Automatic Politeness Classification in Japanese Text

Shusuke Hashimoto
Indiana University Bloomington
Department of Linguistics

Abstract

This paper presents a sentence-level classification system for Japanese politeness, leveraging a transformer-based model fine-tuned on a dataset of 6,000 sentences evenly labeled as *polite*, *neutral*, or *impolite*. The dataset combines structured honorific annotations from the KeiCO Corpus with descriptive text from Wikipedia to capture a range of politeness phenomena. While the model achieves high classification accuracy (98%), qualitative analysis using attention heatmaps reveals a tendency to focus on surface-level polite endings (e.g., *masu*), while overlooking deeper pragmatic cues such as respectful verbs, humble forms, and beautification prefixes. These results highlight both the potential and the limitations of attention-based models in capturing the complexity of Japanese politeness, and point to the importance of dataset design and context diversity in future work on socially aware language systems. All code and resources are available at: https://github.com/shuhashi0352/Japanese-Politeness-Classification.

1 Introduction

Politeness is a fundamental aspect of Japanese communication, intricately woven into its grammatical system through honorifics, humble expressions, and polite verb endings. These linguistic forms serve to index social relationships, hierarchy, and formality, making Japanese a particularly rich but challenging language for automatic politeness classification.

While politeness classification has been explored in English and other languages, Japanese presents unique challenges due to the coexistence of multiple politeness strategies within a single sentence and the context-sensitive nature of their interpretation. A sentence may include respectful, humble, and beautification elements simultaneously, and their perceived politeness often depends on the social or pragmatic setting.

This study investigates sentence-level politeness classification in Japanese using a fine-tuned transformer-based model. We construct a dataset of 6,000 sentences evenly distributed across three labels—*polite*, *neutral*, and *impolite*—sourced from the KeiCO honorific corpus and Wikipedia. The model is evaluated using both accuracy metrics and attention heatmaps to examine whether it captures the nuanced cues of Japanese politeness.

Despite achieving strong accuracy, qualitative analysis reveals that the model often relies on surface-level polite forms while under-attending to more complex respectful or humble constructions. These findings raise important considerations about dataset balance, linguistic coverage, and the interpretability of attention-based classifiers for pragmatically rich tasks like politeness recognition in Japanese.

2 Related Work

Research on politeness in Japanese has evolved at the intersection of sociolinguistics, computational linguistics, and, more recently, deep learning. Politeness in Japanese is deeply embedded in its grammatical structure, especially through the use of honorifics, and varies significantly by gender, social hierarchy, and context, posing unique challenges for automatic classification systems.

2.1 Sociolinguistic Foundations

Early sociolinguistic studies argued that politeness in Japanese cannot be fully explained using Brown and Levinson's theory of face, which assumes universal pragmatic principles. Instead, Ide introduced the concept of "discernment" (wakimae), positing that Japanese politeness stems from social norms and roles rather than individual volition (Ide, 1982). Politeness also intersects with gender, as shown in research on women's language in Japanese, where specific particles and verb forms encode both identity and expected deference (Matsumoto, 1988).

Politeness interpretation in Japanese is also highly contextual and indexical. Yoshimi (Okamoto, 2021) demonstrates that honorifics can carry contradictory meanings depending on the social or pragmatic context—what appears polite may, under certain conditions, be interpreted as impolite or sarcastic.

2.2 Honorific Corpora and Formality Resources

To support computational modeling of politeness, Liu and Kobayashi developed the KeiCO corpus, a large-scale honorific dataset annotated using Systemic Functional Linguistics (SFL) (Liu and Kobayashi, 2022). The corpus includes 10,007 Japanese sentences annotated with four levels of honorifics (from highly formal to informal) and additional tags such as speaker-listener roles, respectful (尊敬語—*sonkei-go*), humble (謙譲語—*kenjou-go*), and polite (丁寧語—*teinei-go*) expressions. This resource enables more detailed classification and style transformation tasks in Japanese, where traditional corpora often overlook pragmatic context and honorific nuance.

