Synchronizing Approach in Designing Annotation Guidelines for Multilingual Datasets: A COVID-19 Case Study Using English and Japanese Tweets

Kiki Ferawati
Nara Institute of Science and Technology
Japan
kiki.ferawati.kb6@is.naist.jp

Wan Jou She
Kyoto Institute of Technology
Japan
wjs2004@kit.ac.jp

Shoko Wakamiya and **Eiji Aramaki**
Nara Institute of Science and Technology
Japan
{wakamiya, aramaki}@is.naist.jp

Abstract

The difference in culture between the U.S. and Japan is a popular subject for Western vs. Eastern cultural comparison for researchers. One particular challenge is to obtain and annotate multilingual datasets. In this study, we utilized COVID-19 tweets from the two countries as a case study, focusing particularly on discussions concerning masks. The annotation task was designed to gain insights into societal attitudes toward the mask policies implemented in both countries. The aim of this study is to provide a practical approach for the annotation task by thoroughly documenting how we aligned the multilingual annotation guidelines to obtain a comparable dataset. We proceeded to document the effective practices during our annotation process to synchronize our multilingual guidelines. Furthermore, we discussed difficulties caused by differences in expression style and culture, and potential strategies that helped improve our agreement scores and reduce discrepancies between the annotation results in both languages. These findings offer an alternative method for synchronizing multilingual annotation guidelines and achieving feasible agreement scores for cross-cultural annotation tasks. This study resulted in a multilingual guideline in English and Japanese to annotate topics related to public discourses about COVID-19 masks in the U.S. and Japan.

1 Introduction

The close political bond and distinct cultural viewpoints have resulted in the U.S.-Japan comparison being a frequently studied topic among researchers interested in cultural contrasts and variations in Western and Eastern societal conventions. This extends to how they behave in daily life and how they responded to major world events, such as the pandemic, making it a great target for a multilingual annotation case study. In the early stages of COVID-19, many health personnel and epidemiologists advocated the importance of mask in curbing the infection (Zeng et al., 2020; Leung et al., 2020; Asadi et al., 2020). However, the U.S. and Japan displayed contrasting attitudes regarding policy implementation to control the spread of the disease and adherence to mask mandates, as demonstrated by their respective governments and citizens (Netburn, 2021; KYODO, 2023; Reich, 2020). This was also observed by Tso and Cowling (2020), who summarized the general use of masks from several countries, including the U.S. and Japan, and suggested further measure to improve mask effectiveness. While the general population recognized the importance of mask in the U.S., whether they actually wore them largely related to their demographics and their beliefs about societal value (Bir and Widmar, 2021). On the other hand, wearing masks was common in Japan even before COVID-19, as observed in the majority of respondents in Suppasri et al. (2021) who had no issue with wearing masks.

Further information about how the two countries reacted to COVID-19 mask-wearing mandate can also be observed from social media. A study by Lin et al. (2022) found that mask mandates and mask-wearing, as obtained from geo-tagged Twitter (now X) images, had a strong association in the U.S. Tweets also provided insights on public opinion about mask-wearing and their reasons for not wearing any (He et al., 2021). Another study identified mask as one of the most frequently used words in tweets from Korea and Japan (Lee et al., 2020), and it also appeared in the list of top concerns expressed through social media during the

Proceedings of the 2nd Workshop on Cross-Cultural Considerations in NLP, pages 32–41

pandemic in Japan (Kamba et al., 2024). Undoubtedly, the vast number of social media posts is often considered a valuable resource for understanding, analyzing, and even informing policymakers about societal perceptions or attitudes toward certain events. However, a significant challenge for cross-cultural comparison studies is obtaining a comparable data from two different languages with different cultural backgrounds.

To get such kind of data, we applied a synchronizing approach in annotating our data. Given the differing responses to mask mandates as a measure against COVID-19 in the U.S. and Japan, we are interested in exploring the variations in the public responses. For instance, we are interested in whether individuals actually wear masks despite their stated stance against masking. However, this presents a complex challenge in the annotation process, hence the decision to split the question into a simpler form. Annotation plays a crucial role in many natural language processing (NLP) research, as the annotation results will provide a foundation for future model training work. We adopted a synchronized approach in which we borrowed linguistic nuances and annotation insights from both languages to design and refine the guidelines to ensure the guidelines achieve a desirable level of agreement, accuracy, and effectiveness in capturing issues related to COVID-19 masking we aim to assess in both cultures. The multilingual guidelines can be accessed as supplementary material of this paper [1].

2 Related work

Utilizing human annotated data in training machine learning classifiers was a practice often employed in Twitter studies (O'dea et al., 2015; Mozetič et al., 2016). Previous COVID-19 studies have also employed such a method, as observed in Klein et al. (2021) who used annotated tweets for COVID-19 tracking to identify potential tweets reporting COVID-19 cases in the U.S.

