How do Encoder-only LMs Predict Closeness and Respect from Thai Conversations?

Pakawat Nakwijit[1]
[1]Queen Mary University of London
{p.nakwijit, m.purver}@qmul.ac.uk

Attapol T. Rutherford[2]
[2] Chulalongkorn University
Bangkok, Thailand
attapol.t@chula.ac.th

Matthew Purver[1,3]
[3]Jožef Stefan Institute
Ljubljana, Slovenia

Abstract

This study explores how encoder-only Language Models (LMs) recognize social relationships from textual data, examining both the models' behaviour and structure. Behaviourally, we analyze word importance, determined by SHAP values, to identify which lexical features—such as pronouns, sentence-final particles, and spelling variations—most influence the model's predictions in different conversational settings. Our findings confirm the use of these lexical features in the model's predictions albeit with varying degrees of contribution. We also validate our results by demonstrating a significant correlation between SHAP values and human evaluations. Structurally, we explore the impact of spelling variations on the structure of the encoder-generated word embeddings in social dimensions; closeness and respect. Using our projection approach, we observe a shift along both social dimensions when spelling variations are introduced in pronouns. Overall, this study sheds light on the mechanisms underlying the encoder model's social relationship recognition and contributes to verifying the alignment between the lexical features used by the model and human intuition.

1 Introduction

Our communication style, including word choice and tone, plays a crucial role in expressing our social identity and relationships, such as closeness and respect (Halliday, 1978; Poynton, 1991). Recognizing these social cues is, however, highly contextual and difficult to identify using traditional methods, particularly in Thai, a language that places a strong emphasis on social harmony and linguistic propriety (Knutson et al., 2003).

The advent of powerful architectures like the Transformer (Vaswani et al., 2017) and its derivatives, such as BERT (Devlin et al., 2019), initiated a new era, achieving remarkable performance across various NLP tasks, including social relationship recognition. However, their complex "black-box" nature renders their inner workings opaque, posing challenges for model interpretation and potentially leading to the generation of harmful content or hallucinations (Weidinger et al., 2021). Therefore, developing explainability mechanisms is critical to elucidate how these models operate enabling users to understand the rationale behind predictions or generated text, fostering trust, accountability, and responsible deployment across various NLP applications (Zhao et al., 2024; Doshi-Velez and Kim, 2017).

This study aims to address these challenges by developing a model proficient in recognizing closeness and respect using encoder-only Language Models (LMs), while simultaneously illuminating the underlying reasoning processes of these models through behaviour and structure aspects of the model. Firstly, we investigated word importance, estimated by SHAP value, to observe what lexical features (including pronouns, sentence-final particles and spelling variation) contribute the most to the model's predictions. We compared them across different conversational settings (private/public conversations, self-reported/perceived labels). In the end, we can confirm that all three lexical features contribute to the model's predictions. It, however, contributes to a different degree in different settings. For instance, first-person pronouns emerge as the primary contributor to the model's predicted closeness across all conversational contexts, surpassing other pronoun types. Conversely, singular pronouns only contribute to perceived closeness. Similarly, words with morphophonemic variation only influence predicted respect within private conversations.

Secondly, we explored the structure of the encoder-generated word embeddings in a social context by projecting the model's word embeddings onto dimensions representing closeness and respect. We presented our work on the investigation of how the introduction of spelling variations affects the

Proceedings of the 28th Workshop on the Semantics and Pragmatics of Dialogue, September, 11–12, 2024, Trento, Italy.

model's embeddings. Our findings demonstrate that introducing spelling variations in pronouns does not alter the overall shape of the projected distribution of word embeddings along the dimensions of closeness and respect. However, there is a notable shift towards increased closeness and decreased respect, as confirmed by the Mann-Whitney U test on the mean values. This underscores the model's sensitivity to linguistic nuances in shaping social perceptions.

2 Related Works

In this section, we review various explanation techniques tailored for LMs, categorizing them into two subsections based on their targeted facets of explainability. The first subsection delves into methods designed to provide an explanation from input features to determine the importance of each input token, for a given prediction (Behavioural). The second subsection explores methods that delve into the internal representation of LMs, seeking to discern its correlations with linguistic features (Structural).