2.3 Deep Learning and Response Generation

Politeness has also been modeled using deep learning for response generation. Resmi and Naseer (P and C, 2019) proposed a two-stage system: a BLSTM-based politeness classifier and an encoder-decoder model trained to generate contextually relevant polite responses. Their model achieved 87.5% accuracy in politeness classification and outperformed several baselines, demonstrating the viability of sequence-based approaches in polite dialogue generation.

Despite these advances, the lack of large-scale parallel datasets for style transfer remains a major limitation. Most Japanese politeness research must rely on weak supervision or external corpora annotated for other stylistic properties. The need for better alignment between linguistic style and neural representations continues to motivate corpus construction and model adaptation.

2.4 Large Language Models and Prompt Sensitivity

Recent studies have revealed that prompt politeness can significantly affect the performance of large language models (LLMs). Lin et al. (Yin et al., 2024) conducted a cross-lingual evaluation and found that overly polite prompts in Japanese often result in degraded task performance. LLMs tend to misinterpret excessive formality as uncertainty or vagueness, especially in zero-shot or instruction-based tasks.

These findings highlight the tension between naturalistic input and task clarity in multilingual LLMs. Similar issues arise in machine translation: models frequently fail to preserve the level of formality or politeness when translating between English and Japanese (Pituxcoosuvarn et al., 2024). Without explicit control mechanisms or context-aware decoding, outputs may violate socio-pragmatic norms critical in Japanese communication.

3 Methodology

3.1 Dataset

The dataset used in this study consists of 6,000 Japanese sentences, equally divided into three labels: *polite*, *neutral*, and *impolite*. Each label comprises 2,000 instances.

3.1.1 Annotation Criteria and Linguistic Basis

Each sentence in our dataset was manually assigned one of three labels: *polite*, *neutral*, or *impolite*, based on linguistic characteristics of Japanese honorifics and pragmatics. These labels reflect not only grammatical forms, but also the pragmatic force conveyed by the speaker, taking into account social expectations, situational appropriateness, and conventionalized register.

Polite: Sentences labeled as *polite* exhibit explicit use of honorific or deferential expressions. This includes:

- **Respectful (尊敬語)** forms used to elevate the subject's actions (e.g., "なさる" (*nasaru*), "いらっしゃる" (*irassharu*)).
- **Humble (謙譲語)** forms where the speaker lowers their own actions to respect the listener (e.g., "伺う" (*ukagau*), "申す" (*mousu*)).
- **Polite (丁寧語)** forms such as the "です・ます" (*desu・masu*) endings, which add politeness regardless of hierarchy.
- **Word beautification (美化語 - *bika-go*)** forms, which involve the use of beautifying

prefixes like “お” (*o*) or “ご” (*go*) (e.g., “お料理” (*o-ryouri*), “ご案内” (*go-annai*)) to make expressions more elegant or refined. While word beautification forms may not directly encode hierarchy, they contribute to a softened, polite register often used in customer service, announcements, and everyday politeness.

Importantly, sentences may mix these types; e.g., a sentence might contain a respectful verb and polite endings simultaneously. Therefore, rather than separating by honorific type, we grouped all such sentences under the *polite* label, as they functionally signal deference or formality at the sentence level.

Neutral: Sentences labeled as *neutral* are marked by the absence of honorifics or informal speech. These typically exhibit plain (non-polite) verb endings, such as the short-form dictionary form (“行く” (*iku*), “する” (*suru*)), and do not signal any overt social hierarchy. Such sentences are commonly found in encyclopedic entries, news reports, or factual statements where politeness is not pragmatically required. The lack of indexical markers of status or emotion places these utterances in a neutral register.

Impolite: The *impolite* label was applied to sentences that include:

- Informal or blunt verb endings in contexts where politeness is normally expected (e.g., using “食え” (*kue*) instead of “召し上がってください” (*meshiaga-tte-kudasai*)).
- Casual contractions or slang (e.g., “だよね” (*dayone*), “じゃん” (*jan*)).
- Omission of honorific markers in socially inappropriate contexts.

