Do Language Models Understand Honorific Systems in Javanese?

Mohammad Rifqi Farhansyah[α β *] **Iwan Darmawan**[β *] **Adryan Kusumawardhana**[β]
Genta Indra Winata[γ † ‡] **Alham Fikri Aji**[δ ‡] **Derry Tanti Wijaya**[β ζ ‡]
[α]Institut Teknologi Bandung [β]Monash Indonesia [ζ]Boston University [γ]Capital One [δ]MBZUAI
{mrifqifarhansyah, adryan.kusumawardhana}@gmail.com idar0006@student.monash.edu
genta.winata@capitalone.com alham.fikri@mbzuai.ac.ae
derry.wijaya@monash.edu

Abstract

The Javanese language features a complex system of honorifics that vary according to the social status of the speaker, listener, and referent. Despite its cultural and linguistic significance, there has been limited progress in developing a comprehensive corpus to capture these variations for natural language processing (NLP) tasks. In this paper, we present UNGGAH-UNGGUH[1], a carefully curated dataset designed to encapsulate the nuances of *Unggah-Ungguh Basa*, the Javanese speech etiquette framework that dictates the choice of words and phrases based on social hierarchy and context. Using UNGGAH-UNGGUH, we assess the ability of language models (LMs) to process various levels of Javanese honorifics through classification and machine translation tasks. To further evaluate cross-lingual LMs, we conduct machine translation experiments between Javanese (at specific honorific levels) and Indonesian. Additionally, we explore whether LMs can generate contextually appropriate Javanese honorifics in conversation tasks, where the honorific usage should align with the social role and contextual cues. Our findings indicate that current LMs struggle with most honorific levels, exhibiting a bias toward certain honorific tiers.

Ngoko	Mbak Srini *mangan* pecel *ajange* pincuk.
Ngoko Alus	Mbak Srini *dhahar* pecel *ambenge* pincuk.
Krama	Mbak Srini *nedha* pecel *ajangipun* pincuk.
Krama Alus	Mbak Srini *dhahar* pecel *ambengipun* pincuk.
Translation 🇮🇩	Mbak Srini *makan* pecel *dengan ajang* pincuk.
Translation 🇬🇧	Mbak Srini *ate* pecel *with* pincuk.
Ngoko	Bu Haiti *arep tuku* gula selawe kilo, *wis dak-dolane*.
Ngoko Alus	Bu Haiti *arep mundhut* gula selawe kilo, *wis adalem dolane*
Krama	Bu Haiti *badhe tumbas* gendhis selangkung kilo, *sampun kula sadeanipun*.
Krama Alus	Bu Haiti *badhe mundhut* gendhis selangkung kilo, *sampun adalem sadeanipun*.
Translation 🇮🇩	Bu Haiti *akan membeli* gula dua puluh lima kilo, *sudah saya jualnya*.
Translation 🇬🇧	Mrs. Haiti *is going to buy* twenty-five kilos of sugar, *I have already sold it*.

Table 1: UNGGAH-UNGGUH examples with four different honorific levels from the most informal (Ngoko) to the most refined (Krama Alus). The *italicized text* highlights the phrases that vary according to the honorific level. More examples can be found in Appendix B.1.

1 Introduction

In many languages, honorifics play a crucial role in addressing others and maintaining social relationships (Agha, 1994). Honorific registers are formally discrete yet functionally stratified systems, meaning that a seemingly fixed set of linguistic forms enables speakers to navigate multiple aspects of the pragmatic context in which language is used (Agha, 1998). Using the appropriate honorific registers in the right situations fosters culturally appropriate conversations and enhances the overall propriety of interactions. One language that exemplifies a highly complex honorific system is Javanese, where honorific usage is deeply intertwined with cultural norms and historical traditions. In Javanese, the use of honorific language is inseparable from the broader concept of linguistic politeness (Rahayu, 2014).

The Javanese language, spoken by over 98 million people,[2] features a distinctive honorific system known as *Unggah-Ungguh Basa*[3], which is essential for conveying respect, social hierarchy, and formality in conversations. The four primary levels of Javanese honorifics (Table 1)—Ngoko, Ngoko Alus, Krama, and Krama Alus—each represent different degrees of formality and respect, playing a crucial role in facilitating appropriate so-

[*]Equal contribution.
[†]The work was conducted outside Capital One.
[‡]Senior authors.
[1]We released our models & dataset in https://huggingface.co/JavaneseHonorifics

[2]https://www.langcen.cam.ac.uk/resources/langj/javanese.html.
[3]More details about Unggah-Ungguh Basa can be seen in Appendix 6

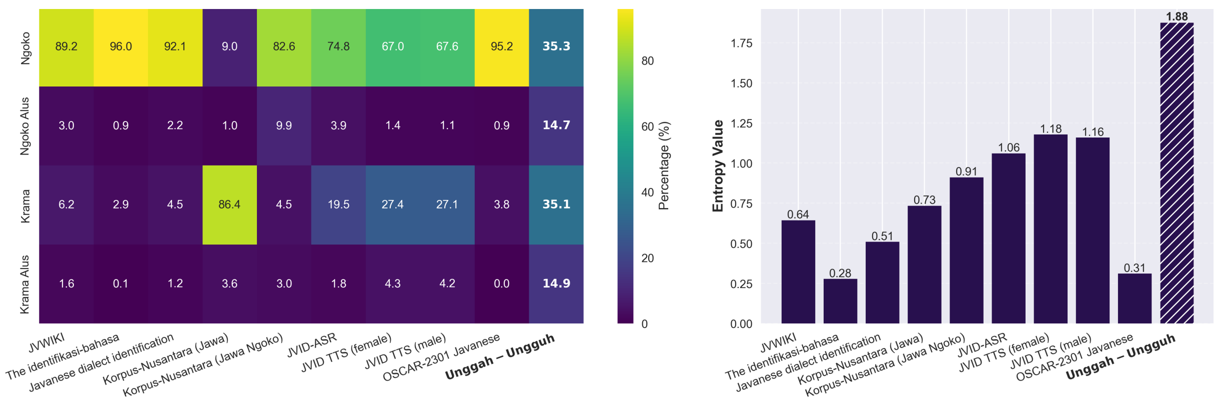

Figure 1: A heatmap **(left)** illustrates the distribution (%) of Javanese honorific levels across various corpora including ours (Unggah-Ungguh). An accompanying entropy bar chart **(right)** quantifies the variability within each corpus using Shannon's Entropy. A higher entropy value indicates a more diverse and well-balanced corpus.

cial interactions within Javanese society. Despite their cultural significance, there is a notable lack of comprehensive linguistic resources that accurately capture these distinctions for natural language processing (NLP) applications.

Current research indicates that existing models struggle to accurately interpret and generate Javanese honorifics due to the absence of a well-annotated corpus (Marreddy et al., 2022), which in turn hinders the development of effective NLP tools capable of handling their complexity. Moreover, as illustrated in Figure 1, most existing Javanese corpora exhibit an imbalanced distribution of honorific levels, further limiting model performance. Addressing this issue is critical, particularly as language models increasingly serve as personal assistants across various domains, adapting to user expectations (Kong et al., 2024). Given that honorifics shape social interactions, an LM's ability to process and generate them appropriately is essential for maintaining perceived status and formality, fostering trust and likability (Jeon, 2022), and ultimately enabling more natural and culturally sensitive human-AI interactions.

In this paper, we aim to bridge the existing gap by systematically evaluating the ability of LMs to comprehend and generate appropriate Javanese honorifics. We conduct a series of evaluations to identify potential biases toward specific honorific levels, assess cross-lingual performance in handling Javanese honorifics, and analyze models' ability to generate conversational text with contextually appropriate honorific usage based on the speaker's role and context. Our contributions are:

- We introduce UNGGAH-UNGGUH, the first multi-cultural honorific corpus for the Javanese language. Each sentence in this corpus is annotated with one of the four honorific levels.
- We leverage this corpus for four downstream NLP tasks: honorific-level classification, honorific style change, cross-lingual honorific translation, and conversation generation with honorific personas.
- We explore a diverse set of LMs to address these tasks, including English-centric, multilingual, and Southeast Asian regional models, as well as models trained on Indonesian and its regional languages. Our approaches include encoder-based models, generative models, and rule-based methods.

Beyond advancing NLP research, this corpus serves as a vital resource and an important first step toward accurately digitizing the nuanced aspects of Javanese cultural heritage. By making it available, we aim to support the development of more accurate and culturally sensitive NLP models for the Javanese language while also encouraging future research on other low-resource languages with similarly complex sociolinguistic structures.

2 Related Work

Research on low-resource languages, particularly Javanese, has made significant progress, making it highly relevant to this study. In the domain of language-specific modeling and processing, work on Javanese language modeling (Wongso et al., 2021a; Cahyawijaya et al., 2024; Owen et al., 2024), machine translation (Sujaini, 2020; Wibawa et al., 2013a,b), part-of-speech tagging (Pratama et al., 2020; Ramadhan et al., 2020; Noor et al., 2020; Enrique et al., 2024), sentiment analy-

Source	Title	Author/Editor	Publication Year	Publisher
Dictionary	Kamus Unggah-Ungguh Basa Jawa	Drs. Haryana Harjawiyana, S.U. & Drs. Th. Supriya	2009	PT Kanisius
Book	Unggah-ungguh Basa Jawa	Umi Kuntari, S.S.	2017	Pustaka Widyatama
	Mempelajari Unggah-Ungguh Bahasa Jawa	Drs. Imam Riyadi, M.KPd.	2019	Penerbit PARAMARTA Trenggalek
	Unggah-ungguh Basa	H. Soemardi	2016	CV Satubuku

Table 2: List of all data sources used to construct UNGGAH-UNGGUH.

sis (Winata et al., 2023; Lucky et al., 2023), and dialect identification (Hidayatullah et al., 2020; Filby et al., 2024) have addressed some of the inherent complexities in the Javanese language. Furthermore, research on Javanese dependency parsing (Ghiffari et al., 2023) has further enhanced our understanding of Javanese linguistic structures. These studies provide a foundation for developing specialized corpora that capture Javanese honorifics.

Dataset creation and curation play a crucial role in advancing research on low-resource languages. Notable efforts in Javanese include Wibawa et al. (2018), who developed Javanese and Sundanese corpora through community collaboration, leveraging local expertise to build comprehensive datasets. The Japanese Honorific Corpus by Liu and Kobayashi (2022) provides a valuable comparative framework for integrating honorific distinctions into linguistic corpora, offering insights applicable to the Javanese context.

Ensuring data quality and rigorous evaluation is essential, particularly for developing reliable and culturally relevant corpora. Previous studies on human evaluation of web-crawled corpora (Ramírez-Sánchez et al., 2022) and dataset audits (Kreutzer et al., 2022) emphasize the importance of stringent quality control measures in corpus development. Collectively, these studies inform our approach to constructing a Javanese honorific corpus, highlighting the need for culturally informed data collection, meticulous annotation, and robust evaluation methods. Drawing from these insights, the development of a comprehensive corpus for the Javanese honorific system based on *Unggah-Ungguh Basa* can significantly enhance the accuracy and cultural relevance of NLP applications in Javanese.