Research involving multiple languages, especially in tweets, is often done in comparative studies where two or more target populations are using different languages. For example, Zotova et al. (2020) compared stance detection in two languages (Catalan and Spanish), designing their guidelines based on the approach introduced by Bosco et al. (2016), with annotations performed by two annotators who are skilled in both languages.

Another study by Jahan (2020) completed a task of offensive language comparison of tweets in five different languages using the English translation of the tweets. Translating all the tweets into English seems feasible to a degree and can benefit from a powerful English-based language model. Similarly, Chen et al. (2022) studied COVID-19 vaccination attitudes using translated tweets from four Western European countries, resulting in a dataset with potential use for COVID-19 analysis. However, using translation tools might result in the inaccurate conveyance of the sentiment (Lohar et al., 2017).

Analyzing the tweet in its native language ensures that the annotators get to comprehend the original meaning, as well as the cultural nuances of the texts, which could easily get lost through translation. Considering the positive and negative aspects of involving translation in dealing with multi language dataset, we decided to focus on annotating the tweets in their native language, utilizing our synchronized approach on guidelines creation and annotation to ensure the comparability of our datasets.

3 Dataset

3.1 Data

The data collection period spanned from January 1st, 2020 and December 31st, 2022 (36 months), using Academic Research Access X API (formerly Twitter API), which had been revoked in mid of 2023. To ensure that the tweets originated from Japan and the U.S., we applied extraction criteria to filter tweets with geo-tag metadata. Using the country location filter of `US` for the U.S. and `JA` for Japan, we obtained 1,102,876 English tweets and 589,927 Japanese tweets.

3.2 Preprocessing of tweets

The first preprocessing step was to validate each tweet's location tag. Since there are several types of location data in geo-tagged tweets, we focused on city-level information in the 'full-name' entity of the tweets. We removed instances where the geo-tagged location failed to match the city lists in the U.S. and Japan (simplemaps, 2022; MIT, 2019). Afterward, we proceeded to the text contents of the tweet. We removed links and changed all the usernames into a common '@username'. In this step, we kept the emoticon, punctuation, and capitalization of the letter in tweets to preserve unobstructed

[1] https://doi.org/10.6084/m9.figshare.26104597

information for the annotators. We also made sure that the tweets contain 'mask' for English tweets and 'マスク' (mask in Japanese) for Japanese tweets. In some cases, the keywords appeared as usernames or links, so they did not provide enough information or relevance to the topics and hence were eliminated in the process.

Stowe et al. (2018) noted in their study that most short English tweets are irrelevant and did not contribute to the annotation agreement. Hence, we excluded English tweets that were less than 25 characters. Due to the usage of kanji in Japanese writing system, which was considered as one character but can contain meanings equivalent to an English word, we applied a lower threshold of 10 characters for Japanese tweets. As a final step before preparing the annotation sample, we removed duplicates and NA's from our data.

4 Method

4.1 Initial exploration of the sample

We prepared a small set of sample tweets from both languages as a basis to create annotation guidelines. The sampling strategy consists of these criteria:

- Each User ID can only be included once in the sample.
- Randomized sampling based on unique User ID.

We began the pilot annotation phase for our English tweets and designed the draft of annotation guidelines in English to align our thoughts. The initial process included a review of tweets sample and a discussion of the potential annotating target. We then translated the guidelines into Japanese and pilot-annotated the multilingual dataset based on the primary guidelines. We went through two rounds of iterations and revisions to examine the annotation agreement for the pilot phase and unite the multilingual guidelines. On translating the guidelines and its revision, we also considered the context of the instructions and found the more close-fitting examples from Japanese tweets if necessary. The process with English tweets is described in Step 1, and the next process for Japanese tweets is described in Step 2 of Figure 1.

The following is the summary of annotation guidelines from the pilot phase, also shown in Figure 2. These are the stages of annotation carried out by the annotators:

1. Tweet relevancy, in which we classify the tweets based on their relevance to COVID-19 mask discussion.
   - Relevant: mentioning mask and related to COVID-19, such as mask-wearing, mask-related policies, masking as COVID-19 prevention measures, etc. Example: "Please wear your mask to prevent the spread of the virus!"
   - Non relevant: tweets mentioning mask but not related to COVID-19, such as beauty face mask, mask as a verb, figurative expression using mask, etc. Example: "This person mask their intention well" (not an actual face mask to prevent virus)
2. Consider only relevant tweets from the first stage. There are two parts to this second stage:
   (a) Stance stage: classify whether a tweet reflects the user's stance toward supporting, opposing mask-wearing, or unclear.
      - Supporting stance: expressing that they are willing to wear mask, promoting its benefits, positive opinion about masks, etc.
      - Opposing stance: expressing disapproval, skeptical, negative opinion about mask, listing disadvantages of wearing mask, etc.
      - Unclear: tweets without enough context to classify as supporting or opposing.
   (b) Mask-wearing stage: for tweets with a clear stance, mark whether the specific user is wearing a mask, not wearing mask, or unknown.