While such sentences may not always be intentionally rude, they lack the grammatical and pragmatic markers that signal consideration for the hearer, especially in formal or hierarchical situations. In Japanese, this absence can itself be interpreted as impolite or overly casual, depending on the social context.

Contextual Considerations: All labels were assigned with attention to likely speaker–listener relationships implied in the text, even if the direct context was limited. Because the task is sentence-level classification, annotations were based on the assumption of a generic social interaction, defaulting to business or formal interactions when judging borderline cases. When a sentence included mixed or ambiguous cues, the dominant pragmatic effect was used as the basis for labeling.

3.1.2 Data Sources

The **polite** and **impolite** data were extracted from the KeiCO Corpus (Liu and Kobayashi, 2022), a dataset of Japanese honorific expressions constructed using Systemic Functional Linguistics. In this corpus, every sentence is assigned one of four levels of honorific usage: Level 1 (The Highest Honorific Level), Level 2 (Secondary Honorific Level), Level 3 (Third Honorific Level), and Level 4 (No Honorifics Used). For the *polite* label, we collected 2,000 instances from sentences labeled as Levels 1 through 3. The *impolite* class consists of 2,000 instances from sentences labeled as Level 4 in the KeiCO Corpus.

The **neutral** data were independently collected from Japanese Wikipedia. We randomly sampled 2,000 sentences spanning multiple genres. While topic diversity cannot be guaranteed due to domain limitations, care was taken to avoid topic bias by sampling from a wide range of article categories.

3.2 Preprocessing

All sentences were tokenised using `fugashi`, a MeCab wrapper for Japanese morphological analysis. Stratified sampling was applied to ensure label balance across training (80%), validation (10%), and test (10%) splits.

3.3 Model and Training

To perform politeness classification, we fine-tuned the LINE-distilled Japanese BERT model[1]. This model was selected for its efficiency and robust performance on sentence-level classification tasks in Japanese. Fine-tuning was conducted using the following hyperparameters:

3.4 Evaluation

Quantitative Evaluation: Model performance was evaluated using **accuracy** as the primary metric. Since the dataset is evenly balanced across the three classes—*polite*, *neutral*, and *impolite*—

[1] `https://huggingface.co/line-corporation/line-distilbert-base-japanese`

Hyperparameter	Value
Number of epochs	3
Batch size	16
Optimizer	AdamW
Learning rate	2e–5
Weight decay	0.01
Warmup steps	500
Max sequence length	128

Table 1: Model training hyperparameters

accuracy is an appropriate and interpretable measure for assessing classification performance in this task.

Qualitative Evaluation through Attention Analysis: To gain insight into the internal behavior of the model, we conducted an attention-based qualitative analysis. Specifically, attention heat maps were generated using the attention weights from the **last layer of the first attention head**. Rather than analyzing which tokens attend to others, we focused on the attention distribution received by each token from the [CLS] token, which serves as the aggregate representation used for classification.

This approach allows us to identify which parts of the input sentence contributed most to the model's final prediction, as inferred from the attention allocation of the [CLS] vector. By visualizing these attention weights, we can examine whether the model prioritizes linguistically meaningful components—such as honorific verbs, polite endings, humble expressions, or beautification prefixes—when determining politeness level.

4 Results

4.1 Quantitative Evaluation

The model's performance was assessed on the training, validation, and test sets using accuracy and loss metrics. The results are summarized in Table 2.

Metric	Score
Final training loss	0.0140
Validation accuracy	0.9817
Test accuracy	0.9800

Table 2: Model performance across training, validation, and test sets.

The model achieves a high test accuracy of 0.9800, demonstrating strong generalization to unseen examples. This suggests that it successfully captures surface-level patterns associated with different politeness categories in Japanese.

