to ensure that the correspondence between a text and an emoji is uniquely determined even if in the case where there exist similar emojis.

Thus, emojis in Table 2 are grouped accordingly, and each emoji is replaced with its representative as shown in Tabel 3. There are many approaches finding similar emojis and grouping them. One common method is utilizing vectors, meaning that emojis are placed into a vector space and emojis' similarity is determined based on vector similarity. In fact, emoji embeddings, which is

Japanese X post (and its English translation)	emoji	group name
ササミさん、ありがとうございます。無事 4000 人突破しました。(Sasami-san, thank you very much. We've successfully surpassed 4,000 people.)	😊	joy
間に合ってます爆笑 (We made it in time LMAO.)	😂	fun
寝れない。(I can't sleep.)	😭	sadness
あにちん結婚おめでとう。(Congratulations on your marriage, Anichin.)	🎉	celebration
おいしいですよね。私も大好きです。(It's delicious, isn't it? I love it too.)	🥰	love
涼しいと思ったけど最寄駅まで歩いたら暑くなった (I thought it was cool, but after walking to the nearest station, I started feeling hot.)	💦	sweat
もっと頑張ろうと思ったらもう終わってた (I was about to put in more effort, but it was already over.)	😇	angel
どんな味がするんだろ (I wonder what it tastes like.)	🤔	question

Table 4: Example posts with each of the 8 representative emojis in M_8

called emoji2vec, have been studied (Eisner et al., 2016) and utilized to cluster emojis (Lee et al., 2022). However, emoji2vec can only be applied to English and thus we do not have Japanese emoji embeddings. In this paper, we leave the issue of grouping Japanese emojis based on Japanese emoji embeddings as a future work and decided that emojis are grouped based on a survey conducted by human subjects. The details of the method of emoji grouping through a survey by human subjects are as follows.

A survey on emoji grouping 13 participants were provided with the 20 emojis shown in Table 2 without any other information about those emojis. They are asked to group those 20 emojis as follows, "これらの 20 種類の絵文字を自由にグループ化してください。ただし、使い方もしくは意味が近い絵文字同士が一つのグループになるようにグループ化を行ってください。各絵文字は一回しか使えません。(Group these 20 emojis as you like. However, please group them such that those with similar usages or meanings are grouped together. Each emoji can only be used once.)". We then compiled the obtained survey results by counting the number of occurrences of each pair of emojis among all survey responses. If the number of occurrences is seven or larger (i.e., more than half of the 13 participants agreed to group the pair together) we consider the two emojis eventually belong to the same group. We examined the number of occurrences of all the pairs, and the final emoji-grouping result is shown in Table 3.

While there exist initially 20 emojis before replacement, through replacement shown in Table 3, 8 emoji groups and their representative emojis with group names are obtained as in Table 3. The resulting dataset after replacement is referred to as M_8, where, as shown in the last column of Table 3, M_8 has a biased distribution in that nearly one-third of it belongs to the "joy" emoji group. In the evaluation of section 6, this bias makes the pre-trained models advantageous when evaluated against the similarly biased test dataset, because they are all trained with those biased training dataset. ChatGPT, on the other hand, is fine-tuned with the unbiased training dataset with the uniform distribution. For each of the 8 representative emojis, Table 4 shows an example post found in M_8 and its English translation.

4.2 Training and Test Datasets

While emojis are grouped, it is still difficult for models to properly predict emojis for X posts because posts annotated with emojis may happen to express emotions that do not semantically coincide with the texts. Therefore, we decided to develop datasets of posts that are emoji-predictable by humans from M_8. The procedure of creating the datasets is outlined as follows, where the target amount of posts for a predictable dataset is given as N and the steps 1, 2, and 3 below are repeated until N is reached.

1. A random post p is selected from the dataset M_8, where its text is denoted as t_p and its emoji as e_p.
2. The first author examines only the text t_p and predicts the most appropriate emoji $\hat{e}$ for the text t_p.
3. If $\hat{e}$ equals to e_p, the post p is added to the dataset consisting of posts that are emoji-predictable by humans.