3 UNGGAH-UNGGUH Corpus

The UNGGAH-UNGGUH Corpus is a meticulously curated dataset focused on capturing the nuanced use of honorific language in Javanese.[4] The dataset ensures comprehensive coverage of honorific usage by drawing from a diverse range of reputable sources, including dictionaries and books. This diverse sourcing results in a context-rich collection.

3.1 Honorific System

The Javanese language features a distinctive system of speech levels known as *undha usuk*, which reflects the social hierarchy and relationships between speakers. Traditionally, this system was divided into various levels; however, in modern times, it has been streamlined into four primary levels: *Ngoko*, *Ngoko Alus*, *Krama*, and *Krama Alus*. Each level possesses unique linguistic characteristics and specific contexts of use, making them essential for accurate annotation and analysis in the development of a Javanese honorific corpus.

Ngoko. Ngoko is the most informal speech level, used in everyday communication among equals or when addressing someone of lower social status. It consists entirely of ngoko words.

Ngoko Alus. Ngoko Alus is a refined version of ngoko, incorporating polite or respectful words. It is typically used when speaking to someone of slightly higher status or to show a degree of respect while maintaining the informal structure of ngoko.

Krama. Krama is a more formal speech level, used when addressing someone of higher social status or someone unfamiliar. It is characterized by the use of distinct krama vocabulary, which differs significantly from ngoko.

Krama Alus. Krama Alus is the most refined and polite speech level, reserved for interactions with individuals of significantly higher social status or in very formal settings. It involves the use of additional honorific words and phrases that further elevate the politeness of speech.

3.2 Corpus Creation

Table 2 summarizes the main sources used to construct the Javanese Honorific Corpus. The dataset primarily originates from Kamus Unggah-Ungguh Basa Jawa (Harjawiyana et al., 2001), a dictionary

[4] Our dataset will be licensed under CC-BY-NC 4.0 for research purposes only.

Text Label	# Sentences	Avg. Sentence Length	# Word Tokens	# Word Types	Yule's characteristic K
Ngoko	1,419	9.26	13,142	3,486	118.80
Ngoko Alus	590	10.07	5,944	1,527	108.13
Krama	1,414	9.60	13,572	3,280	124.61
Krama Alus	601	10.13	6,088	1,530	115.14
Overall	4,024	9.63	38,746	6,156	105.43

Table 3: Data Statistics of UNGGAH-UNGGUH.

containing example sentences labeled by honorific level. Since the source is not digitized, we scanned relevant pages, applied OCR, and manually corrected errors through a two-stage verification process involving native Javanese speakers.

In the first stage, one of the authors (native Javanese speaker) corrected OCR errors by comparing outputs with the original text. In the second stage, another native speaker independently reviewed all sentences and transcribed Indonesian translations, identifying and fixing 58 errors (1.5%) out of 4,024 sentences. Additional example sentences were sourced from other Javanese language references and manually verified to ensure accuracy and relevance in representing Javanese honorifics.

3.3 Data Statistics

Table 3 provides a statistical overview of the Javanese honorific corpus, comprising 4,024 sentences across four honorific levels: Ngoko (1,419 sentences), Ngoko Alus (590), Krama (1,414), and Krama Alus (601). The UNGGAH-UNGGUH corpus shows significant variability and a relatively balanced distribution compared to other datasets, as indicated by a Shannon entropy[1] value of 1.88, surpassing nine other datasets, as illustrated in Figure 1. This result highlights the corpus's balanced diversity across different honorific levels, making it a valuable resource for studying linguistic diversity in Javanese.

The average sentence length varies slightly across honorific levels, ranging from 9.26 to 10.13 words, reflecting the increased linguistic complexity at higher levels. The corpus contains 38,746 word tokens and 6,156 unique word types, with Yule's characteristic K[2] value (Yule, 1944) calculated as 105.43.

Yule's K measures word repetition, with higher values indicating more frequent repetition (lower lexical diversity) and lower values reflecting less repetition (higher lexical diversity). In this study, our Javanese Honorific Corpus has a Yule's K value of 105.43, compared to 125.54 in the Japanese Honorific Corpus (Liu and Kobayashi, 2022), indicating that the Javanese Honorific Corpus exhibits greater lexical diversity than its Japanese counterpart.

3.4 Tasks

In our benchmark, we want to train and evaluate LMs' understanding to various language styles in different honorific levels. Our benchmark comprises four downstream NLP tasks:

Task 1: Honorific Level Classification. The task is to classify a text into one of the honorific levels. Given an input text x, we intend to use some LMs θ to map x to an honorific-level label $\hat{h}$ from four honorific levels. This evaluation is crucial for assessing the capability of LMs in recognizing specific honorific level within Javanese sentence.

Task 2: Honorific Style Change. The task is to style translate from a given text x from a source honorific style h_{src} to a target honorific style h_{tgt} in Javanese, resulting a translated text $\hat{y}$. This task is particularly important for evaluating model's capability to capture nuanced sociolinguistic conventions and generate contextually appropriate text across different levels of politeness and respect.

Task 3: Cross-lingual Honorific Translation. The task is to translate a given sentence x from a specific honorific level h to Indonesian (X $\rightarrow$ ID), $\hat{y}$, and vice versa (ID $\rightarrow$ X). This task is crucial for evaluating how well LMs handle cross-lingual translation between languages with asymmetrical honorific systems (e.g., Javanese (rich honorific system) and Indonesian (lacks explicit honorifics). Furthermore, KL Divergence[3](Kullback and Leibler, 1951) and Jensen Score[4](Lin, 2002) values computed from UNGGAH-UNGGUH (Table 4) between Indonesian and Javanese indicate a substantial overall divergence, with scores reaching 2.26 and 0.34 respectively. Higher divergence values indicate greater lexical shifts during cross-lingual translation, with Jensen Score reflecting mean distributional distance and KL Divergence

	Ngoko-Indonesia	Ngoko Alus-Indonesia	Krama-Indonesia	Krama Alus-Indonesia	**Overall**
KL-Divergence	1.481	1.29	1.72	1.38	2.26
Jensen Score	0.25 ± 0.00007	0.23 ± 0.00014	0.28 ± 0.00007	0.24 ± 0.00014	0.34 ± 0.00005

Table 4: KL Divergence and Jensen Score between Javanese honorific-level distributions and their Indonesian translations in UNGGAH-UNGGUH.

capturing asymmetrical divergence between token distributions. These values suggest significant lexical shifts between Javanese and Indonesian. Consequently, the cross-lingual honorific translation task becomes particularly challenging, as models must effectively bridge this lexical gap while preserving both semantic content and sociolinguistic nuance.

Task 4: Conversation Generation. In this task, we aim to simulate and synthesize conversations between two speakers to assess whether LMs can generate appropriate honorific-specific language based on the speakers' social status. Given the social statuses (e.g., student and teacher) of speakers A and B, respectively, along with a context c, the objective is to generate a conversation $\hat{y}$ featuring one utterance per speaker. Each utterance should correspond to the appropriate honorific level dictated by h_A and h_B, while maintaining coherence with c. For instance, in a conversation between a student and a teacher, the student is expected to speak in a more formal register (Krama Alus), whereas the teacher might use a more casual style (Ngoko). We manually curate 160 different evaluation scenarios, detailed in Appendix B.2. Evaluating this task is crucial for measuring cultural appropriateness in scenarios involving Javanese speakers.

4 Experimental Setup

We describe the experimental[5] details to train and evaluate models for classifying Javanese honorific levels, as well as the experiments conducted for the four honorific-related tasks. Our approach involves using encoder LMs, generative models, and a rule-based model. We utilize two categories of models: *Fine-tuned Models* and *Off-the-Shelf Models*. Fine-tuned models are specifically employed for Task 1, Honorific Level Classification, with the best-performing classifier subsequently used to evaluate the accuracy of the generated text's honorific level in Task 4, Conversation Generation. Details regarding computational resources and model hyperparameters are provided in Appendix F.

[5]Our code are available at `https://github.com/JavaneseHonorifics/Unggah-Ungguh`

4.1 Fine-tuned Models

We fine-tune encoder-based and decoder-based Javanese LMs to classify Javanese sentences into specific honorific levels. Specifically, we fine-tune two encoder-based LMs on UNGGAH-UNGGUH: Javanese BERT and Javanese DistilBERT, both of which are pre-trained on Javanese text corpora (Wongso et al., 2021b). Additionally, we fine-tune a decoder-based LM, Javanese GPT-2, originally designed for movie classification using Javanese IMDB reviews.

To establish baselines, we fine-tune an LSTM model and develop a rule-based classifier. The rule-based model is implemented using a dictionary-based approach, mapping word types to their corresponding honorific levels based on predefined linguistic rules, as outlined in Algorithm 1. The dictionary used in this model is derived from the "Indonesian Javanese Dictionary Starter Kit" (Rahutomo et al., 2018), which categorizes Javanese words into three types: *Ngoko*, *Krama Alus*, and *Krama Inggil*. This dataset enables the mapping of Javanese word types to their respective honorific levels. For fine-tuning, we use 80% of our UNGGAH-UNGGUH for training, while the remaining 20% is reserved for evaluating the honorific level classification models and selecting the best-performing classifier.

4.2 Off-the-Shelf Models

We evaluate various models across four tasks, encompassing both closed-source and open-source options, and categorize them into: English-centric models, multilingual models, Southeast Asian regional models, and models fine-tuned specifically for Indonesian and local languages (eg. Javanese, Sundanese, etc.).

For closed-source models, we utilize GPT-4o (Achiam et al., 2023) and Gemini 1.5 Pro (Team et al., 2024), noted for their outstanding performance and status as state-of-the-art language models. In the open-source category, we employ Gemma 2 9B Instruct (Team et al., 2024) as English-centric model and Llama 3.1 8B Instruct (Dubey et al., 2024) as multilingual model. These models were chosen for their high performance and serve as benchmarks for the SahabatAI [6] [7] models, which are designed for Indonesian contexts. Additionally, we include the Southeast Asian regional model, Sailor2 8B (Dou et al., 2024), as a comparative baseline.

Task Setup. For all experiments, we run once for each task and setting. We describe the experimental setting for each task as following:

1. **Zero-shot classification.** We classify Javanese sentences into specific honorific levels using zero-shot approach. This method is applied to Task (1) and compared against the fine-tuned models.
2. **Zero-shot translation.** In Tasks (2) and (3), we use off-the-shelf models to perform translation between honorific levels using a zero-shot approach.
3. **One-shot generation.** For Task (4), Conversation Generation, we employ a one-shot approach in which off-the-shelf models are provided with a sample conversation as a formatting guide. Furthermore, we conduct experiments with explicit hints about Javanese honorific usage to assess whether such hints can enhance LM performance (refer to Figure 7 in Appendix C).

The prompts are detailed in Appendix C.