We explained the goal of each topic to the annotators during the annotation briefing. We discussed target output and included common patterns found in the tweets, along with annotation examples from the initial exploration step in the guideline. After reflecting on initial results and discussing with the annotators, we decided to refine the annotation guidelines by combining support and unclear stances so we could focus on outlining people against COVID-19 masking (see Figure 2). This resulted in two categories only: against and not against. The term round in this paper refers to the annotation process of incrementally releasing a certain percentage of tweets to the reviewers. One complete annotation process by the annotators was considered as one round. We have four rounds in total: 10%, 50%, 100% and repeating 100% after a final discussion with the annotators.

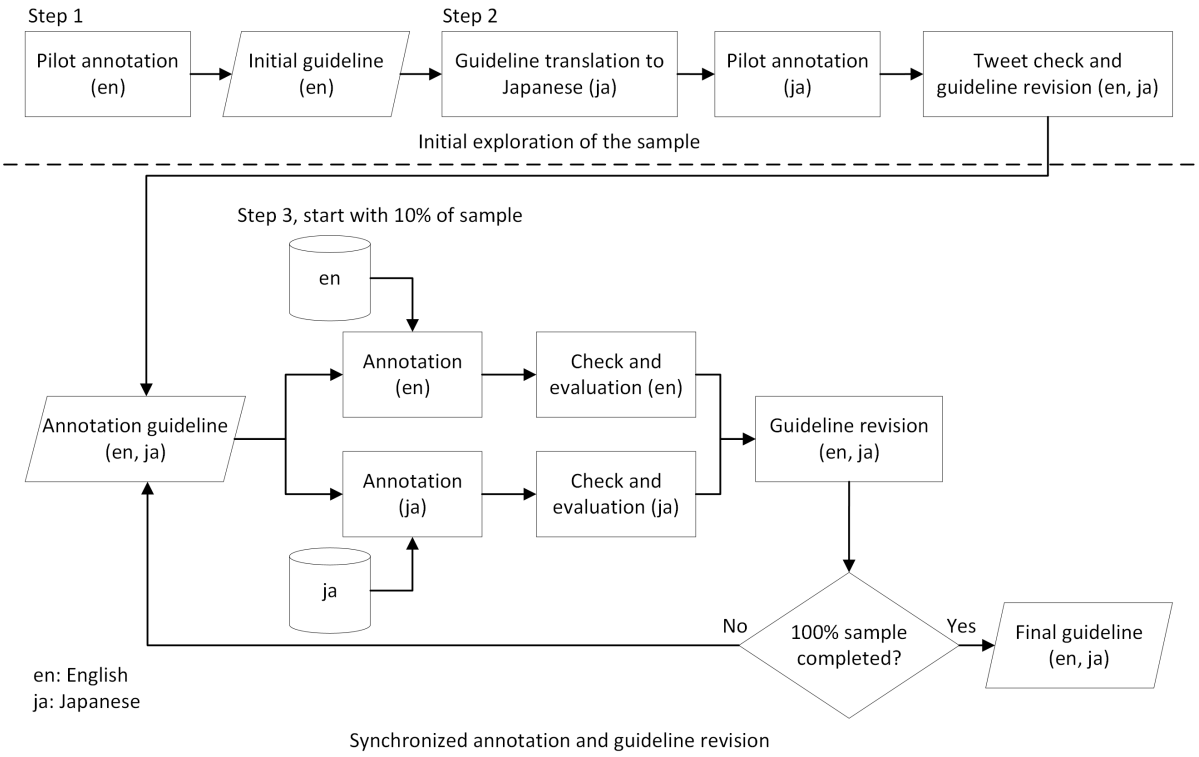

Figure 1: Initial exploration of sample and guideline design (top); Synchronized annotation process by annotators and guideline revision, started with 10% sample, 50%, and 100%, three rounds in total (bottom).

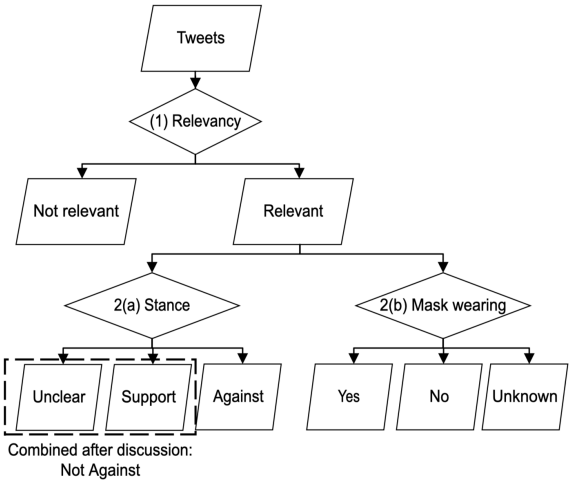

Figure 2: Flowchart of annotation topics: filter relevant tweets and mark the categories for stance and mask-wearing.