Prompts without description of common usages of emojis	Prompts with description of common usages of emojis
Choose an emoji that is the most appropriate to the tweet [*This place is for the actual post for prediction*] from the choices. And answer only the emoji you chose. Choices: 😊, 😂, 😭, 🎉, 🥰, 💦, 😇, 🤔	Choose an emoji that is the most appropriate to the tweet [*This place is for the actual post for prediction*] from the choices. And answer only the emoji you chose. **However, the following rules apply. 🎉 is only used after text like "congratulations". 🥰 presents feelings like "like", "happiness", "wonderful". 💦 represents "stress", "nervousness", "sweat". 😇 represents "it's over", "I'm done", "angel".** Choices: 😊, 🤣, 😭, 🎉, 🥰, 💦, 😇, 🤔

Table 5: Prompts for LLMs (English translation of Japanese prompts).

For the predictable datasets, we develop training, validation, and test datasets to examine the performance of the emoji prediction task. The specific procedure for creating training, validation, and test datasets is outlined below.

Validation and test datasets The number of emoji-predictable posts N is set as 360. Following the procedure mentioned above, emoji-predictable datasets each comprising 360 randomly selected posts are designated as a validation dataset denoted as V_{8h}, and a test dataset denoted as T_{8h}. Additionally, datasets each comprising 360 posts randomly collected from M_8, regardless of whether they are emoji-predictable by humans or not, are designated as a validation dataset V_{8M}, and a test dataset T_{8M}. Furthermore, datasets each comprising 360 posts randomly collected based on a uniform distribution of 8 emojis are designated as a validation dataset V_{8u}, and a test dataset T_{8u}.

Training datasets for ChatGPT From M_8, we first remove the data in all the validation datasets and the test datasets created above, resulting in a subset denoted as R_8. Then, based on a uniform distribution of 8 emojis, randomly selected emoji-predictable sets of 40, 80, 120, and 160 posts are obtained from R_8 to be the training datasets, each denoted as R_8^{40}, R_8^{80}, R_8^{120}, and R_8^{160}, respectively.

Training datasets for pre-trained models From M_8, we remove the data in all the validation datasets and the test datasets created above, denoted as R_{8p} (313,649 posts, the same as R_8), which are then used as the training dataset for fine-tuning XLM, Japanese BERT, and Japanese RoBERTa.

The number of posts within T_{8h} (360 posts) amounts to about 38% of all the posts examined in the steps 1, 2, and 3 above, meaning that the rate of predictability is approximately 38%.

5 Emoji Prediction Methods

We utilize GPT-4o (`gpt-4o-2024-05-13`), GPT-3.5 (`gpt-3.5-turbo-0125`), Claude 3.5 Sonnet (`claude-3-5-sonnet-20240620`), Gemini 1.5 Pro (`gemini-1.5-pro`) as the LLMs, as well as XLM (`FacebookAI/xlm-mlm-17-1280`), BERT (`tohoku-nlp/bert-base-japanese-v3`), and RoBERTa (`rinna/japanese-roberta-base`) as the pre-trained models. The results are evaluated based on classification accuracy (Acc) and F_1 score.

5.1 Large Language Models

5.1.1 Overview

We examine the performance of zero-shot, few-shot, and fine-tuning on the test datasets created in section 4.2 using LLMs.

For zero-shot, only the text of the test data is inputted into LLMs in the form of prompts shown in Table 5. For few-shot, 1 or 2 posts are extracted from R_8^{160} for each of the 8 emojis to create an 8-shot set and a 16-shot set. The 8-shot set and the 16-shot set are inputted also in the form of prompts shown in Table 5 before the test data to demonstrate how the models should predict each emoji.