2.1 Behavioural Analysis

Behavioural analysis often relies on strategically manipulating model inputs to observe their resulting behaviour. This approach leverages the inherent explanatory power of input features in NLP, where inputs directly correspond to human-interpretable elements like words, sub-words, or characters. By identifying the most influential words, researchers can gain valuable insights into the model's internal decision-making processes.

One prominent approach is Local Interpretable Model-Agnostic Explanations (LIME) by Ribeiro et al. (2016). LIME approximates the behaviour of complex models using a simple model trained locally around specific data points. To provide an explanation for an individual data point, a model, often a linear model due to its simplicity, is trained on data sampled locally around that specific instance. This localized training aims to approximate the behaviour of the original complex model within this restricted region of the feature space. This allows for explanations tailored to an individual instance. The authors demonstrated that explanations generated using LIME can accurately reflect the underlying behaviour of the model. However, LIME's explanatory power is limited to individual instances (local explanations). Additionally, Lundberg and Lee (2017) also highlighted potential shortcomings in LIME, including violations of local accuracy and consistency properties. These limitations can lead to counterintuitive explanations in certain scenarios.

Another method, SHapley Additive exPlanations (SHAP) by Lundberg and Lee (2017), built upon the well-established mathematical concept of Shapley values (Shapley, 1952). SHAP treats input features as contributors to a prediction outcome in a cooperative game. It assigns each feature subset a value reflecting its contribution. This approach offers strong expressiveness, particularly for LMs. Unlike LIME, Lundberg and Lee (2017) demonstrated that it satisfies all desirable properties including local accuracy, missingness and consistency. Additionally, SHAP also allows for global interpretations by averaging its values for each feature across a dataset which have been shown to be consistent with the local explanations (Molnar, 2018; Covert et al., 2020). Notably, Wu et al. (2021) exemplified a successful SHAP application in dataset construction by using it as a guide for their experts in designing counterfactual examples. Hayati et al. (2021) used SHAP to investigate how a model predicts linguistic styles by contrasting lexicons highlighted by humans with those exhibiting high SHAP scores. In this work, we employed SHAP in a comparable manner by aggregating importance scores across three lexical features. These scores are then used to evaluate the significance of each lexical feature across different conversational settings and to assess their alignment with human-annotated scores.

2.2 Structural Analysis

Structural analysis aims to observe linguistic knowledge embedded within the internal representations of the model. It is commonly achieved through probing techniques, which use a simple model, often a logistic regression, to determine whether a target linguistic structure can be predicted from the learned representation. Mohebbi et al. (2021) successfully demonstrated that representations in models like BERT encapsulate linguistically relevant information, encompassing both syntactic and semantic aspects. Their findings also suggest that lower layers predominantly capture word-level syntax, while higher layers excel at encoding sentence-level syntax and semantic knowledge, akin to human language processing. However, Belinkov (2022) argued that conclusions drawn from probing techniques may not always be as robust

as they appear. With sufficiently high-dimensional embeddings, complex probes, and large auxiliary datasets, the probes can seemingly learn to extract any information from any embeddings.

An alternative approach to understanding the model's structure involves examining how the model encodes information within its representations. Torroba Hennigen et al. (2020) extended probing techniques by assessing probe performance on different subsets of dimensions to locate the amount of linguistic information encoded within distinct subsets of dimensions. Their research revealed that many morphosyntactic features are reliably encoded by only a small number of neurons. Kozlowski et al. (2019) adopted a different perspective by projecting embeddings to provide visual explanations. They leveraged the principle that word embeddings should be able to capture semantics as arithmetic relationships between embeddings in a high-dimensional space. Their work illustrated that dimensions induced by pre-trained embeddings correspond to dimensions of cultural meaning (e.g. rich/poor). The projection of words onto these dimensions reflects widely shared stereotypes of social class. For instance, words like "golf" and "tennis" are associated with rich individuals, while "boxing" is linked to lower socioeconomic status. In this study, we adopt a similar approach to investigate how the introduction of spelling variations influences the model's embeddings. This analysis aims to reaffirm that lexical information is effectively represented within the model.