4.2 Qualitative Evaluation through Attention Analysis

To complement the quantitative metrics, we analyzed the model's attention behavior by visualizing heat maps from the last layer's first attention head. These maps show which tokens receive the most attention from the [CLS] token, which is used as the aggregate representation for sentence classification. This analysis helps determine whether the model is internally attending to linguistically relevant cues of politeness, such as polite endings, honorific verbs, humble forms, or beautification prefixes.

Seven manually constructed sentences were used to isolate key politeness phenomena, as shown in Figures 1 to 7. The corresponding attention heatmaps are provided in Appendix. Despite the model's high classification accuracy, the attention behavior suggests that it does not always align with linguistically meaningful tokens.

Figure 1 shows a neutral sentence ending with the plain-form verb "する" (*suru*). The [CLS] token barely attends to this verb, despite its relevance to determining formality. In Figure 2, the impolite noun "奴" (*yatsu*)—a casual or rude way of referring to others—receives little attention, even though it strongly indicates impoliteness.

In Figure 3, the sentence ends with the polite verb "ます" (*masu*), and the model does attend to it to some extent, though not predominantly. Figures 4, 5, and 6 show polite expressions using inflected respectful and humble forms—"ませ" (*mase*), "いらっしゃい" (*irasshai*), and "申し上げ" (*moushiage*) respectively. In all three cases, the model's attention tends to favor "ます" (*masu*) over the more nuanced polite markers, potentially due to the embedded nature of these forms in the final verb phrase (e.g., "いらっしゃいます" —*irasshai-masu*).

Figure 7 contains the beautification prefix "お" —*o* (as in "お手洗い" (*o-tearai*)), a common polite construction. The model fails to attend to this prefix altogether, suggesting that beautification forms are not adequately recognized by the model's internal mechanism.

These findings indicate that while the model performs well on classification tasks, its attention behavior lacks sensitivity to the full range of Japanese politeness strategies. In particular, it over-relies on surface-level polite endings such as *masu*, while underutilizing respectful, humble, and beautification forms. It is worth noting that in some domains, the interpretation of forms such as *irasshai* or *moushiage* can be ambiguous when inflected; however, this project adopts a contextual view of politeness that treats these forms as polite when used functionally within an utterance.

5 Discussion

The experimental results highlight a notable gap between the model's strong quantitative performance and its internal interpretability. With over 98% accuracy on the test set, the model demonstrates a strong ability to generalize across the three politeness classes, likely due to the consistent presence of surface-level markers such as polite endings (e.g., *masu*, *desu*).

However, the attention heatmaps reveal that the model does not consistently attend to deeper linguistic indicators of politeness. Expressions involving respectful verbs (e.g., *irasshai*), humble forms (e.g., *moushiage*), and beautification prefixes (e.g., *o* in *o-tearai*) often receive little to no attention from the `[CLS]` token, even when they are semantically and pragmatically central to the sentence's politeness level. This suggests that while the model can accurately predict labels, it may do so by relying on shallow statistical regularities rather than nuanced syntactic or morphological cues.

The handling of inflected forms such as *irasshai-masu* or *moushiage-masu* reveals an additional challenge. These expressions integrate respectful or humble roots with polite verb endings, which may obscure their internal structure from the model's view. This complexity complicates classification and suggests that future work should consider more sophisticated representations of Japanese honorific morphology.

These findings suggest that the interpretability limitations observed in the attention analysis may be less about the inadequacy of attention mechanisms themselves, and more reflective of the characteristics of the training data. In particular, the polite class combined a wide range of expressions —including respectful, humble, and beautification forms—without ensuring an even or representative distribution. As a result, the model may have disproportionately learned to rely on more frequent and surface-level features such as polite verb endings (e.g., *masu*) while under-attending to less common but semantically significant markers.

Similarly, the neutral class consisted exclusively of sentences drawn from Wikipedia, which typically exhibit formal but descriptively oriented language. This narrow domain coverage may have constrained the model's understanding of neutrality in more varied pragmatic settings. Broadening the data sources and increasing diversity in expression types could help the model develop a more nuanced understanding of contextual politeness beyond overt morphological cues.