For fine-tuning, since fine-tuning of GPT-4o was not available at the time we conducted this experiment, we only fine-tune GPT-3.5 in this paper. Fine-tuning of GPT-3.5 is performed using

model	without description of common usages of emojis	with description of common usages of emojis
	T_{8h} (Acc / F_1)	T_{8h} (Acc / F_1)
GPT-4o (zero-shot)	0.70 / 0.54	0.66 / 0.51
GPT-4o (8-shot)	0.69 / 0.53	0.68 / 0.55
GPT-4o (16-shot)	0.68 / 0.53	0.69 / 0.54
GPT-4o (fine-tuning)	0.71 / 0.54	0.70 / 0.53
GPT-3.5 (zero-shot)	0.66 / 0.49	0.67 / 0.50
GPT-3.5 (8-shot)	0.56 / 0.43	0.58 / 0.48
GPT-3.5 (16-shot)	0.53 / 0.42	0.60 / 0.49
GPT-3.5 (fine-tuning)	**0.74** / **0.56**	0.69 / 0.56
Claude 3.5 Sonnet (zero-shot)	0.69 / 0.53	0.75 / 0.53
Claude 3.5 Sonnet (8-shot)	0.73 / **0.56**	**0.78** / **0.61**
Claude 3.5 Sonnet (16-shot)	0.72 / 0.54	0.75 / 0.59
Gemini 1.5 Pro (zero-shot)	0.64 / 0.49	0.66 / 0.50
Gemini 1.5 Pro (8-shot)	0.69 / 0.51	0.71 / 0.54
Gemini 1.5 Pro (16-shot)	0.67 / 0.50	0.70 / 0.55
XLM	0.73 / **0.56**	N/A
BERT	0.70 / 0.54	N/A
RoBERTa	0.71 / 0.51	N/A

Table 6: Acc and F_1 scores of the test dataset that is emoji-predictable by humans (T_{8h}). Bold text indicates the highest Acc and F_1 scores of each setting.

the training data created in section 4.2. The optimal settings for the number of training data and epochs are explored using the validation dataset. After fine-tuning, only the text of the test data is inputted into the model with the optimal setting, and the Acc and F_1 score of the prediction results are measured.

5.1.2 Prompts

To address the issue of LLMs' misuse or misrecognition of emojis in Japanese, we provide LLMs with description of common usages of emojis in Japanese posts in the prompts. In order to keep the prompts short, we provide description of common usages only for four emojis that LLMs frequently misuse or misrecognize instead of providing all. Both "Prompts without description of common usages of emojis" setting and "Prompts with description of common usages of emojis" setting are shown in Table 5. Bold text indicates the description of common usages of the four emojis that LLMs frequently misuse or misrecognize.

5.2 Pre-Trained Models

Fine-tuning of pre-trained models are conducted with the dataset R_{8p} created in section 4.2. Optuna[8] is utilized to search for optimal settings of batch size and learning rate using the validation dataset. After fine-tuning, only the text of the test data is inputted into the model with the optimal setting, and the Acc and F_1 score of the prediction results are then measured.

[8] https://optuna.org/

6 Evaluation Results

The results on the test dataset that is emoji-predictable by humans (T_{8h}) are shown in Table 6. Because the pre-trained models do not use prompts, description of common usages of emojis is not available. Overall, Claude (8-shot), which achieved an Acc of 0.78 and an F_1 score of 0.61, performs the best. Fine-tuning on GPT-3.5 is confirmed effective in terms of the emoji prediction task, where it outperformed any other models when without description of common usages of emojis. Table 7 and Table 8 show Acc and F_1 scores of the test datasets that are created regardless of whether they are emoji-predictable by humans or not (T_{8M} and T_{8u}).