4.2 Synchronized annotation process

There were four annotators in total, two for each language. All annotators were graduate students at our university and demonstrated a good command of the target annotating language (two in Japanese and two in English). They also had experience in natural language processing tasks and studies.

Initial round with 10% sample

Using the same sampling method described in the previous subsection, we obtained 1,100 sample tweets for each language to be annotated. We explained the annotation guidelines and the expected results of the annotation. In the first round, the annotators were asked to annotate 10% tweets based on their initial understanding of the guidelines. The results were then evaluated, and the first inter-annotator agreement (IAA) was calculated. Figure 1 explained the summary of the synchronized iterative annotation process to revise the guideline for English and Japanese, illustrated in Step 3 in the figure.

Review disagreed tweets to clarify the guidelines

To determine whether disagreements were caused by miscomprehension of guidelines or genuine differences in the interpretation of the tweets, we highlighted disputed tweets and discussed them with annotators during each annotation round. Each meeting lasted about an hour, with the authors and annotators working together to identify the potential cause of miscomprehension in the guidelines. Based on these discussions, we proceeded to revise

the multilingual guidelines to prevent future misunderstandings. Changes in one language were also reflected in the other, ensuring that the new instructions remained relevant to both languages. This approach was guided by the belief that contextual nuances and clarified wording, though originating from one language, would enhance understanding in both English and Japanese. We started and completed the discussions and annotations of both English and Japanese tweets at the same timeline.

Repeat the synchronization process for the second and third rounds

After the first round of guidelines revision, the annotators were asked to annotate 50% tweets according to the refined guidelines (including re-checking the first 10% tweets). To avoid biasing the annotators, while annotators and we discussed the agreement ratios and clarified the misunderstanding of the guidelines, the team did not review the individual tweets and left the decision to change their annotation results to the annotators themselves. We then compared the IAA for the first 10% and 50%. We then repeated the annotation steps for the rest of the sample data set, 1,100 tweets, and obtained the IAA for all the samples. During the annotation phase, we continuously had discussions and revisions for the guidelines based on the feedback and annotation results of the annotators. We had a final discussion with the annotators after they completed annotating the entire sample data set and re-annotated the data after the discussion as the final round of the annotation.

4.3 Annotator agreement evaluation

We used Cohen's kappa, which is a commonly used measure to evaluate IAA for nominal annotation where the annotators operated independently (Cohen, 1960). We evaluated the IAA for each round of annotations and compared the results from each round for both languages. For the agreement calculation on the stance and mask-wearing stage, we only included tweets where both annotators agreed as relevant. The interpretation for agreement results was based on McHugh (2012). We considered a minimum threshold of weak agreement (range of 0.40 to 0.59) as an acceptable agreement level in this study.

5 Results

5.1 Inter-annotator agreement

The summary of IAA for the sample data is shown in Table 1. In Round 1, while the English tweets showed moderate agreement for tweet relevancy (0.51), the Japanese tweets demonstrated almost perfect agreement (0.91) between annotators. We discussed the guidelines and clarified the ambiguous points with the annotator. Most of the concerns were about criteria for relevant tweets and tweets containing sarcasm, such as "mask doesn't protect you, right" and other examples as shown in Table 2. We then revised the guidelines based on the feedback from the annotators. The final Cohen's kappa for the relevancy stage is at a similar value of 0.65 for English tweets and 0.67 for Japanese tweets, both in the moderate agreement level.

In the second stage of the annotation, we focused on the stance about COVID-19 masking. Both English and Japanese tweets showed a weak agreement in the first round, with 0.52 for English and 0.58 for Japanese. The agreement for Japanese tweets was consistent in the following rounds, even with the guideline revision, still hovering around the weak agreement (0.59 for the final round). These results suggested that the stance on masking was challenging for annotators to agree upon in both English and Japanese.

The last topic, mask-wearing behavior, showed a very low Cohen's kappa score in the earliest round, spotting minimal agreement with 0.21 for English and 0.35 for Japanese. However, the final agreement showed an agreement of 0.55 for English and 0.47 for Japanese, a desirable increase compared to the value in the first round. The final results show a similar value for the stance stage, although only a weak agreement. Our result suggested that although determining behavior tendency (e.g., mask-wearing) from tweets was also challenging for the annotators, the synchronized approach proposed by us did help to improve the overall agreement through stages and achieve desirable agreement scores for both languages.