4.3 Evaluation Metrics

All models are evaluated using distinct metrics tailored to the nature of each task. For classification-related tasks, we employ accuracy (Acc.), precision (Prec.), recall (Rec.), and F1-score (F_1) to comprehensively assess the models' ability to correctly classify honorific levels. For translation tasks (Task 2 and Task 3), evaluation is conducted using BLEU (Papineni et al., 2002) and CHRF++ (Popović, 2017), as both metrics are well-suited for assessing machine translation (MT) quality. In CHRF++, the parameter β is set to 2, as this configuration has been shown to achieve the highest Kendall's τ segment-level correlation with human relative rankings (RR), ensuring more reliable evaluation results (Popović, 2016).

[6] https://huggingface.co/GoToCompany/llama3-8b-cpt-sahabatai-v1-instruct
[7] https://huggingface.co/GoToCompany/gemma2-9b-cpt-sahabatai-v1-instruct

5 Results and Analysis

The experimental results for the fine-tuned models in Task 1 are presented in Table 5. Additionally, the performance of the zero-shot classification task using Off-the-Shelf models is summarized in Table 7. The results of the honorific style translation task, evaluated using BLEU, are reported in Table 9, while the complementary results based on the CHRF++ metric are provided in Table 20 in the Appendix. Similarly, the results of the honorific cross-lingual translation task, evaluated using BLEU, are detailed in Table 10, with the corresponding CHRF++ results presented in Table 22 in the Appendix. Finally, the performance of the models on the conversation generation task is presented in Table 11 and further illustrated in Figure 2.

5.1 Baseline and Fine-tuned Honorific Level Classification

Model	Acc.	Prec.	Rec.	F1
Baseline				
Dictionary-Based Model	88.37	88.92	88.37	88.64
LSTM Model	93.47	91.56	92.78	91.34
Fine-tuning				
Javanese BERT	93.91	94.09	93.91	93.97
Javanese GPT2	92.42	92.48	92.42	92.43
Javanese DistilBERT	**95.65**	**95.66**	**95.65**	**95.66**

Table 5: Performance comparison of different fine-tuned models on Task 1: Honorific Level Classification.

The rule-based dictionary model achieved an accuracy of 88.37%, as shown in Table 5. While less flexible than machine learning approaches, this model provided a useful baseline for comparison. Additionally, the LSTM model, employed as a baseline deep learning approach, achieved an accuracy of 93.47% (Table 5). This result highlights the effectiveness of traditional neural networks in classifying Javanese honorific levels when trained on a well-annotated corpus. The strong performance of the LSTM model may be partially attributed to the relatively small dataset, as LSTM models tend to perform well with limited data (Ezen-Can, 2020).

Under identical experimental conditions, fine-tuned transformer-based models demonstrated superior performance compared to the baselines. `Javanese DistilBERT` achieved the highest accuracy at 95.65%, surpassing both `Javanese BERT` (93.91%) and `Javanese GPT-2` (92.42%). The fine-tuned `Javanese DistilBERT` outperformed all other classifiers (Table 5 & Table 7), making it the most suitable choice to automatically assess the honorific level of the generated text in Task 4.

To assess robustness when classifier models are trained and tested on sentences with different meanings, Table 6 presents the results of cross-validation with grouping. We group sentences based on meaning before splitting them into folds. This ensures that sentences with the same meaning but different honorific levels are kept in the same fold, either all in the training set or all in the test set. As a result, models are trained to focus on differences in honorific levels rather than meaning correlations. The results demonstrate the model's strong generalization capabilities across different manners of splitting data.

Model	Acc.	Prec.	Rec.	F1
Cross Validation				
Javanese BERT	**93.99**	**94.04**	**93.99**	**93.96**
Javanese GPT2	91.48	91.48	91.48	91.40
Javanese DistilBERT	93.10	93.14	93.10	93.05

Table 6: Performance of fine-tuned models on Javanese honorific classification with cross-validation settings, where models are evaluated using 5-fold cross-validation on the training set to assess consistency and reliability.

5.2 Off-the-Shelf Honorific Level Classification

In Table 7, among the closed-source and open-source models, `GPT-4o` achieved the highest accuracy (53.50%) in Task 1: Honorific Level Classification. Meanwhile, `Gemini 1.5 Pro` obtained the highest scores in precision, recall, and F1-score.

An analysis of evaluation metrics per honorific label reveals a strong tendency among Off-the-Shelf models to classify texts into the *Ngoko* label. This bias is particularly evident in `Gemini 1.5 Pro`, which achieved the highest precision (90.10%) and F1-score (86.70%) for the *Ngoko* label, while `GPT-4o` attained the highest recall (91.10%) for the same label. Additionally, the *Ngoko* label consistently yielded the highest evaluation metric scores across nearly all Off-the-Shelf models, further reinforcing the classification bias toward this honorific level. Since these models perform poorly on other honorific levels, their overall performance remains lower than that of the fine-tuned models (Table 5). This bias toward the *Ngoko* honorific level may be attributed to the predominance of *Ngoko* text in existing Javanese language corpora used to pre-train these LLMs.

Model		Accuracy	Precision	Recall	F1
Closed-source					
GPT-4o		**53.50**	40.30	49.80	40.70
Gemini 1.5 Pro		50.70	**54.60**	**54.20**	**45.40**
Open-source					
Sailor2 8B		33.60	20.90	24.00	16.40
Gemma2 9B Instruct		28.40	42.30	23.00	17.80
Llama3.1 8B Instruct		43.00	20.60	30.50	24.00
SahabatAI v1 Instruct (Llama3 8B)		48.50	38.00	35.10	31.00
SahabatAI v1 Instruct (Gemma2 9B)		47.50	35.00	34.70	30.00
Honorific-Level Evaluation					
Model		**Level**	**Precision**	**Recall**	**F1**
GPT-4o		Ngoko	78.00	**91.10**	84.00
		Ngoko Alus	0	0	0
		Krama	53.50	26.00	35.00
		Krama Alus	29.90	82.40	43.80
Gemini 1.5 Pro		Ngoko	**90.10**	80.00	**84.70**
		Ngoko Alus	26.10	43.10	32.50
		Krama	70.60	10.50	18.30
		Krama Alus	31.70	83.40	45.90
Sailor2 8B		Ngoko	34.40	86.30	49.20
		Ngoko Alus	4.00	0.30	0.6
		Krama	37.50	8.20	13.50
		Krama Alus	7.50	1.30	2.30
Gemma2 9B Instruct		Ngoko	88.90	6.00	1.10
		Ngoko Alus	5.60	10.50	7.30
		Krama	36.80	72.30	48.80
		Krama Alus	37.70	8.70	14.10
Llama3.1 8B Instruct		Ngoko	46.70	87.70	61.00
		Ngoko Alus	0	0	0
		Krama	35.80	34.40	35.10
		Krama Alus	0	0	0
SahabatAI v1 Instruct (Llama3 8B)		Ngoko	61.20	71.50	65.90
		Ngoko Alus	19.40	3.40	5.80
		Krama	40.60	64.20	49.70
		Krama Alus	30.80	1.30	2.60
SahabatAI v1 Instruct (Gemma2 9B)		Ngoko	51.80	86.30	64.70
		Ngoko Alus	0	0	0
		Krama	41.00	45.60	43.20
		Krama Alus	47.20	7.00	12.20

Table 7: Comparison of zero-shot overall and per-honorific level classification (Task 1) performance across different Off-the-Shelf models. The upper section reports overall Accuracy, Precision, Recall, and F1-score, while the lower section presents performance per honorific level. Additionally, detailed information about 0 result could be seen in Appendix.T 8.

Model	Level	Misinterpreted Label Ngoko (%)	Krama (%)	Krama Alus (%)
GPT-4o	Ngoko Alus	31.36	24.41	44.24
Llama3.1 8B Instruct	Ngoko Alus	50.00	50.00	0
SahabatAI v1 Instruct	Ngoko Alus	43.22	53.22	3.56
Model	**Level**	**Ngoko (%)**	**Ngoko Alus (%)**	**Krama (%)**
Llama3.1 8B Instruct	Krama Alus	32.95	0	67.05

Table 8: Percentage of label misclassification in the honorific level classification experiment where the evaluation metrics—Accuracy, Precision, and Recall—yield a value of 0.

Further analysis in Table 7 and Table 8 reveals that the *Ngoko Alus* label is frequently misclassified, with errors spread across the other three labels. No instances are predicted as *Ngoko Alus*, suggesting that models struggle to identify this honorific level. Misclassification is also observed for *Krama Alus*, particularly in `LLaMA3.1 8B Instruct`,

Model	Ngoko→X			Ngoko Alus→X			Krama→X			Krama Alus→X		
	Ngoko Alus	Krama	Krama Alus	Ngoko	Krama	Krama Alus	Ngoko	Ngoko Alus	Krama Alus	Ngoko	Ngoko Alus	Krama
Copy Baseline	30.33	10.07	8.23	30.00	8.66	28.08	10.00	8.68	31.46	8.15	28.07	31.37
GPT-4o	23.74	**22.88**	**19.72**	**38.83**	**16.83**	**34.95**	**36.80**	**17.70**	**32.60**	**29.92**	24.30	25.55
Gemini-1.5-pro	2.52	**11.77**	4.36	24.71	5.97	5.28	**27.39**	**12.94**	0.78	**21.73**	15.49	9.36
Sailor2 8B	1.02	1.13	1.15	1.01	1.04	0.89	1.19	0.90	0.40	1.24	1.12	0.59
Gemma2 9B Instruct	0.73	0.21	0.14	1.01	0.33	0.83	0.31	0.22	0.76	0.20	0.79	1.22
Llama3.1 8B Instruct	0.36	0.10	0.08	0.20	0.09	0.33	0.17	0.11	0.33	0.12	0.29	0.32
SahabatAI v1 (Llama3 8B)	0.72	0.21	0.20	0.63	0.27	0.78	0.39	0.23	0.81	0.30	0.61	0.56
SahabatAI v1 (Gemma2 9B)	0.47	0.23	0.39	0.69	0.13	0.34	0.61	0.36	0.35	0.47	0.49	0.29

Table 9: Performance of honorific style translation across different models and different honorific levels measured using BLEU. The **Copy Baseline** represents the BLEU score obtained by directly comparing the input text with the ground truth output, without any model inference. Bold text indicate model performance that surpass the Copy Baseline, indicating some style change capabilities.

where it is often confused with *Krama*, indicating difficulty in distinguishing between adjacent honorific levels.

5.3 Honorific Style Change

In this task, the `GPT-4o` model outperforms all other Off-the-Shelf models. Its performance surpasses some of the copy baseline, where inputs are directly compared to ground truth outputs without inference. This suggests that `GPT-4o` possesses some degree of style translation capability.

Table 9 shows that when translation tasks are grouped by honorific level, the best results are observed in *Ngoko* and *Krama Alus*. Style translation toward these levels consistently exceeds the copy baseline, indicating that models perform better when distinguishing between honorific levels with significant differences, such as *Ngoko* (the lowest level) and *Krama Alus* (the highest). In contrast, the models struggle to differentiate between honorific levels that are closer in hierarchy, e.g., *Ngoko* to *Ngoko Alus* or *Krama Alus* to *Krama*.