In contrast to T_{8h} and T_{8M} where XLM performs the best among those three pre-trained models, in T_{8u}, RoBERTa performs the best, achieving an Acc of 0.38 and an F_1 score of 0.35. A probable reason why XLM underperforms RoBERTa is due to the number of parameters of the models. XLM (570M parameters) carries more parameters than RoBERTa (110M parameters) do. Considering that T_{8u} contain more posts that are

model	without description of common usages of emojis	with description of common usages of emojis
	T_{8M} (Acc / F_1)	T_{8M} (Acc / F_1)
GPT-4o (zero-shot)	0.35 / 0.25	0.34 / 0.25
GPT-4o (8-shot)	0.33 / 0.24	0.35 / 0.26
GPT-4o (16-shot)	0.32 / 0.22	0.34 / 0.25
GPT-4o (fine-tuning)	0.33 / 0.23	0.33 / 0.22
GPT-3.5 (zero-shot)	0.30 / 0.23	0.32 / 0.22
GPT-3.5 (8-shot)	0.31 / 0.24	0.34 / 0.25
GPT-3.5 (16-shot)	0.33 / 0.21	0.34 / 0.23
GPT-3.5 (fine-tuning)	0.33 / 0.24	0.33 / 0.24
Claude 3.5 Sonnet (zero-shot)	0.35 / 0.27	0.39 / 0.26
Claude 3.5 Sonnet (8-shot)	0.34 / 0.27	**0.40 / 0.28**
Claude 3.5 Sonnet (16-shot)	0.35 / 0.27	0.39 / 0.25
Gemini 1.5 Pro (zero-shot)	0.33 / 0.25	0.34 / 0.25
Gemini 1.5 Pro (8-shot)	0.33 / 0.27	0.37 / 0.26
Gemini 1.5 Pro (16-shot)	0.32 / 0.25	0.33 / 0.25
XLM	**0.48 / 0.38**	N/A
BERT	0.46 / 0.33	N/A
RoBERTa	0.47 / 0.36	N/A

Table 7: Acc and F_1 scores of the test dataset T_{8M}. Bold text indicates the highest Acc and F_1 scores of each setting.

not emoji-predictable by humans, XLM may require more training data when trained with emoji-unpredictable posts than when trained with emoji-predictable posts, resulting in that XLM underperforms RoBERTa against T_{8u}.

The major cause of why XLM achieved almost the same performance as GPT-3.5 (fine-tuning) can be explained from distribution of datasets. As we mentioned in section 4.1, both the training dataset of XLM and the test dataset T_{8h} have the biased distribution with the dominant "joy" emoji class, while GPT-3.5 is fine-tuned with the unbiased training dataset with the uniform distribution.

This is contrastive with the evaluation results of the test dataset T_{8u} (having the uniform distribution) in Table 8, where RoBERTa outperforms GPT-3.5 (fine-tuning). This is also because both the training dataset of GPT-3.5 fine-tuning and the test dataset T_{8u} have the unbiased uniform distribution, while the training dataset of RoBERTa is still biased.

In all settings except GPT-4o and GPT-3.5 (fine-tuning), the results of "with description of common usages of emojis" are generally better than those of "without description of common usages of emojis". However, the difference is small in most cases. The probable reason why performance cannot be improved through description of common usages of emojis is because before given description, GPT-4o and GPT-3.5 (in GPT-3.5's case, through fine-tuning) has already gained more knowledge about usage of emojis than the description. Therefore, it can happen that the given description did not contribute to improving the models' performance. On the other hand, as easily expected, the results of the test datasets that are emoji-predictable by humans are far more better than the test datasets that are created regardless of whether they are emoji-predictable by humans or not. Unlike previous works on emoji prediction, this paper experimentally confirmed that it is easier to predict emojis of posts that are emoji-predictable by humans than those that are not. Regarding posts that are emoji-unpredictable by humans, they may contain emotions that do not semantically coincide with the texts, which prevents them from being correctly emoji-predicted. The analysis of usages of emojis used in these posts and their characteristics is our future work.