5.2 Guideline improvement

There was a consistent increase in the relevancy stage after each round of annotation and guideline iteration. On the same 10% part of the sample, the agreement score of English tweets increased from 0.51 to 0.67 and finally to 0.71. While Japanese tweets yielded a high score initially (0.91), leaving

		Stage 1: Relevancy		Stage 2(a): Stance		Stage 2(b): Mask-wearing	
Round	**Sample**	**English**	**Japanese**	**English**	**Japanese**	**English**	**Japanese**
1	10%	0.51	0.91	0.52	0.58	0.21	0.35
2	10%	0.67	0.95	0.55	0.56	0.45	0.39
	50%	0.58	0.74	0.48	0.55	0.57	0.57
3	10%	0.71	0.95	0.53	0.55	0.46	0.35
	50%	0.65	0.87	0.46	0.58	0.56	0.54
	100%	0.65	0.67	0.42	0.55	0.53	0.49
Final	100%	0.79	0.92	0.46	0.59	0.55	0.47

Table 1: Cohen's kappa for the annotation results. Percentage in each round shows the number of samples annotated. The agreement is calculated for each round and each part of samples.

little rooms to improve, there was still a slight increase of the agreement score (0.95) in the second round. Further observation also showed that the addition of the samples results in a slight decrease (between 0 to 0.07) in the Cohen's kappa score for both languages, except for Japanese tweets in Stage 1, Round 3, with 0.2 decrease from 50% of sample to 100%. The score for 100% English tweets sample was 0.65, the same as the score obtained from 50% sample, suggesting that the latest version of the guideline helped the annotators to achieve a consistent outcome. However, such an improvement was not replicated in the Japanese samples. The agreement for the final 100% sample surprisingly decreased to 0.67 in the moderate category while demonstrating stable improvement in the 50% samples (0.87). After a final discussion with the annotators and a last modification of the guidelines, the re-annotated sample resulted in a higher agreement of 0.79 and 0.92 for English and Japanese sample, respectively.

Regarding the annotation for stance on COVID-19 mask (Stage 2(a)), the increase on each rounds of annotation is not apparent, with the Cohen's kappa score staying in the similar range throughout the annotation process. In the mask-wearing stage (Stage 2(b)), there was a notable increase in the 10% of the sample from 0.21 to 0.45 and 0.46 for the English tweets. The final score for 100% sample capped at 0.55, though it did not differ much from the previous set of samples. The Japanese annotation scores were fluctuating between 0.35 and 0.39 for the initial 10% samples regardless of the iterative procedure, while showing a good improvement for the final full sample (0.47). All annotators both settled at around weak agreement for this round, with English agreement scores slightly surpassed the Japanese ones in the final round.

6 Discussions

6.1 The process of guidelines synchronization

Our findings showed that revising and synchronizing the guidelines after each round of annotation significantly enhanced the IAA score. We reviewed the guidelines after each round and clarified any ambiguous or misleading instructions based on disagreements found in both English and Japanese annotations. During discussions with annotators, we intentionally incorporated linguistic nuances (e.g., expressions and examples) and disagreement rationales from both languages to ensure consistent understanding across all annotation contexts. Additionally, disagreements in other languages appeared to complement each other, enabling us to clarify guidelines and rectify potentially ambiguous explanations. Annotators were given the opportunity to review previous samples in subsequent rounds and adjust their decisions according to the latest guidelines.

Our annotators found that marking guideline revisions in a different color helped them identify key changes across rounds. Since annotation was conducted simultaneously in both languages, guideline updates were synchronized, incorporating findings and suggestions from both languages. For instance, additional instructions based on Japanese tweets were translated into English with appropriate examples to maintain synchronization. While reviewing the guidelines, we also noticed some similarities observed in the source of annotation mismatch between the two languages despite the differences in culture, such as sarcastic and ambiguous tweets which appeared in both Japanese and English tweets (discussed in the following section).

However, there appears to be a limitation when applying it to annotating individuals' stances reflected in the tweets. For instance, in Stage 2(a), we compared the annotation results and attempted to synchronize the guidelines by clarifying the ambiguous explanations that were causing disagreements in both languages. While we were able to improve the agreement score of the annotation outcome, there was only a limited increase in the score, making it hard to justify the benefit of such an approach. Perhaps individuals from different cultures interpret the concept of "stance" differently, and researchers should caution about annotating the concepts that are not identical in different cultures using such an approach.

Overall, the increase in the IAA after each round of annotation and revision suggested that the newer version of guidelines showed a better performance for mask-related tweets in the U.S. and Japan. By synchronizing the annotation guidelines after each annotation rounds, we managed to incorporate all the changes and suggestions from two different languages in a single guideline. Considering the difficulty of annotating a multilingual data set for intercultural comparison, we believe such an approach was critical in offering an operationalizable practice to achieve a stable performance.

6.2 Annotation difficulty

Sarcasm and ambiguous expression

We observed that tweets containing "sarcastic expressions" and "culturally specific expressions" posed additional challenges for annotators in determining whether the tweet implies a positive or negative stance toward mask-wearing policies. Below we dive into the nuances of each linguistically challenging annotation tasks.