5.4 Cross-lingual Honorific Translation

In this task, `GPT-4o` and `Gemini 1.5 Pro` are the top-performing models, while the others exhibit significantly lower performance. `GPT-4o` outperforms all models in nearly every translation category, except for *Indonesian → Krama*, where `Gemini 1.5 Pro` leads with a marginal BLEU score advantage of 0.55. In all other categories, `GPT-4o` achieves the highest performance, notably in *Ngoko → Indonesian* translation, where it attains 38.28 BLEU.

As shown in Table 10, translation ***from Indonesian*** generally yields lower performance compared to translation ***into Indonesian***. For example, in the *Indonesian → Krama Alus* and *Krama Alus → Indonesian* categories, the *Indonesian → Krama Alus* translation achieves a BLEU score of only 15.56,

Model	ID → X: Ngoko	Ngoko Alus	Krama	Krama Alus
GPT4o	**27.81**	**14.52**	14.31	**15.56**
Gemini-1.5-pro	16.76	6.99	**14.86**	7.05
Sailor2 8B	4.1	1.2	0.89	0.98
Gemma2 9B Instruct	0.43	0.29	0.12	0.11
Llama3.1 8B Instruct	0.04	0.07	0.04	0.04
SahabatAI v1 Instruct (Llama3 8B)	0.78	0.37	0.11	0.13
SahabatAI v1 Instruct (Gemma2 9B)	6.27	3.49	1.06	1.59
Model	**X → ID: Ngoko**	**Ngoko Alus**	**Krama**	**Krama Alus**
GPT4o	**38.28**	**30.58**	**33.31**	**27.47**
Gemini-1.5-pro	32.74	23.82	29.34	23.81
Sailor2 8B	7.27	6.36	6.71	5.8
Gemma2 9B Instruct	3.29	2.22	1.55	1.77
Llama3.1 8B Instruct	0.29	0.23	0.16	0.23
SahabatAI v1 Instruct (Llama3 8B)	4.51	3.64	3.19	3.07
SahabatAI v1 Instruct (Gemma2 9B)	15.0	11.5	10.39	9.27

Table 10: Cross-lingual evaluation using BLEU scores for translation between Indonesian (ID) and Javanese honorific levels.

whereas the *Krama Alus → Indonesian* translation yields a significantly higher BLEU score of 27.47. This finding confirms that translating from Javanese to Indonesian is generally easier for language models than the reverse.

Among translations to Javanese, those targeting *Ngoko* achieve the highest performance across models, mirroring the trend observed in Tasks 1 to 3, where LLMs exhibit a bias toward *Ngoko*.

5.5 Conversation Generation

Model	Without Hint		With Hint	
	A	B	A	B
GPT-4o	43.12	**60.00**	42.50	60.00
Gemini-1.5-pro	46.25	**60.00**	42.50	**60.62**
Sailor2 8B	**48.75**	58.12	**45.00**	56.25
Gemma2 9B Instruct	42.50	58.12	41.25	58.12
Llama3.1 8B Instruct	40.62	55.00	42.50	60.00
SahabatAI (Llama3 8B)	41.88	57.50	41.25	57.50
SahabatAI (Gemma2 9B)	41.25	58.75	43.12	58.75

Table 11: Honorific level generation accuracy for conversation generation task.

The `Sailor` model achieved the highest accu-

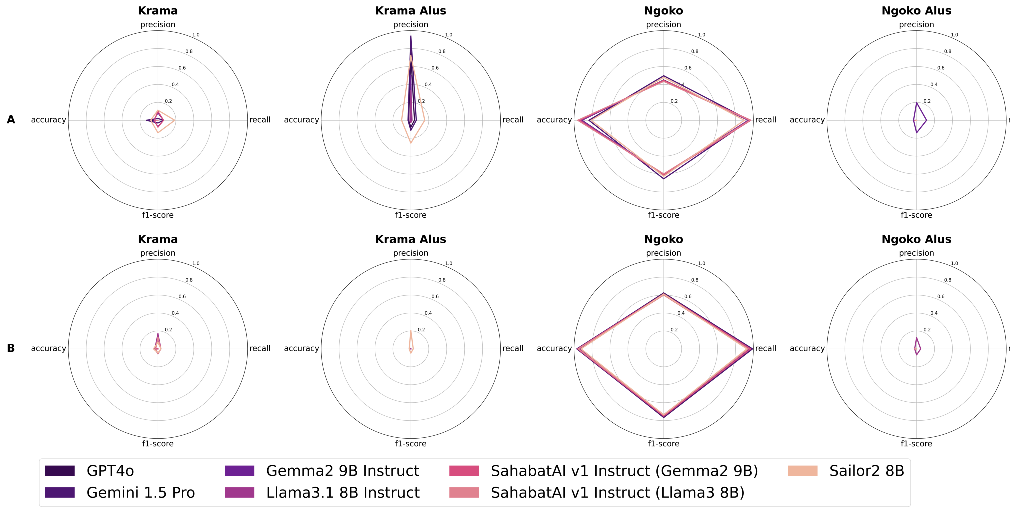

Figure 2: A comparative analysis of conversation generation quality across models, categorized by different honorific labels and the utterance (A or B).

racy in classifying speaker A's utterance (the first speaker) in both experiments, with and without hints (Table 11). For speaker B's utterance, the highest accuracy was attained by the `Gemini 1.5 Pro` and `GPT-4o` models. Providing hints in the form of usage guidelines for each honorific level did not consistently lead to improvements.

Additionally, Figure 2 presents the evaluation metrics for each honorific label, showing that the *Ngoko* label consistently achieves the highest scores. These findings suggest that the models continue to struggle with fully understanding the nuanced usage of different honorific levels in conversation. Specifically, they tend to generate dialogues predominantly in *Ngoko*, indicating a preference for this honorific over others.

Notably, the models consistently perform better when generating utterances for speaker A than for speaker B. As shown in Figure 2, speaker B surpasses speaker A only in the *Ngoko* label. This suggests that models face greater difficulty in generating appropriate honorific-level speech when responding to a conversation, whereas they demonstrate greater proficiency in initiating dialogue with the correct honorific level.

6 Conclusion and Future Work

We introduce the Javanese Honorific Corpus (UNGGAH-UNGGUH) as a valuable linguistic resource and evaluate the ability of modern NLP models to handle Javanese honorific-related tasks, including classification, machine translation, and text generation. Our experimental results reveal that LLMs exhibit a bias toward a specific honorific level (*Ngoko*), primarily due to the imbalanced distribution of honorific tiers in existing Javanese datasets.

Additionally, cross-lingual performance on Javanese text with distinct honorific levels remains poor, and LLMs struggle to accurately interpret and generate honorifics in conversational contexts. This study's methodologies and datasets provide a foundation for future research on low-resource languages with complex honorific systems. Future work may explore dialectal variations, finer linguistic nuances, or larger datasets to enhance model performance and generalizability.

Limitations

Corpus Size The UNGGAH-UNGGUH contains 4,024 sentences and 160 conversation cases, which may not capture the full complexity of Javanese honorific. Expanding the corpus with more diverse sources, including spoken language, would improve generalizability.

Dialect The corpus does not account for regional dialects of Javanese so it limits model accuracy in different linguistic regions. Future work should include dialectal variations for broader coverage.

Ethical Consideration

Misclassification of honorifics can result in disrespectful interactions within Javanese social contexts. It is crucial for users of these models to be aware of their limitations in handling language features that are culturally sensitive.

Acknowledgements

This research was funded by Lembaga Pengelola Dana Pendidikan (LPDP), the Ministry of Education, Culture, Research and Technology of the Republic of Indonesia through the Indonesia-US Research Collaboration in Open Digital Technology program, and Monash University's Action Lab. Their support for this research is deeply appreciated, and we acknowledge their vital role in the successful completion of this work. The findings and conclusions presented in this publication are those of the authors and do not necessarily reflect the views of the sponsors.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Asif Agha. 1994. Honorification. *Annual Review of Anthropology*, pages 277–302.

Asif Agha. 1998. Stereotypes and registers of honorific language. *Language in society*, 27(2):151–193.

Samuel Cahyawijaya, Holy Lovenia, Fajri Koto, Rifki Afina Putri, Emmanuel Dave, Jhonson Lee, Nuur Shadieq, Wawan Cenggoro, Salsabil Maulana Akbar, Muhammad Ihza Mahendra, et al. 2024. Cendol: Open instruction-tuned generative large language models for indonesian languages. *arXiv preprint arXiv:2404.06138*.

Longxu Dou, Qian Liu, Guangtao Zeng, Jia Guo, Jiahui Zhou, Wei Lu, and Min Lin. 2024. Sailor: Open language models for south-east asia.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Gabriel Enrique, Ika Alfina, and Evi Yulianti. 2024. Javanese part-of-speech tagging using cross-lingual transfer learning. *IAES International Journal of Artificial Intelligence*, 13(3):3498–3509.

Aysu Ezen-Can. 2020. A comparison of LSTM and BERT for small corpus. *CoRR*, abs/2009.05451.

Brilliant Filby, Utomo Pujianto, Jehad AH Hammad, and Aji Prasetya Wibawa. 2024. Dialect classification of the javanese language using the k-nearest neighbor. *Journal of Information Technology and Cyber Security*, 2(2):111–122.

F. A. A. Ghiffari, I. Alfina, and K. Azizah. 2023. Cross-lingual transfer learning for javanese dependency parsing. *Unpublished Manuscript*.

Haryana Harjawiyana, Th Supriya, et al. 2001. Kamus unggah-ungguh basa jawa. *(No Title)*.

R. Haryanya Harjawiyana and Theodorus Supriya. 2009. *Kamus unggah-ungguh basa Jawa*, edisi revisi 2009 edition. PT Kanisius, Yogyakarta. Cetakan pertama tahun 2001, with consultation by Prof. Dr. Soepomo Poedjosoedarmo.

A. F. Hidayatullah, S. Cahyaningtyas, and R. D. Pamungkas. 2020. Attention-based cnn-bilstm for dialect identification on javanese text. *Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control*, 5(4):317–324.

Yongwoog Andrew Jeon. 2022. Let me transfer you to our ai-based manager: Impact of manager-level job titles assigned to ai-based agents on marketing outcomes. *Journal of Business Research*, 145:892–904.

Aobo Kong, Shiwan Zhao, Hao Chen, Qicheng Li, Yong Qin, Ruiqi Sun, Xin Zhou, Enzhi Wang, and Xiaohang Dong. 2024. Better zero-shot reasoning with role-play prompting. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4099–4113, Mexico City, Mexico. Association for Computational Linguistics.

J. Kreutzer, I. Caswell, L. Wang, A. Wahab, D. van Esch, N. Ulzii-Orshikh, A. Tapo, N. Subramani, A. Sokolov, C. Sikasote, M. Setyawan, S. Sarin, S. Samb, B. Sagot, C. Rivera, A. Rios, I. Papadimitriou, S. Osei, P. O. Suarez, and M. Adeyemi. 2022. Quality at a glance: An audit of web-crawled multilingual datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72.

Solomon Kullback and Richard A Leibler. 1951. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86.

Jianhua Lin. 2002. Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory*, 37(1):145–151.