6 Future Work

While this study focuses on sentence-level politeness classification, future work should explore the integration of this capability into dialogue systems, particularly chatbots designed for Japanese users. In human conversation, especially in Japanese, politeness is not solely governed by the level of formality used by the interlocutor. It can also shift dynamically based on the emotional state or situational context.

For example, a user may use casual or even impolite language due to frustration or distress. In such cases, a chatbot that merely mirrors the user's level of formality may come across as inappropriate or even insensitive. Instead, it may be more desirable for the chatbot to maintain a respectful tone —responding in polite or neutral forms regardless of the user's speech register—to preserve a sense of professionalism or emotional safety.

This highlights the importance of combining politeness classification with sentiment analysis. A chatbot equipped with both capabilities could infer not only how something is said (politeness) but also why it is said that way (sentiment or emotion). This would allow the system to adapt its speech strategy accordingly, effectively "sensing the atmosphere"and choosing an appropriate register that reflects both social norms and the user's emotional state.

Developing such dual-sensitive systems could enhance the naturalness and social appropriateness of chatbot interactions in Japanese, and potentially extend to other languages with rich politeness systems.

7 Conclusion

This study presents a sentence-level classification system for Japanese politeness using a fine-tuned transformer model and a balanced dataset of polite, neutral, and impolite utterances. By leveraging the KeiCO Corpus and additional Wikipedia data, we constructed a 6,000-sentence dataset and evaluated model performance both quantitatively and qualitatively.

The classifier achieved high accuracy, demonstrating its effectiveness in capturing surface-level indicators of politeness. However, attention-based analysis revealed that the model tends to focus on overt polite forms such as "ます" (*masu*) and "です" (*desu*), while underutilizing respectful, humble, and beautification expressions. These findings point to limitations in how the model internalizes and interprets the rich morphological and pragmatic structures of Japanese politeness.

We also discussed the challenges posed by imbalanced representation of honorific forms in the dataset and the restricted domain of neutral examples. These factors may have contributed to the model's shallow interpretability despite strong predictive accuracy.

Ultimately, this work highlights both the feasibility and challenges of politeness classification in Japanese, offering insights into data design, model interpretation, and the linguistic subtleties of honorifics. The results lay the groundwork for future developments in adaptive dialogue systems, particularly in combining politeness with sentiment understanding to support context-aware and socially appropriate language generation.

References

Sachiko Ide. 1982. Japanese sociolinguistics politeness and women's language. *Lingua*, 57(2):357–385.

Muxuan Liu and Ichiro Kobayashi. 2022. Construction and validation of a Japanese honorific corpus based on systemic functional linguistics. In *Proceedings of the Workshop on Dataset Creation for Lower-Resourced Languages within the 13th Language Resources and Evaluation Conference*, pages 19–26, Marseille, France. European Language Resources Association.

Yoshiko Matsumoto. 1988. Reexamination of the universality of face: Politeness phenomena in japanese. *Journal of Pragmatics*, 12(4):403–426.

Shigeko Okamoto. 2021. Your politeness is my impoliteness. *East Asian Pragmatics*, 6(1):39–64.

Resmi P and Naseer C. 2019. A deep learning approach for polite dialogue response generation. In *Proceedings of the International Conference on Systems, Energy & Environment (ICSEE)*, GCE Kannur, Kerala. Available at SSRN: https://ssrn.com/abstract=3438132 or http://dx.doi.org/10.2139/ssrn.3438132.

Mondheera Pituxcoosuvarn, Wuttichai Vijitkunsawat, and Yohei Murakami. 2024. Addressing sociolinguistic challenges in machine translation: An llm-based approach for politeness and formality. In *2024 8th International Conference on Information Technology (InCIT)*, pages 757–762.

Ziqi Yin, Hao Wang, Kaito Horio, Daisuke Kawahara, and Satoshi Sekine. 2024. Should we respect llms? a cross-lingual study on the influence of prompt politeness on llm performance. *Preprint*, arXiv:2402.14531.