Sarcasm and tweets written in ambiguous expressions in English were one of the sources of disagreement between annotators. Number 1-2 in Table 2 show an example of sarcastic tweets. Non-native and native English differ in their ability to identify written sarcasm (Techentin et al., 2021). While language understanding and cultural differences could also impact annotation, previous results in sarcasm classification research showed that the difficulties experienced by the annotators did not result in significant degradation to the expected result (Joshi et al., 2016).

While both languages showed a number of tweets containing sarcasm, there were some differences in how they are written. A study by Prichard and Rucynski (2022) shows that English sarcasm is difficult to identify by Japanese students, suggesting that the type of sarcasm is different between the languages. Our study results were also aligned with the previous literature. As suggested by Obana and Haugh, sarcasm in Japanese sometimes include the inappropriate use of honorifics which, depending on the use cases, can be interpreted as sarcastically polite, such as using higher level of honorifics than necessary (Obana and Haugh, 2021).

Regarding tweets that contained ambiguous expression, annotators showed disagreement because they could not comprehend what the actual meaning of the tweets were. The ambiguous expressions appear more frequently in Japanese, which was also identified by a study conducted by Suzuki et al. (2017). To overcome the challenge of these two types of tweets, we asked the annotators to note down the type of ambiguous tweets in the process of annotation. We incorporated the information from annotators, created a more detailed explanations on actions involved in the tweets. The difference in frequency of the problem appeared between languages was shown clearly in the first round of annotation, where there were big observable differences in the agreement score.

Lack of context and cultural nuances

Another reason for the difficulty was caused by the word limit and short nature of tweets, which often offered insufficient to no information for critical contexts. For example, if a tweet was part of a thread or conversation, the context of discussion might be missing and causing difficulty to interpret, which also applies to our data in terms of relevancy or stance, as listed in number 3-6 in Table 2. This problem of insufficient context and little content in tweets was also a concern mentioned in Stowe et al. (2018). Furthermore, our study indicated that the modified communication style conveyed in the short form of texts could present extra difficulty for the annotation tasks.

The findings of our study confirmed that cultural aspect of the community, such as individualistic or collectivist, is one of the source of the differences in communication style in social media (Garcia-Gavilanes et al., 2013). The differences originated from the cultural background between the two countries were also apparent in the study by Acar and Deguchi (2013), where the habit of the users also shapes the posts type. Tweets

No	Tweet example	Annotation difficulty
1	Don't worry I'm always masked! Always hot!	sarcastic tweet
2	Yes, mask doesn't protect you, right	sarcastic tweet
3	@username how is it going through the mask?	not enough context to infer what the user stance is about COVID-19 masking
4	Order your mask now!	not enough context on which type of mask is talked about
5	people need to learn how to wear mask in public place like this!	tweets implying a complain about people not wearing mask, but no clear indication whether the user is actually wearing any
6	mask! #citylife #lovemycity #enjoy	not enough context

Table 2: Difficulty example, showing tweets and the reasons why the particular tweet is categorized as difficult. Examples were obtained from both languages but shown in English.

from the U.S. reflects more question-like type as a way to connect with others, while in the Japanese tweets' case, questions is perceived as a sign of disharmony. Japanese users favor a relatively reserved and courteous communication style (Midooka, 1990), which was also reflected in their online posts, such as preferring harmony while tweeting even when they intended to express their opinions (Acar and Deguchi, 2013). However, these communication styles easily resulted in users posting ambiguous tweets or using indirect expressions. Annotators should be debriefed such a potential tendency when referring to the guidelines.

Annotation topics

The topics of annotations could imply various difficulties because they demand navigating intricate linguistic textures, contextual nuances, and diverse expressions within constrained character limits. In our tasks, the task to identify relevance of COVID-19 were deemed less challenging; whereas, the tasks of identifying the stance and mask-wearing status were deemed highly challenging. A potential explanation for why relevance assessment posed a low difficulty lies in annotators' ease in judging the presence or absence of the target topic (COVID-19).

As mentioned earlier, relevancy was the easiest to mark and distinguish, as shown by the agreement results between annotators for both languages. The instructions for relevancy are also fairly straightforward, with unclear tweets marked as irrelevant. On the other hand, the stance stage shows an overall weak agreement. This annotation topic was also a concern in other research involving annotation, where Mohammad et al. (2016) mentioned that determining stance can be difficult for human annotators without a proper understanding of the full context of the text. Addawood et al. (2017) noticed the consistent result of classification and feedback from human annotators having difficulty deciding between favor and neutral category. Determining the stance of the user was proven to be difficult, especially if the source is a short text in the form of tweets which sometimes lack of enough context to be inferred, as we mentioned before. Sometimes the annotators cannot be sure which category of stance the tweets are in.