Gang Liu and Jiabao Guo. 2019. Bidirectional lstm with attention mechanism and convolutional layer for text classification. *Neurocomputing*, 337:325–338.

M. Liu and I. Kobayashi. 2022. *Construction and Validation of a Japanese Honorific Corpus Based on Systemic Functional Linguistics*. Unpublished Manuscript.

Raymasterio Vera Lucky, Baladiffa Aurora Herkanyaka, Vieri Ferdian Putra Basuki, Ivan Halim Parmonangan, et al. 2023. Unveiling sentiments in javanese text: A study on sentiment analysis for the javanese language. In *2023 IEEE 9th Information Technology International Seminar (ITIS)*, pages 1–6. IEEE.

Mounika Marreddy, Subba Reddy Oota, Lakshmi Sireesha Vakada, Venkata Charan Chinni, and R. Mamidi. 2022. Am i a resource-poor language? data sets, embeddings, models and analysis for four different nlp tasks in telugu language. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22:1 – 34.

Atik Zilziana Muflihati Noor, Arie Ardiyanti Suryani, and Ibnu Asror. 2020. Part of speech tagging menggunakan bahasa jawa dengan metode condition random fields. *eProceedings of Engineering*, 7(2).

Louis Owen, Vishesh Tripathi, Abhay Kumar, and Biddwan Ahmed. 2024. Komodo: A linguistic expedition into indonesia's regional languages. *arXiv preprint arXiv:2403.09362*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Maja Popović. 2016. chrF deconstructed: beta parameters and n-gram weights. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 499–504, Berlin, Germany. Association for Computational Linguistics.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Ryan Armiditya Pratama, Arie Ardiyanti Suryani, and Warih Maharani. 2020. Part of speech tagging for javanese language with hidden markov model. *Journal of Computer Science and Informatics Engineering (J-Cosine)*, 4(1):84–91.

Ely Triasih Rahayu. 2014. Comparison of honorific language in javanese and japanese speech community. *International Journal on Studies in English Language and Literature (IJSELL)*, 2(7):140–146.

Faisal Rahutomo, Yoppy Yunhasnawa, Agustaf Fanisnaini Narolis, and Muhammad Rifky Prayanta. 2018. Indonesian javanese dictionary starter kit. `https://doi.org/10.17632/y3hstv4bfn.2`.

Fa'iq Askhabi Ramadhan, Arie Ardiyanti Suryani, and Moch Arif Bijaksana. 2020. Part of speech tagging in javanese using support vector machine method. *eProceedings of Engineering*, 7(2).

G. Ramírez-Sánchez, M. Bañón, J. Zaragoza-Bernabeu, and S. Ortiz Rojas. 2022. Human evaluation of web-crawled parallel corpora for machine translation. In *Proceedings of the 2nd Workshop on Human Evaluation of NLP Systems (HumEval)*, pages 32–41.

Herry Sujaini. 2020. Improving the role of language model in statistical machine translation (indonesian-javanese). *International Journal of Electrical and Computer Engineering*, 10(2):2102.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Aji Wibawa, Andrew Nafalski, A Effendi Kadarisman, and Wayan F Mahmudy. 2013a. *Indonesian-to-Javanese machine translation*. Ph.D. thesis, IACSIT-Internal Association of Computer Science.

Aji Prasetya Wibawa, Andrew Nafalski, Jeffrey Tweedale, Neil Murray, and Ahmad Effendi Kadarisman. 2013b. Hybrid machine translation for javanese speech levels. In *2013 5th International Conference on Knowledge and Smart Technology (KST)*, pages 64–69. IEEE.

J. A. E. Wibawa, S. Sarin, C. Li, K. Pipatsrisawat, K. Sodimana, O. Kjartansson, A. Gutkin, M. Jansche, and L. Ha. 2018. Building open javanese and sundanese corpora for multilingual text-to-speech. In *Unpublished Manuscript*.

Genta Indra Winata, Alham Fikri Aji, Samuel Cahyawijaya, Rahmad Mahendra, Fajri Koto, Ade Romadhony, Kemal Kurniawan, David Moeljadi, Radityo Eko Prasojo, Pascale Fung, et al. 2023. Nusax: Multilingual parallel sentiment dataset for 10 indonesian local languages. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 815–834.

W. Wongso, D. S. Setiawan, and D. Suhartono. 2021a. Causal and masked language modeling of javanese language using transformer-based architectures. In *2021 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, pages 1–7.

Wilson Wongso, David Samuel Setiawan, and Derwin Suhartono. 2021b. Causal and masked language modeling of javanese language using transformer-based architectures. In *2021 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, pages 1–7. IEEE.

G. Udny Yule. 1944. *The statistical study of literary vocabulary*. University Press, Cambridge [England].

A Unggah-Ungguh Basa System

The Unggah-Ungguh Basa system (Figure 8) is a Javanese linguistic framework designed to guide the appropriate use of varying honorific levels depending on conversational context. Based on the work of Harjawiyana and Supriya (2009) and detailed in Table 12, this system comprises four main components: Word Type Levels (Tataran Tembung), Affixes (Tataran Rimbag), and Sentence Type (Warnaning Ukara).

Word Type Levels (Tataran Tembung). This categorization is based on speech formality, ranging from informal (Ngoko) to polite (Krama), and extending to the most formal expressions (Krama Inggil).

Affixes (Tataran Rimbag). This process involves modifying the root of a word by adding a prefix, infix, or suffix, and encompasses three distinct honorific affixes:

- Less formal affixes (Rimbag Ngoko): “dakjupuk”, “kokjupuk”, “dijupukake”.
- Formal affixes (Rimbag Krama): “kulah-pêndhet”, “sampêyan-pêndhet”, “dipundhê-taken”.
- The most formal affixes (Rimbag Krama Inggil): “adalem̩-pêndhet”, “panjenengan-pundhut”.

Sentence Type (Warnaning Ukara). This system encompasses various sentence types, such as commands, requests, narratives, and questions. Responses can be either "tanggap" (active) or "tanduk" (passive).

Language Hierarchical Structure (Undha-Usuk Basa). This represents the hierarchical structure of social interaction or communication as follows:

- Ngoko: Emerges from informal (Ngoko) speech combined with informal discourse (Rimbag ngoko).
- Ngoko Alus: Combines informal (Ngoko) words with high or polite speech levels to reflect politeness.
- Krama: Involves polite (Krama) discourse combined with either informal or polite words, and speech levels of Krama or Ngoko.
- Krama Alus: Incorporates formal speech with polite words, indicating a high level of politeness.

This system highlights the complex interplay of formality and social interaction in Javanese language.

B Dataset

B.1 UNGGAH-UNGGUH's Samples

The UNGGAH-UNGGUH dataset comprises four columns: label, Javanese sentence, group, and Indonesian sentence (Table 13). The label column represents the **honorific level of each Javanese sentence** instance, where label "0" corresponds to the *ngoko* honorific level, label "1" to *ngoko alus*, label "2" to *krama*, and label "3" to *krama alus*. Sentences with the same meaning are grouped under a single group ID, as specified in the "group" column.

B.2 Conversation's Role and Context

Honorific level	A Utterance	B Utterance	Total
Ngoko	69	97	166
Ngoko alus	9	11	20
Krama	8	15	23
Krama alus	74	37	111

Table 14: Honorific level distribution of conversation's status and context dataset.

The *Conversation's Role and Context* dataset contains **160 instances** and is utilized for evaluating language models (LMs) in the task of **conversation generation** involving two speakers (A and B). It comprises **5 primary columns**, which are used to assess the model's ability to perform the task, along with **2 additional columns** that provide example dialogues for both speakers. The **primary columns** include: role A, role B, context, and honorific level for Speaker A's utterance & B's utterance. **Role A, Role B, and Context** columns are written in **English** to enhance accuracy when used in English-language prompts during LM inference. Meanwhile, the distribution of honorific level instances for **Speaker A's** and **Speaker B's utterance** columns can be seen in Table 14. Samples of the *Conversation's Role and Context* primarily column dataset are presented in Table 15. These evaluation scenarios will be released publicly along with the UNGGAH-UNGGUH dataset.

Honorific Level	Affixes (Tataran Rimbag)			Word Type Levels (Tataran Tembung)		
	Krama Inggil	Krama	Ngoko	Krama Inggil	Krama	Ngoko
Krama Alus	✓	✓	✗	✓	✓	✓
Krama	✗	✓	✗	✗	✓	✓
Ngoko Alus	✓	✗	✓	✓	✗	✓
Ngoko	✗	✗	✓	✗	✗	✓

Table 12: The mapping of Honorific Levels includes their corresponding affixes and word type levels.

Label	Javanese Sentence	Group	Indonesian Sentence	English Sentence
0 1 2 3	Nggaawa jeruk kuwi! Panjenengan ngampil jeruk kuwi! Sampeyan mbekta jeram menika! Panjenengan ngampil jeram menika!	1	Membawalah jeruk itu.	Bring that orange.
0 1 2 3	Tustele rusak, digawa ya ora ana gunane. Tustele rusak, dipamila ya ora ana gunane. Tustelipun risak, dipunbektaa inggih boten wonten ginanipun. Tustelipun risak, dipunamila inggih boten wonten ginanipun.	2	Tustelnya rusak, seandainya dibawa tidak ada gunanya.	The camera is broken, so there's no point in bringing it.
0 1 2 3	Mobile dak-undurake, abanana! Mobile adalem undurake, panjenengan paringi dhawu! Mobilipun kula unduraken, sampeyan abani! Mobilipun adalem unduraken, panjenengan paringi dhawu!	3	Mobil saya undurkan, berilah aba-aba!	I am reversing my car, give me a signal!
0 1 2 3	Yen wedange arep diladekake, ngabanana! Yen unjukane arep diladekake, panjenengan mringi dhawu! Yen wedangipun badhe dipunladosaken, sampeyan ngabani! Yen unjukanipun badhe dipunladosaken, panjenengan maringi dhawu!	4	Jika minuman akan disajikan, memberilah isyarat!	If the drinks are going to be served, give a signal!

Table 13: Samples from UNGGAH-UNGGUH Dataset.

C Prompt Template

We utilize **6 distinct straightforward prompts** (Fig. 3, 4, 5, 6, and 7), each designed for a specific objective, covering four tasks along with several variations of certain tasks. The prompts used for the *honorific style change* and *honorific cross-lingual translation* tasks share a dominant structural similarity, with only minor modifications. Meanwhile, the **hints** included in the *conversation generation* prompts serve as guidelines for the appropriate usage of different levels of Javanese honorific. These hints are intended to facilitate the conversation generation process, as the context provided to the model includes information extracted from the hints, particularly regarding the relationship between role A and role B. It is important to note that text highlighted in red within the prompts represents input that needs to be provided.