Appendix

This appendix presents the attention heatmaps corresponding to the qualitative analysis discussed in Section 4. Each heatmap visualizes the attention distribution from the [CLS] token to all input tokens, based on the last layer's first attention head. The manually constructed example sentences were designed to isolate specific politeness-related expressions such as plain forms, respectful verbs, humble verbs, polite endings, and beautification prefixes. These visualizations serve to clarify which tokens the model attends to when making sentence-level politeness classifications.

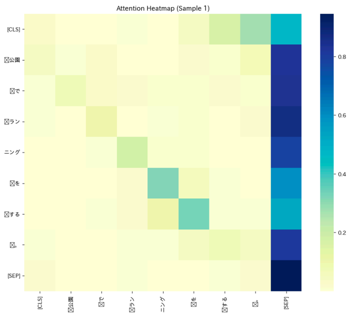

Figure 1: Attention heatmap for neutral sentence: 公園でランニングをする。

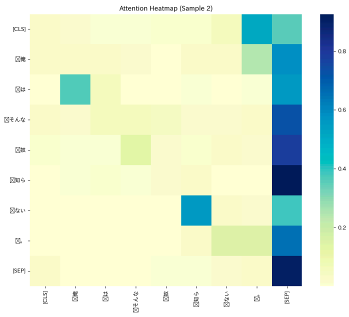

Figure 2: Attention heatmap for impolite sentence: 俺はそんな奴知らない。

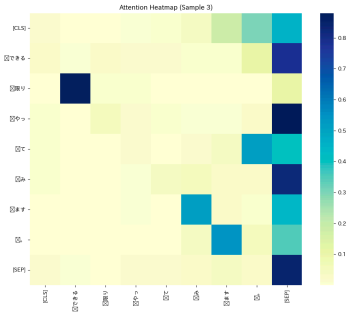

Figure 3: Attention heatmap for polite sentence: できる限りやってみます。

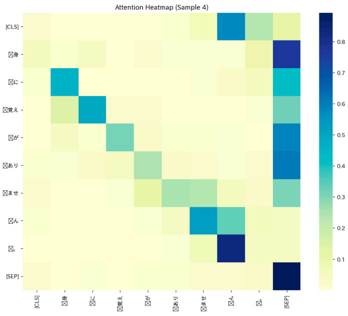

Figure 4: Attention heatmap for inflected polite form: 身に覚えがありません。

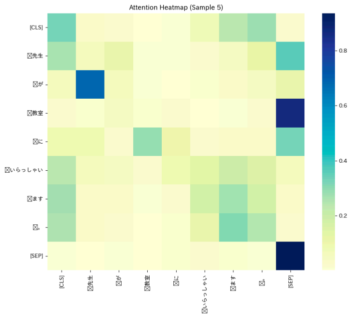

Figure 5: Attention heatmap for respectful form: 先生が教室にいらっしゃいます。

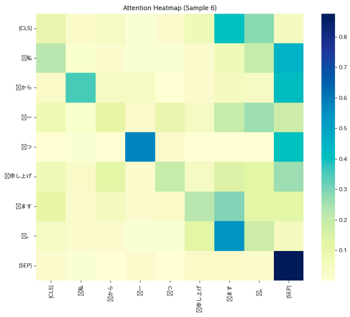

Figure 6: Attention heatmap for humble form: 私から一つ申し上げます。

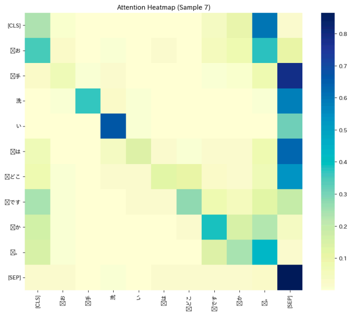

Figure 7: Attention heatmap for beautification prefix: お手洗いはどこですか。