3 Conversation Corpus

The corpus utilized in this study was collected from Nakwijit et al. (2024). The corpus comprises a diverse collection of Thai conversational texts sourced from two sources; 1,234 private conversations specifically curated from their study and 2,496 public conversations from X (formerly Twitter). The corpus is organised into two tasks, including closeness and respect, with three conversational settings, including

- Setting 1: Private Conversations with Self-Reported Relationships (Private-Self)
- Setting 2: Public Conversations with Perceived Relationships (Public-Perceived)
- Setting 3: Private Conversations with Perceived Relationships (Private-Perceived)

They also provided a set of lexicons from 15 lexical features. In this study, we only focus on three lexical features including pronouns, sentence-final particles and spelling variations. Throughout the experiments, we linearized the utterances in a conversation and marked the beginning of each utterance with *[sys]* or *[usr]* to indicate those who initiated the conversation and the respondent. More detailed descriptions of the corpus and lexical features can be found in Appendix A and Appendix B.

4 Social Relationship Models

In this section, we outline our experiments concerning the construction of a social relationship model. Subsequently, the best model according to the F1 score from each setting was selected for further analysis in the subsequent sections of the study.

4.1 Experimental Setup

Before model training and analysis, the corpus underwent standard preprocessing procedures, converting text to lowercase, replacing repeated characters with a *[REP]* token, and tokenizing the text using PyThaiNLP's tokenizer (Phatthiyaphaibun et al., 2023). Following the original paper, we confined our target labels to three levels of closeness and respect, discarding the minority. Labels for closeness and respect were then normalized to a continuous range between -1 and 1, where -1 and 1 denote the lowest and highest degrees of closeness or respect in that setting.

Lastly, we randomly shuffled the corpus and partitioned it into 80% for training, 10% for validation, and 10% for testing. Standard machine learning protocols were followed: training was conducted on the training set, hyperparameters were tuned on the validation set for optimal F1-score, and final metrics were reported based on the test set. The final predictions were discretized back into three labels using thresholds of -0.5 and 0.5 accordingly.

4.2 Selected Models

We experimented with 6 models; 3 simple baselines, and 3 LMs, which are as follows:

Majority-class Model: This model serves as the simplest approach by predicting solely the majority class. It sets a minimum baseline performance that accounts for label imbalances.

Naive Bayes Classifier: It is a probabilistic model based on Bayes' theorem. It operates under the naive assumption of conditional indepen-

dence between individual words, given the class label. This simplification makes it suitable as a baseline model when it is constrained to employ only surface-level lexical information. In essence, it gauges the extent to which closeness and respect levels can be predicted solely based on observable lexicons.

Logistic Regression: An Ordinary Least Squares regression (OLS) model was employed, utilizing 15 linguistic features as predictors such as the number of unique words, number of turns, number of long words, and average number of words per utterance. This model served as a baseline to gauge the predictive power conferred solely by the linguistic features of the conversation.

Fine-tuned XLM-R: It is a multilingual language model designed for understanding and generating text across 100 languages (Conneau et al., 2020).

Fine-tuned WangChanBERTa: It is a monolingual language model trained on a Thai corpus (Lowphansirikul et al., 2021).

Fine-tuned PhayaThaiBERT: It is an extended version of WangChanBERTa via vocabulary transfer to compensate for a lack of foreign vocabulary and orthographic variations in the previous models (Sriwirote et al., 2023).

All three encoders were selected for their status as competitive models, which can leverage pre-trained common-sense knowledge, surface-level lexical information, and broader contextual information. Although they all utilize the RoBERTa architecture (Liu et al., 2019), they vary in terms of their multilingual capabilities (for XLM-R) versus monolingual capabilities (for WangChanBERTa and PhayaThaiBERT) and in the size of their vocabularies, ranging from small (25k words in WangChanBERTa) to large (250k words in PhayaThaiBERT).