7 Evaluation on English X Posts

In order to evaluate the performance of our emoji prediction models against an existing English posts dataset for emoji prediction, we evaluate the pre-trained models BERT, RoBERTa and XLM applied to our Japanese datasets with an English dataset (Baziotis et al., 2018). We avoid apply-

model	without description of common usages of emojis	with description of common usages of emojis
	T_{8u} (Acc / F_1)	T_{8u} (Acc / F_1)
GPT-4o (zero-shot)	0.33 / 0.30	0.31 / 0.29
GPT-4o (8-shot)	0.30 / 0.27	0.33 / 0.30
GPT-4o (16-shot)	0.32 / 0.25	0.32 / 0.31
GPT-4o (fine-tuning)	0.32 / 0.27	0.33 / 0.31
GPT-3.5 (zero-shot)	0.26 / 0.24	0.29 / 0.26
GPT-3.5 (8-shot)	0.20 / 0.20	0.24 / 0.23
GPT-3.5 (16-shot)	0.21 / 0.20	0.23 / 0.21
GPT-3.5 (fine-tuning)	0.32 / 0.28	0.31 / 0.28
Claude 3.5 Sonnet (zero-shot)	0.34 / 0.32	0.31 / 0.27
Claude 3.5 Sonnet (8-shot)	0.33 / 0.31	0.34 / 0.31
Claude 3.5 Sonnet (16-shot)	0.31 / 0.30	**0.32 / 0.30**
Gemini 1.5 Pro (zero-shot)	0.31 / 0.28	0.31 / 0.27
Gemini 1.5 Pro (8-shot)	0.33 / 0.28	0.30 / 0.27
Gemini 1.5 Pro (16-shot)	0.31 / 0.26	0.30 / 0.24
XLM	0.36 / 0.32	N/A
BERT	0.33 / 0.30	N/A
RoBERTa	**0.38 / 0.35**	N/A

Table 8: Acc and F_1 scores of the test dataset T_{8u}. Bold text indicates the highest Acc and F_1 scores of each setting.

	SVM	FacebookAI/ xlm-mlm-17-1280 (XLM)	google-bert/ bert-base-cased (BERT)	FacebookAI/ roberta-base (RoBERTa)
Acc / F_1	0.45 / 0.31	0.46 / 0.31	0.49 / 0.35	0.51 / 0.37

Table 9: Acc and F_1 scores of emoji prediction for English datasets.

ing LLMs because the number of the test data is too large. We then reexperiment on the English dataset (Baziotis et al., 2018) by applying SVM that was evaluated in the prior study (Çöltekin and Rama, 2018) and achieves the best F_1 score in SemEval 2018 Task 2 (Barbieri et al., 2018a). For the pre-trained models BERT and RoBERTa, we specifically evaluate their English versions (Devlin et al., 2019; Liu et al., 2019). The dataset consists of 491,665 training data, 50,000 trial data and 50,000 test data and we conducted the experiment in the same manner as described in section 5.2. Their evaluation results are shown in Table 9, where the pre-trained models BERT, RoBERTa and XLM outperform SVM that performed the best in SemEval 2018 Task 2 (Barbieri et al., 2018a).

8 Conclusion

This paper examined the performance of emoji prediction for Japanese X posts utilizing large language models and compared their performance with pre-trained models. By grouping emojis and replacing them with representative ones while selecting posts that are emoji-predictable by humans, we achieved high Acc and F_1 score. It turns out that overall, Claude performs the best among all the models used in this paper. Additionally, we discovered that, in some cases, by inputting description of common usages of emojis into prompts, we can achieve slightly better performance. On the other hands, for posts that are emoji-unpredictable by humans, it is necessary to analyze usages of emojis used in these posts and their characteristics to discover the reason why models fail to predict emojis of those posts. As mentioned in section 4.1, emojis are grouped based on opinions of 13 survey participants. This could create some biases, so we plan to group emojis according to certain embeddings of emojis. It is also another significant future work to extend our experiment to a multi-label task since some posts can contain multiple emotions and can be followed by multiple emojis.