D Generation Result

D.1 Honorific Style Change

Table 16 presents the inference results of the Off-the-Shelf models on the honorific style change task. The GPT-4o and Gemini 1.5 Pro models produce outputs that are relatively comparable to the ground truth for each instance. In contrast, the other five models exhibit significantly lower performance. The Sailor2 8B model, when tested on instances from the fourth group of the UNGGAH-UNGGUH dataset, generates outputs that are nearly correct but

Role A	Role B	Context	A Utterance Category	B Utterance Category
Teacher	Student	A asked speaker B about what he had eaten today. Speaker A has a higher status or position than Speaker B.	0	2
Peer	Peer	Speaker A asks speaker B about what Speaker B's father ate today. Speaker A and Speaker B have equal status or position and have familiar interactions.	1	1
Older sibling	Younger sibling	Speaker A asks speaker B about what Speaker B learned at school today. Speaker B has a lower status or position than Speaker A.	0	1
Grandchild	Grandmother	Speaker A gives him speaker B's glasses. Speaker A has a lower status or position than Speaker B.	3	0

Table 15: Primary column samples of conversation's status and context dataset.

contain subordinate clauses that can be identified as hallucinations, as they deviate from the given input context. Meanwhile, the Gemma2 9B Instruct and SahabatAI v1 Instruct (Llama3 8B) models show a similar tendency of failing to modify the input appropriately, often producing translations that include vocabulary outside the Javanese language. Additionally, the Llama3.1 8B Instruct and SahabatAI v1 Instruct (Llama3 8B) models exhibit a comparable pattern, frequently translating the given input into Indonesian rather than transforming it into the intended honorific level.

D.2 Honorific Cross-lingual Translation

The inference results for the honorific cross-lingual translation task, presented in Tables 17 & 18, indicate that translation performance for the "**to Indonesia**" direction is generally better than for the "**from Indonesia**" direction. The GPT-4o and Gemini 1.5 Pro models continue to demonstrate superior performance, while the other models tend to produce lower-quality outputs. These models frequently exhibit hallucinations, generating translations that are entirely unrelated to the input text. Additionally, they often misinterpret certain words, including common vocabulary. For instance, the SahabatAI v1 Instruct (Gemma2 9B) model incorrectly translates *jeruk* (orange) as *jambu* (guava).

D.3 Conversation Generation

Table 19 presents the generated outputs from the conversation generation task for each Off-the-Shelf model using specific instances. In general, the inference results across all models exhibit errors, particularly the inclusion of Indonesian vocabulary within the generated conversational sentences. Commonly misused words (Indonesian vocabulary) include apa kabar (how are you), saya (I), dimakan (eaten), mangga (mango), nasi (rice), aku (I), baru (new), and disini (here). Additionally, grammatical errors are observed in the model outputs. For instance, the inference result from the Gemini 1.5 Pro model contains a sub-clause that reads, "Iyo, sing sinau sregep...", which has an incorrect grammatical structure. The correct phrasing should be "Iyo, sinau sing sregep...".

E Additional Results

E.1 Honorific Style Change

Table 20 presents the performance evaluation of various models in translating between different Javanese honorific levels, measured using CHRF++ scores. The Copy Baseline represents the CHRF++ score obtained by directly comparing the input text with the ground truth without model inference, serving as a reference point for assessing model effec-

Model	Input (*Ngoko*)	Generation	Gold (*Krama Alus*)
GPT 4o	Rambutmu ambune kok wangi, nganggo sampo apa?	Rambut panjenengan ambune kok wangi, nganggo sampo punapa?	Rikma panjenengan gandanipun kok wangi, ngagem sampo menapa?
Gemini 1.5 Pro	Bu Haiti arep tuku gula selawe kilo, wis dak-dolane.	Ibu Haiti badhé mundhut gula kalih dasa gangsal kilo, sampun kula damelaken.	Bu Haiti badhe mundhut gendhis selangkung kilo, sampun adalem sadeanipun.
Sailor2 8B	Yen wedange arep diladekake, ngabanana!	Yen wedangipun badhe dipunladosaken, kula aturi dhateng ing papan.	Yen unjukanipun badhe dipunladosaken, panjenengan maringi dhawuh!
Gemma2 9B Instruct	Adhimu iku alise nanggal sapisan.	Nubarkah iku alise nanggal sapisan.	Ingkang rayi menika imbanipun nanggal sapisan.
Llama3.1 8B Instruct	Aku arep matur ana ing ngarepmu kabeh.	Saya harap sekali dapat menyapa kawan-kawan seluruhnya.	Adalem badhe munjuk wonten ing ngarsa panjenengan sadaya.
SahabatAI v1 Instruct (Llama3 8B)	Para kawula njaluk supaya Raden Wijaya madeg ratu ing Majapait.	Para satriya ngajak agar Prabhu Jayanagara dadi raja ing Majapahit.	Para kawula nyuwun supados Raden Wijaya jumeneng ratu ing Majapait.
SahabatAI v1 Instruct (Gemma2 9B)	Mobile dak-undurne, abanana!	‘Mobile‘ipun ngaktipaken banjur dipununduh.	Mobilipun adalem unduraken, panjenengan paringi dhawuh!

Table 16: Generation samples of honorific style change task (*Ngoko*→*Krama Alus*).

Model	Input (*Ngoko*)	Generation	Gold (*Indonesian*)
GPT 4o	Abang-abang kae klambi apa serbet?	Merah-merah itu baju apa serbet?	Merah-merah itu baju atau serbet?
Gemini 1.5 Pro	Menakjinggane kae, sing raine diabang-abang.	Hebat sekali dia, yang mukanya merah padam.	Menakjingganya itu, yang wajahnya diberi warna merah!
Sailor2 8B	Mobile dak-undurne, abanana!	Ia tidak mundur dari yang sudah-sudah,	Mobil saya undurkan, berilah aba-aba!
Gemma2 9B Instruct	Abot-abot nggawa tela iku arep dianggo apa?	Apa yang akan dibawa oleh semua orang itu untuk digunakan?	Berat-berat membawa ketela itu akan dipergunakan untuk apa?
Llama3.1 8B Instruct	Koran-koran sing wis ora kanggo iku dak-dolane wae.	Terakhir sekali koran itu tidak berguna.	Koran-koran yang sudah tidak terpakai itu saya jualinya saja.
SahabatAI v1 Instruct (Llama3 8B)	Tustele rusak, digawaa ya ora ana gunane.	Kerusakan terjadi pada bagian yang tidak berguna.	Tustelnya rusak, seandainya dibawa tidak ada gunanya.
SahabatAI v1 Instruct (Gemma2 9B)	Nggawaa jeruk kuwi!	Ambil jeruk itu!	Membawalah jeruk itu.

Table 17: Generation samples of honorific cross-lingual translation task (*Ngoko*→*Indonesian*).

Model	Input (*Indonesian*)	Generation	Gold (*Ngoko*)
GPT 4o	Merah-merah itu baju atau serbet?	Abang-abang kuwi klambi apa serbet?	Abang-abang kae klambi apa serbet?
Gemini 1.5 Pro	Menakjingganya itu, yang wajahnya diberi warna merah!	Nggilani kuwi, sing raine diwenehi werna abang!	Menakjinggane kae, sing raine diabang-abang.
Sailor2 8B	Mobil saya undurkan, berilah aba-aba!	Mobile kulo munduri.,	Mobile dak-undurne, abanana!
Gemma2 9B Instruct	Berat-berat membawa ketela itu akan dipergunakan untuk apa?	kono ketale iki akan dipergunakan untuk apa	Abot-abot nggawa tela iku arep dianggo apa?
Llama3.1 8B Instruct	Koran-koran yang sudah tidak terpakai itu saya jualinya saja.	wong-wong wis ora, iki aku tuku.	Koran-koran sing wis ora kanggo iku dak-dolane wae.
SahabatAI v1 Instruct (Llama3 8B)	Tustelnya rusak, seandainya dibawa tidak ada gunanya.	Palingan kanggo nggemborke wong-wong neng omah.	Tustele rusak, digawaa ya ora ana gunane.
SahabatAI v1 Instruct (Gemma2 9B)	Membawalah jeruk itu.	Gawawo woh jambu iku!	Nggawaa jeruk kuwi!

Table 18: Generation samples of honorific cross-lingual translation task (*Indonesian*→*Ngoko*).

tiveness. The results indicate that GPT-4o achieves the highest performance, surpassing the baseline across all translation tasks. Gemini-1.5-pro also demonstrates notable performance, particularly in Ngoko → Krama and Krama → Krama Alus translations. In contrast, models such as Sailor2 8B, Gemma2 9B Instruct, and Llama3.1 8B Instruct exhibit significantly lower scores, suggesting limited

Model	Role A	Role B	Context	A Utterance	B Utterance
GPT 4o	Schoolmate	Schoolmate	The speaker talked with his schoolmates at a reunion event. Speaker A and Speaker B have equal status or position and have familiar interactions.	Aku uga apik. Kangen karo jaman sekolah mbiyen, ya?	Iyo, kangen banget. Kowe isih eling wektu kita sering dolanan bareng sak kelas?
Gemini 1.5 Pro	Grandfather	Grandchild	Speaker A gives pocket money to speaker B to provide for going to school. Speaker A has a higher status or position than Speaker B.	Iyo, sing sinau sregep. Aja dolan wae. (Yes, study diligently. Don't just play)	Inggih, Mbah. Kula badhe sregep sinau
Sailor2 8B	A friend from the same village who has not met each other for a long time	A friend from the same village who has not met each other for a long time	Speaker A asks speaker B about the news from speaker B. Speaker A and Speaker B have equal status or position and have familiar interactions.	Hai, apa kabar? Wis suwe aku ora ketemu kowe.	Kabarku apik-apik wae Mas. Sampeyan piye?
Gemma2 9B Instruct	People who do not know each other	People who do not know each other	Speaker A asked speaker B about the direction to Gadjah Mada University. Speaker A has a higher status or position than Speaker B	Mbak, ngerti ora mboten arah kanggo Universitas Gadjah Mada?	Maaf Pak, aku kurang tau kok. Aku baru di sini.
Llama3.1 8B Instruct	Older sibling	Younger sibling	Speaker A asks speaker B about what Speaker B learned at school today. Speaker B has a lower status or position than Speaker A.	Saya kudu tahu apa kang wedhe siswamu sabene dina, ndak?	Mbok, saya iki wis sinaos nulis pethi ing buku lan nyimpening ngisor majelise.
SahabatAI v1 Instruct (Llama3 8B)	Peer	Peer	Speaker A asks speaker B about what Speaker B's father ate today. Speaker A and Speaker B have equal status or position and have familiar interactions.	Apa sing dimakan bapakmu dina iki?	Mbok ya nasi goreng, tapi aku ora ngerti pasti
SahabatAI v1 Instruct (Gemma2 9B)	Teacher	Student	Speaker A asked speaker B about what he had eaten today. Speaker A has a higher status or position than Speaker B.	Nggih, nak? Sampeyan wis mangan apa dina iki?	Mangga lan roti goreng

Table 19: Samples of conversation generation task (A and B utterance).

effectiveness in handling honorific style transformations. Table 21 provides the distribution of sentence instances for each honorific level translation pair. The dataset is balanced across the different transformations, with the highest number of instances (1,412) observed in the Ngoko ↔ Krama pair. This distribution ensures that models have sufficient training data for evaluating performance across both direct and reverse translations. However, the lower number of instances for Ngoko Alus transformations may contribute to the models' difficulty in accurately translating this level.