Slightly different compared to the other stages, the main difficulty for the mask-wearing stage is determining the subject of the tweets, as tweets often do not follow a proper sentence pattern. In Japanese, subject is sometimes omitted and not mentioned clearly, which was also found in tweets data as observed by Akahori et al. (2021). The main source of differences found in this study is that sometimes even if the tweets clearly mention someone is wearing a mask, the annotators are not sure as to who is the one wearing the mask. However, the latest version of the guideline improved the agreement on this point in reaching a weak agreement instead of the minimal agreement on the first version of the guidelines.

7 Conclusion

Creating a guideline covering more than one language can be challenging. Our approach consists of simultaneous annotation to synchronize the guideline in order to achieve a reliable and comparable dataset. The final guideline designed in this study shows adequate results for both languages, even if

the two sets of tweet sources come from different cultural backgrounds.

This study resulted in multilingual annotation guidelines in English and Japanese for classifying tweets' relevancy, stance, and mask-wearing status. The final version of the guideline can be utilized to obtain more annotated sample for the future work on the comparison between mask opinion in the two countries.

Limitations

This study is currently limited to English and Japanese for short and informal text, i.e., tweet posts. Furthermore, we imposed a dichotomized options pair (e.g., against vs not against) and omitted neutral option because we observed a lot of ambiguity and unclear tweets in the second stage, especially the stance stage. The stage are difficult to analyze due to the ambiguity and relatively short text, sometimes insufficient to detect the stances, so we decided not to include a neutral opinion as our option and limited our choice for against or not. Population-wise, since we used geo-tagged tweets only, our sample is also limited to people who chose to provide their location information in their tweets.

Ethical consideration

This study did not require participants to be involved in any physical or mental intervention. The data in this study also did not use personally identifiable information, thus exempted from institutional review board approval in accordance with the Ethical Guidelines for Medical and Health Research Involving Human Subjects stipulated by the Japanese national government.

We made sure that the annotators could work comfortably throughout the annotation period, with reasonable working flexibility. The annotators also received compensation based on the standard rate of part-time work at our university.

Acknowledgements

This work was supported by JST SICORP Grant Number JPMJSC2107, and JSPS KAKENHI Grant Number JP22K12041, Japan.

References

Adam Acar and Ayaka Deguchi. 2013. Culture and social media usage: Analysis of japanese twitter users. *International Journal of Electronic Commerce Studies*, 4(1):21–32.

Aseel Addawood, Jodi Schneider, and Masooda Bashir. 2017. Stance classification of twitter debates: The encryption debate as a use case. In *Proceedings of the 8th international conference on Social Media & Society*, pages 1–10.

Tatsuki Akahori, Kohji Dohsaka, Masaki Ishii, and Hidekatsu Ito. 2021. Efficient creation of japanese tweet emotion dataset using sentence-final expressions. In *2021 IEEE 3rd Global Conference on Life Sciences and Technologies (LifeTech)*, pages 501–505. IEEE.

Sima Asadi, Christopher D Cappa, Santiago Barreda, Anthony S Wexler, Nicole M Bouvier, and William D Ristenpart. 2020. Efficacy of masks and face coverings in controlling outward aerosol particle emission from expiratory activities. *Scientific reports*, 10(1):1–13.

Courtney Bir and Nicole Olynk Widmar. 2021. Societal values and mask usage for covid-19 control in the us. *Preventive Medicine*, 153:106784.

Cristina Bosco, Mirko Lai, Viviana Patti, Manuel Rangel Pardo Francisco, Rosso Paolo, et al. 2016. Tweeting in the debate about catalan elections. In *Proceedings of the Workshop on Emotion and Sentiment Analysis*, pages 67–70. European Language Resources Association (ELRA).

Ninghan Chen, Xihui Chen, and Jun Pang. 2022. A multilingual dataset of covid-19 vaccination attitudes on twitter. *Data in Brief*, 44:108503.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Ruth Garcia-Gavilanes, Daniele Quercia, and Alejandro Jaimes. 2013. Cultural dimensions in twitter: Time, individualism and power. In *Proceedings of the international AAAI conference on web and social media*, volume 7, pages 195–204.

Lu He, Changyang He, Tera L Reynolds, Qiushi Bai, Yicong Huang, Chen Li, Kai Zheng, and Yunan Chen. 2021. Why do people oppose mask wearing? a comprehensive analysis of us tweets during the covid-19 pandemic. *Journal of the American Medical Informatics Association*, 28(7):1564–1573.

Md Saroar Jahan. 2020. Team oulu at semeval-2020 task 12: Multilingual identification of offensive language, type and target of twitter post using translated datasets. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1628–1637.

Aditya Joshi, Pushpak Bhattacharyya, Mark Carman, Jaya Saraswati, and Rajita Shukla. 2016. How do cultural differences impact the quality of sarcasm annotation?: A case study of indian annotators and american text. In *Proceedings of the 10th SIGHUM*

Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities, pages 95–99.