References

Francesco Barbieri, Miguel Ballesteros, and Horacio Saggion. 2017. Are emojis predictable? In *Proc. 15th EACL*, pages 105–111.

Francesco Barbieri, Jose Camacho-Collados, Francesco Ronzano, Luis Espinosa-Anke, Miguel Ballesteros, Valerio Basile, Viviana Patti, and Horacio Saggion. 2018a. SemEval 2018 task 2: Multilingual emoji prediction. In *Proc. SemEval*, pages 24–33.

Francesco Barbieri, Luis Espinosa-Anke, Jose Camacho-Collados, Steven Schockaert, and Horacio Saggion. 2018b. Interpretable emoji prediction via label-wise attention LSTMs. In *Proc. EMNLP*, pages 4766–4771.

Christos Baziotis, Athanasiou Nikolaos, Athanasia Kolovou, Georgios Paraskevopoulos, Nikolaos Ellinas, and Alexandros Potamianos. 2018. NTUA-SLP at SemEval-2018 task 2: Predicting emojis using RNNs with context-aware attention. In *Proc. SemEval*, pages 438–444.

Spencer Cappallo, Thomas Mensink, and Cees G.M. Snoek. 2015. Image2emoji: Zero-shot emoji prediction for visual media. In *Proc. 23rd ACMMM*, page 1311–1314.

Çağrı Çöltekin and Taraka Rama. 2018. Tübingen-Oslo at SemEval-2018 task 2: SVMs perform better than RNNs in emoji prediction. In *Proc. SemEval*, pages 34–38.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Proc. 33rd NeurIPS*, volume 32, page 1–11.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proc. NAACL*, pages 4171–4186.

Ben Eisner, Tim Rocktäschel, Isabelle Augenstein, Matko Bošnjak, and Sebastian Riedel. 2016. emoji2vec: Learning emoji representations from their description. In *Proc. 4th SocialNLP*, pages 48–54.

Bjarke Felbo, Alan Mislove, Anders Søgaard, Iyad Rahwan, and Sune Lehmann. 2017. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. In *Proc. EMNLP*, pages 1615–1625.

Preni Golazizian, Behnam Sabeti, Seyed Arad Ashrafi Asli, Zahra Majdabadi, Omid Momenzadeh, and Reza Fahmi. 2020. Irony detection in Persian language: A transfer learning approach using emoji prediction. In *Proc. 12th LREC*, pages 2839–2845.

Jingun Kwon, Kobayashi Naoki, Hidetaka Kamigaito, Hiroya Takamura, and Manabu Okumura. 2022. Joint modeling of emoji position and its label for better understanding in social media. *Journal of NLP*, 29(2):467–492.

SangEun Lee, Dahye Jeong, and Eunil Park. 2022. Multiemo: Multi-task framework for emoji prediction. *Knowledge-Based Systems*, 242:108437.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized bert pretraining approach. arXiv:1907.11692.

Weicheng Ma, Ruibo Liu, Lili Wang, and Soroush Vosoughi. 2020. Multi-resolution annotations for emoji prediction. In *Proc. EMNLP*, pages 6684–6694.

Gopendra Vikram Singh, Dushyant Singh Chauhan, Mauajama Firdaus, Asif Ekbal, and Pushpak Bhattacharyya. 2022. Are emoji, sentiment, and emotion Friends? A multi-task learning for emoji, sentiment, and emotion analysis. In *Proc. 36th PACLIC*, pages 166–174.

Toshiki Tomihira, Atsushi Otsuka, Akihiro Yamashita, and Tetsuji Satoh. 2018. What does your tweet emotion mean? Neural emoji prediction for sentiment analysis. In *Proc. 20th iiWAS*, page 289–296.

Toshiki Tomihira, Atsushi Otsuka, Akihiro Yamashita, and Tetsuji Satoh. 2020. Multilingual emoji prediction using bert for sentiment analysis. *International Journal of Web Information Systems*, 16:265–280.