E.2 Honorific Cross-lingual Translation

Table 22 presents the CHRF++ scores for evaluating the cross-lingual translation between Indonesian (ID) and different Javanese honorific levels. The results indicate that GPT-4o consistently achieves the highest scores across all translation directions, demonstrating superior performance in both ID→Javanese and Javanese→ID tasks. Gemini-1.5-pro follows closely, particularly excelling in ID→Krama and Krama→ID translations. In contrast, models such as Sailor2 8B, Gemma2 9B Instruct, and Llama3.1 8B Instruct exhibit substantially lower scores, suggesting challenges in handling both Indonesian-to-Javanese and Javanese-to-Indonesian honorific translations. Additionally, Table 23 provides the number of sentence instances available for each translation process. The dataset is balanced across the different honorific levels, with the highest number of instances (1,412) observed in the Ngoko Alus↔ID pair. This distribution ensures that models are evaluated on a sufficiently diverse set of translation tasks.

F Computational Resources & Hyper-parameters

We utilized four A40 40GB GPUs for the experiments related to downstream tasks.

Model	Ngoko→X			Ngoko Alus→X			Krama→X			Krama Alus→X		
	Ngoko Alus	**Krama**	**Krama Alus**	**Ngoko**	**Krama**	**Krama Alus**	**Ngoko**	**Ngoko Alus**	**Krama Alus**	**Ngoko**	**Ngoko Alus**	**Krama**
Copy Baseline	50.16	34.71	28.63	54.69	31.58	59.61	38.81	32.48	56.55	33.12	63.35	58.36
GPT-4o	47.00	**52.66**	**49.02**	**59.79**	**45.34**	**65.28**	**60.80**	**39.51**	**60.32**	**53.03**	48.21	52.89
Gemini-1.5-pro	24.14	**41.89**	22.47	49.28	**32.55**	30.07	**51.88**	**36.36**	16.05	**46.01**	41.55	37.35
Sailor2 8B	19.57	20.17	21.10	18.68	19.75	21.72	20.72	19.42	18.97	20.54	20.22	17.79
Gemma2 9B Instruct	14.23	12.82	11.99	14.44	12.61	15.40	12.56	12.62	15.51	11.79	15.73	16.49
Llama3.1 8B Instruct	13.26	13.15	12.61	11.88	12.96	15.81	12.31	12.31	14.42	11.28	15.38	14.26
SahabatAI v1 (Llama3 8B)	15.86	14.72	14.62	13.92	14.80	18.93	14.98	14.05	17.27	14.11	17.12	15.83
SahabatAI v1 (Gemma2 9B)	17.19	16.29	17.90	16.88	15.42	18.97	17.85	15.91	18.12	16.25	18.38	16.18

Table 20: Performance evaluation of style translation models across different Javanese honorific levels, measured using CHRF++ scores. The **Copy Baseline** represents the CHRF++ score obtained by directly comparing the input translation text with the ground truth, without undergoing model inference. Bold text indicate model performance surpassing the Copy Baseline, demonstrating effective translation capabilities.

Translation Pairs (Bidirectional) Label 1	Label 2	Number of Instances (Sentences)
Ngoko	Ngoko Alus	585
Ngoko	Krama	1412
Ngoko	Krama Alus	595
Ngoko Alus	Krama	584
Ngoko Alus	Krama Alus	588
Krama	Krama Alus	591

Table 21: Number of sentence instances for each honorific level translation process. Each translation process is bidirectional, meaning the count for a given pair (e.g., Ngoko ↔ Krama) is identical in both directions.

F.1 Honorific Level Classification

The model is configured with temperature 0.1, top-p 0.9, and a maximum of 50 new tokens. A low temperature (0.1) ensures highly deterministic outputs, which is crucial for accurate classification. The top-p (0.9) parameter helps retain relevant token probabilities while allowing slight flexibility in word selection. The max_new_tokens (50) limit prevents unnecessary token generation, ensuring concise and precise classification.

F.2 Honorific Style Changes

For translating from one honorific level to another, the model generates a single output sequence to maintain consistency. A temperature of 0.7 ensures a balance between randomness and coherence, while top-p (0.9) and top-k (50) filtering help retain relevant word choices. The repetition penalty (1.2) discourages redundant text generation, and early stopping prevents unnecessary token extension. The min_length (10) and max_new_tokens (100) parameters control the length of generated text, ensuring meaningful responses without excessive verbosity.

F.3 Honorific Cross-lingual Translation

When translating between Javanese and Indonesian across different honorific levels, the hyperparameters remain consistent to preserve translation fairness. The model generates a single sequence while using temperature (0.7) and top-p (0.9) to maintain lexical variety. The top-k (50) value helps focus on relevant words while avoiding excessive randomness. Early stopping enhances efficiency, and a repetition penalty (1.2) ensures fluency. The minimum length constraint (10) prevents overly short outputs, while the maximum token limit (100) maintains reasonable translation length.

F.4 Conversation Generation

For generating honorific-level conversations in Javanese, the hyperparameters are set to balance fluency and creativity. The model generates one response per prompt, while temperature (0.7) introduces controlled randomness to simulate natural conversations. Top-p (0.9) and top-k (50) work together to refine token selection, preventing the model from choosing unlikely continuations. The repetition penalty (1.2) avoids repetitive responses, and early stopping enhances efficiency. The length constraints (min_length = 10, max_new_tokens = 100) help maintain coherent and meaningful dialogue structures.

F.5 LSTM

The model's architecture includes an embedding layer with a vocabulary size of 10,000 and an embedding size of 64. Input sequences are limited to a maximum length of 128. A convolutional layer with 32 filters and a kernel size of 3 is applied, followed by a max pooling layer to reduce dimensionality. The key component is a bidirectional LSTM with 32 units, capturing sequential

Model	ID → Ngoko	ID → Ngoko Alus	ID → Krama	ID → Krama Alus
GPT4o	**54.14**	**36.99**	42.59	**46.16**
Gemini-1.5-pro	47.49	32.33	**45.55**	31.66
Sailor2 8B	26.50	20.36	17.99	19.85
Gemma2 9B Instruct	11.85	11.45	10.98	11.03
Llama3.1 8B Instruct	11.60	12.47	12.15	12.38
SahabatAI v1 Instruct (Llama3 8B)	16.19	15.06	14.13	13.93
SahabatAI v1 Instruct (Gemma2 9B)	31.02	25.68	21.47	23.26
Model	**Ngoko → ID**	**Ngoko Alus → ID**	**Krama → ID**	**Krama Alus → ID**
GPT4o	**61.23**	**53.25**	**56.61**	**51.49**
Gemini-1.5-pro	58.44	51.65	55.83	51.22
Sailor2 8B	32.06	29.76	30.52	29.68
Gemma2 9B Instruct	20.43	17.70	15.88	15.27
Llama3.1 8B Instruct	12.99	12.39	12.46	12.37
SahabatAI v1 Instruct (Llama3 8B)	29.07	26.66	25.97	25.10
SahabatAI v1 Instruct (Gemma2 9B)	42.12	37.99	37.21	34.90

Table 22: Cross-lingual evaluation using CHRF++ scores for translation between Indonesian (ID) and Javanese honorific levels.

Translation Pairs (Bidirectional) Label 1	Label 2	Number of Instances (Sentences)
Ngoko	ID	585
Ngoko Alus	ID	1412
Krama	ID	595
Krama Alus	ID	584

Table 23: Number of sentence instances for each translation process. Each translation process is bidirectional, meaning the count for a given pair (e.g., ID ↔ Ngoko) is identical in both directions.

dependencies from both directions. This architecture, which combines a bidirectional LSTM with convolutional and max-pooling layers, has proven effective in text classification tasks (Liu and Guo, 2019). A dropout layer with a rate of 0.4 is used to prevent overfitting. The final output layer is a dense layer with 4 units, representing the honorific levels, and uses softmax activation. The model is trained using the Adam optimizer and categorical cross-entropy loss, with early stopping based on validation accuracy to avoid overfitting. Training is conducted for up to 100 epochs with a batch size of 32.

F.6 Finetuned Models

The fine-tuned models demonstrated robust performance across all evaluation metrics. Table 24 details the hyper-parameters used in these experiments, revealing that a learning rate of 5e-5 and a batch size of 16 consistently produced strong results across models. The hyper-parameter tuning was conducted on a validation subset to refine the candidate of classifier models' performance. All fine-tuning experiments were conducted with a fixed random seed (set to 42) to ensure consistent results across runs.

G Others

G.1 Shannon Entropy

$$H = -\sum_{i=1}^{n} p_i \log_2(p_i), \tag{1}$$

where p_i is probability of a specific honorific form.

G.2 Yule's characteristic K value

$$K = 10^4 \times \frac{\sum_m [m^2 V(m, N)] - N}{N^2} \tag{2}$$

where N represents the total number of word tokens, and $V(m, N)$ is a function that calculates the number of word types occurring exactly m times in the corpus.

G.3 Kullback-Leibler Divergence

$$\mathrm{KL}_{\mathrm{sym}}(P, Q) = \frac{1}{2}\sum_i P(i)\log\frac{P(i)}{Q(i)} + \frac{1}{2}\sum_i Q(i)\log\frac{Q(i)}{P(i)} \tag{3}$$

where $P(i)$ and $Q(i)$ denote the probability of token i in the Javanese and Indonesian distributions, respectively. We apply Laplace smoothing to avoid zero probabilities and use a symmetric version to ensure comparability across directions.

G.4 Jensen Score

$$\mathrm{JSD}(P \parallel Q) = \frac{1}{2}\mathrm{KL}(P \parallel M) + \frac{1}{2}\mathrm{KL}(Q \parallel M), \quad \text{where } M = \frac{1}{2}(P + Q) \tag{4}$$

where P and Q are the token distributions of Javanese and Indonesian sentences, and M is the average distribution. The Jensen score corresponds to the Jensen-Shannon divergence and is computed as the squared distance output from `scipy`'s `jensenshannon` function.

G.5 Rule Based Algorithm

Algorithm 1 classifies a given Javanese sentence into one of four honorific levels: *Ngoko* (informal), *Ngoko Alus* (polite informal), *Krama* (formal), or *Krama Alus* (highly formal). It first counts the occurrences of words associated with each speech level and calculates their proportions relative to the total words in the sentence. The classification is determined based on the dominant proportion, with ties defaulting to *Ngoko*. Special words from the *Krama Inggil* lexicon can upgrade the classification, while the presence of *Ngoko* words can downgrade it, ensuring a more nuanced classification.

G.6 Javanese Language Dataset

Table 25 presents an overview of various Javanese language corpora, detailing their distribution across four honorific levels: *Ngoko*, *Ngoko Alus*, *Krama*, and *Krama Alus*. The dataset sizes vary significantly, ranging from 105 to over 185,000 entries, reflecting diverse sources and linguistic compositions. Additionally, Shannon's Entropy is computed for each corpus to quantify variability in honorific usage, where higher entropy values indicate a more balanced distribution across honorific levels, suggesting a richer representation of Javanese linguistic diversity. Notably, the dataset UNGGAH-UNGGUH exhibits the highest entropy (1.8763), implying a well-balanced honorific distribution, while corpora such as *The Identifikasi-Bahasa* and *OSCAR-2301 Javanese* display lower entropy, indicating a more skewed distribution. The data presented in this table corresponds to the visualization depicted in Figure 1, which illustrates the honorific level distributions and entropy values across different corpora.