Masaru Kamba, Wan Jou She, Kiki Ferawati, Shoko Wakamiya, and Eiji Aramaki. 2024. Exploring the impact of the covid-19 pandemic on twitter in japan: Qualitative analysis of disrupted plans and consequences. *JMIR infodemiology*, 4:e49699.

Ari Z Klein, Arjun Magge, Karen O'Connor, Jesus Ivan Flores Amaro, Davy Weissenbacher, and Graciela Gonzalez Hernandez. 2021. Toward using twitter for tracking covid-19: a natural language processing pipeline and exploratory data set. *Journal of medical Internet research*, 23(1):e25314.

KYODO. 2023. Japanese remain largely masked up on 1st day of eased covid rules.

Hocheol Lee, Eun Bi Noh, Sea Hwan Choi, Bo Zhao, and Eun Woo Nam. 2020. Determining public opinion of the covid-19 pandemic in south korea and japan: social network mining on twitter. *Healthcare informatics research*, 26(4):335.

Nancy HL Leung, Daniel KW Chu, Eunice YC Shiu, Kwok-Hung Chan, James J McDevitt, Benien JP Hau, Hui-Ling Yen, Yuguo Li, Dennis KM Ip, JS Peiris, et al. 2020. Respiratory virus shedding in exhaled breath and efficacy of face masks. *Nature medicine*, 26(5):676–680.

Xiaofeng Lin, Georgia Kernell, Tim Groeling, Jungseock Joo, Jun Luo, and Zachary C Steinert-Threlkeld. 2022. Mask images on twitter increase during covid-19 mandates, especially in republican counties. *Scientific Reports*, 12(1):21331.

Pintu Lohar, Haithem Afli, and Andy Way. 2017. Maintaining sentiment polarity in translation of user-generated content. *Prague Bulletin of Mathematical Linguistics*, (108):73–84.

Mary L McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282.

Kiyoshi Midooka. 1990. Characteristics of japanese-style communication. *Media, Culture & Society*, 12(4):477–489.

MIT. 2019. List of cities in japan.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. A dataset for detecting stance in tweets. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3945–3952.

Igor Mozetič, Miha Grčar, and Jasmina Smailović. 2016. Multilingual twitter sentiment classification: The role of human annotators. *PloS one*, 11(5):e0155036.

Deborah Netburn. 2021. A timeline of the cdc's advice on face masks.

Yasuko Obana and Michael Haugh. 2021. (non-) propositional irony in japanese–impoliteness behind honorifics. *Lingua*, 260:103119.

Bridianne O'dea, Stephen Wan, Philip J Batterham, Alison L Calear, Cecile Paris, and Helen Christensen. 2015. Detecting suicidality on twitter. *Internet Interventions*, 2(2):183–188.

Caleb Prichard and John Rucynski. 2022. L2 learners' ability to recognize ironic online comments and the effect of instruction. *System*, 105:102733.

Michael R Reich. 2020. Pandemic governance in japan and the united states: the control-tower metaphor. *Health Systems & Reform*, 6(1):e1829314.

simplemaps. 2022. United states cities database.

Kevin Stowe, Martha Palmer, Jennings Anderson, Marina Kogan, Leysia Palen, Kenneth M Anderson, Rebecca Morss, Julie Demuth, and Heather Lazrus. 2018. Developing and evaluating annotation procedures for twitter data during hazard events. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, pages 133–143.

Anawat Suppasri, Miwako Kitamura, Haruka Tsukuda, Sebastien P Boret, Gianluca Pescaroli, Yasuaki Onoda, Fumihiko Imamura, David Alexander, Natt Leelawat, et al. 2021. Perceptions of the covid-19 pandemic in japan with respect to cultural, information, disaster and social issues. *Progress in Disaster Science*, 10:100158.

Shota Suzuki, Ryohei Orihara, Yuichi Sei, Yasuyuki Tahara, and Akihiko Ohsuga. 2017. Sarcasm detection method to improve review analysis. In *ICAART (2)*, pages 519–526.

Cheryl Techentin, David R Cann, Melissa Lupton, and Derek Phung. 2021. Sarcasm detection in native english and english as a second language speakers. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, 75(2):133.

Ricky V Tso and Benjamin J Cowling. 2020. Importance of face masks for covid-19: A call for effective public education. *Clinical Infectious Diseases*, 71(16):2195–2198.

Nianyi Zeng, Zewen Li, Sherrianne Ng, Dingqiang Chen, and Hongwei Zhou. 2020. Epidemiology reveals mask wearing by the public is crucial for covid-19 control. *Medicine in Microecology*, 4:100015.

Elena Zotova, Rodrigo Agerri, Manuel Nuñez, and German Rigau. 2020. Multilingual stance detection in tweets: The Catalonia independence corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1368–1375, Marseille, France. European Language Resources Association.