Model	lr	Batch Size	Acc.	Prec.	Rec.	F1
Javanese BERT	5e-5	16	93.01	92.98	93.01	92.97
	5e-5	32	**93.19**	**93.16**	**93.19**	**93.16**
	2e-5	16	92.11	92.15	92.11	92.12
	2e-5	32	91.03	91.09	91.03	91.06
	1e-5	16	89.78	89.70	89.78	89.73
	1e-5	32	81.36	80.49	81.36	78.97
Javanese GPT	5e-5	16	**92.65**	**92.67**	**92.65**	**92.65**
	5e-5	32	89.96	90.18	89.96	89.96
	2e-5	16	86.38	86.56	86.38	86.37
	2e-5	32	82.62	83.28	82.62	82.83
	1e-5	16	80.47	80.13	80.47	80.18
	1e-5	32	74.55	73.92	74.55	73.34
Javanese DistilBERT	5e-5	16	**93.19**	**93.12**	**93.19**	**93.13**
	5e-5	32	91.40	91.29	91.40	91.32
	2e-5	16	88.35	88.09	88.35	88.15
	2e-5	32	86.20	86.39	86.20	85.28
	1e-5	16	79.57	79.33	79.57	74.65
	1e-5	32	73.30	65.61	73.30	63.64

Table 24: Performance comparison of different models with hyper-parameter tuning conducted on the validation subset of the Javanese Honorific Corpus. While the **javanese-bert-small-imdb-classifier** model achieved the best overall results with a batch size of 32, the performance with a batch size of 16 was only slightly lower. Across all models, a learning rate of 5e-5 and a batch size of 16 consistently provided strong results, making this the most stable hyper-parameter combination.

Prompt example for Task 1: Honorific Level Classification

Analyze the Javanese sentence enclosed in square brackets.
Determine if it is ngoko, ngoko alus, krama, or krama alus.
Return the answer as the corresponding text label: 0 (ngoko), 1 (ngoko alus), 2 (krama), 3 (krama alus).
Provide only the integer label without any additional explanation. [<SENTENCE>]

Figure 3: Honorific level classification task's prompt.

Dataset	Honorific Level's Distribution Percentage				Size	Shannon Entropy
	Ngoko (%)	Ngoko Alus (%)	Krama (%)	Krama Alus (%)		
UNGGAH-UNGGUH	35.26	14.66	35.14	14.94	4024	1.87
The identifikasi-bahasa	96.05	0.93	2.91	0.11	44109	0.27
Javanese dialect identification	92.09	2.21	4.53	1.17	16498	0.50
Korpus-Nusantara (Jawa)	9.00	1.00	86.40	3.60	1000	0.73
Korpus-Nusantara (Jawa Ngoko)	82.59	9.95	4.49	2.97	6059	0.91
JV-ID-ASR	74.80	3.90	19.49	1.81	185076	1.06
JV-ID TTS (female)	66.97	1.36	27.41	4.26	2864	1.17
JV-ID TTS (male)	67.58	1.15	27.08	4.19	2958	1.15
OSCAR-2301 Javanese	95.24	0.95	3.81	0.00	105	0.31
Jvwiki	89.20	2.95	6.23	1.62	37335	0.64

Table 25: Javanese Language Corpus Statistics

Algorithm 1 Rule-Based Classification of Javanese Speech Levels

Require: A sentence s to classify
Ensure: Classification label $l \in \{0, 1, 2, 3\}$ where:
  0: Ngoko
  1: Ngoko Alus
  2: Krama
  3: Krama Alus

Initialize word counts:
$w_{\text{ngoko}} \leftarrow \text{count\_words}(s, \text{“ngoko”})$
$w_{\text{krama}} \leftarrow \text{count\_words}(s, \text{“krama”})$
$w_{\text{kramaAlus}} \leftarrow \text{count\_words}(s, \text{“krama\_alus”})$
$w_{\text{kramaInggil}} \leftarrow \text{count\_words}(s, \text{“krama\_inggil”})$
$w_{\text{total}} \leftarrow w_{\text{ngoko}} + w_{\text{krama}} + w_{\text{kramaAlus}} + w_{\text{kramaInggil}}$

Calculate proportions:
$p_{\text{ngoko}} \leftarrow w_{\text{ngoko}} / w_{\text{total}}$
$p_{\text{krama}} \leftarrow w_{\text{krama}} / w_{\text{total}}$
$p_{\text{kramaAlus}} \leftarrow w_{\text{kramaAlus}} / w_{\text{total}}$

Initial classification:
if $p_{\text{ngoko}} > \max(p_{\text{krama}}, p_{\text{kramaAlus}})$ **then**
  $l \leftarrow 0$ {Ngoko}
else if $p_{\text{krama}} > \max(p_{\text{ngoko}}, p_{\text{kramaAlus}})$ **then**
  $l \leftarrow 2$ {Krama}
else
  $l \leftarrow 3$ {Krama Alus}
end if

Handle ties:
if $p_{\text{ngoko}} = p_{\text{kramaAlus}}$ **then**
  $l \leftarrow 0$ {Default to Ngoko}
end if

Adjust based on special words:
if $w_{\text{kramaInggil}} > 0$ **then**
  $l \leftarrow \min(l + 1, 3)$ {Upgrade level}
end if
if $w_{\text{ngoko}} > 0$ **then**
  $l \leftarrow \max(l - 1, 0)$ {Downgrade level}
end if

return l

Prompt example for Task 2: Honorific Style Change

Translate this Javanese sentence from <HONORIFIC LEVEL A> to <HONORIFIC LEVEL B>:
Source: <TEXT A>
Target:

Figure 4: Honorific style change task’s prompt.

Prompt example for Task 3: Cross-lingual Honorific Translation

Translate this Indonesian sentence to Javanese using <HONORIFIC LEVEL>:
Indonesian: <INDONESIAN TEXT>
Javanese:

Translate this Javanese sentence to Indonesian:
Javanese: <JAVANESE TEXT>
Indonesian:

Figure 5: Cross-lingual honorific translation task’s prompt.

Prompt example for Task 4: Conversation Generation (without hint)

Create a conversation between A as <ROLE A> and B as <ROLE B> in Javanese language with this context: '<CONTEXT>'
Please follow this format:
A: '<UTTERANCE>'
B: '<UTTERANCE>'

Answer:

Figure 6: Coversation generation task's prompt (without hint).

Prompt example for Task 4: Conversation Generation (with hint)

Create a conversation between A as <ROLE A> and B as <ROLE B> in Javanese language with this context: '<CONTEXT>'
Please follow this format:
A: '<UTTERANCE>'
B: '<UTTERANCE>'

Use this Javanese's honorific level usage as a hint:
1. Ngoko:
- Used for informal conversations with peers or lower-status individuals
- Common in close relationships or familiar interactions
2. Ngoko Alus:
- Adds respect when speaking to equals or higher-status individuals in informal or close relationships
- Flexible for conversations with mixed-status participants
- Talked with a person with equals status about other person who has a higher-status 3. Krama:
- Used for respectful conversations with equals or higher-status individuals, especially when not close
- Suitable for maintaining formality in less familiar interactions
4. Krama Alus:
- Expresses high respect in conversations with higher-status individuals or unfamiliar equals
- Essential for formal interactions requiring utmost politeness
- Talked with a person with higher-status about other person who has a higher-status

Answer:

Figure 7: Coversation generation task's prompt (with hint).

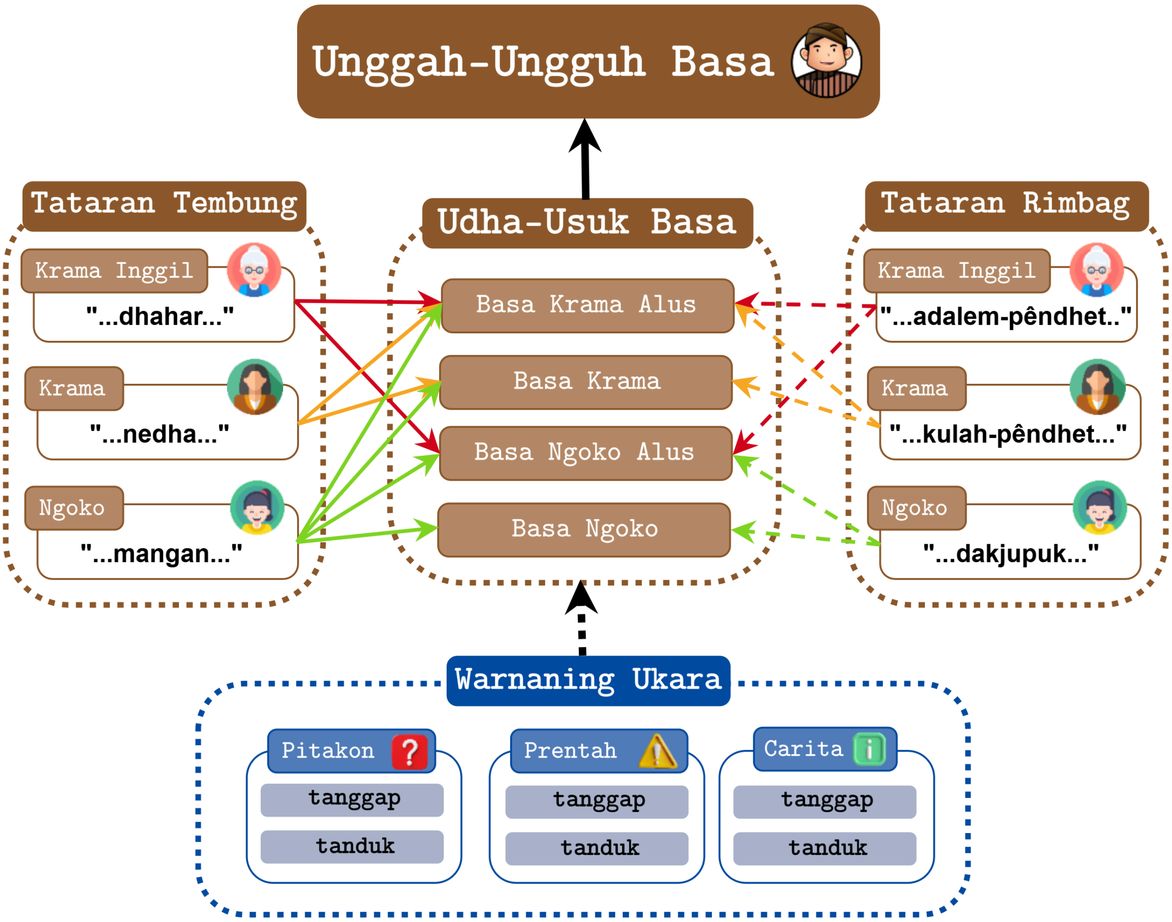

Figure 8: Diagram of the Unggah-Ungguh Basa System.