We followed the standard fine-tuning practice on WangchanBERTa. The fine-tuning parameters for the model were set as follows:

- Learning rate: 2e-5
- Optimiser: Adam
- Weight decay rate: 0.01
- Number of epochs: 20
- Batch size: 16
- Input max length: 128
- Select the best model with F1 score

Each model was trained five times and reported the average results according to F1 score. The numbers are presented in Table 1.

4.3 Model Performance

A noticeable improvement emerges when additional information is incorporated into the model. The Naive Bayes model, with direct access to surface-level information such as word frequency in a conversation, demonstrates decent performance, achieving F1 scores ranging from 0.43 to 0.56 for closeness and 0.47 to 0.67 for respect—constituting 82% to 90% of the best model's performance. This finding aligns with previous research, suggesting that lexicons alone can serve effectively as social markers (Schwartz et al., 2013). Conversely, linear regression on lexical features yields slightly inferior results, ranging from 0.33 to 0.54 for closeness and 0.31 to 0.46 for respect. Our best model, fine-tuned PhayaThaiBERT, effectively predicts closeness labels with F1 scores ranging from 0.50 to 0.67 and respect labels from 0.43 to 0.75 closely followed by fine-tuned WangChanBERTa and XLM-R.

All LMs surpassed other baselines in nearly all settings, highlighting the importance of pre-trained knowledge, such as contextual representations and common ground knowledge. However, it was evident that XLM-R, as a multilingual model, performed considerably worse than the other two monolingual models. Additionally, vocabulary expansion notably enhanced PhayaThaiBERT's performance over WangChanBERTa in 5 out of 6 settings.

Upon closer examination, all models struggled in two specific settings: *Closeness Setting2: Public-Perceived* and *Respect Setting1: Private-Self*, with F1 scores of only 0.50 and 0.43, respectively. One possible reason for this may be unclear guidelines during data collection, as suggested by the notably low validation agreement observed in *Respect Setting1: Private-Self* (Nakwijit et al., 2024). However, this does not fully explain the models' relative success in other settings, given that the same groups of annotators annotated all labels. Another potential explanation could be that while some settings exhibit consistent and clear linguistic patterns, the constructs of self-perceived respect and perceived closeness are inherently more complex and/or subtle than previously understood. Nevertheless, investigating this matter further falls outside the scope of our study.

5 Understanding Model's Behaviour Through SHAP

In this section, our objective is to ascertain the extent to which each lexicon and lexicon type con-

Model	Task1: Closeness			Task2: Respect		
	Setting 1 Private-Self	Setting 2 Public-Perceived	Setting 3 Private-Perceived	Setting 1 Private-Self	Setting 2 Public-Perceived	Setting 3 Private-Perceived
Baseline						
Majority-class Baseline	0.155	0.206	0.401	0.179	0.276	0.308
Naive Bayes Classifier	0.563	0.435	0.542	0.470	0.678	0.535
Logistic Regression	0.400	0.327	0.542	0.314	0.444	0.463
LMs						
XLM-R	0.604	0.420	0.498	0.200	0.675	0.432
WangChanBERTa	0.657	0.490	0.639	0.313	0.748	**0.761**
PhayaThaiBERT	**0.666**	**0.496**	**0.657**	**0.431**	**0.750**	0.712

Table 1: The f1 performance metrics of our social relationship models in the closeness and respect tasks across three conversational settings

tributes to the model's predictions.

5.1 Methodology

In our analysis, SHAP values were computed using our best model (fine-tuned PhayaThaiBERT). The contribution score for each word in the conversations was calculated. These values were then grouped by their respective lexical features, converted into absolute values, averaged, and subsequently reported in Table 2 and Table 3 for closeness and respect tasks.

5.2 Results and Discussion

Based on the SHAP values, pronouns emerge as a pivotal contributor to the prediction process, exhibiting average SHAP values of 1.13, 4.52 and 1.04 per token for closeness tasks and 1.88, 2.93 and 1.71 per token for respect tasks. These values surpass the baseline derived from random tokens in five of six settings. The numbers also suggest that pronouns with different morphosyntactic features, such as grammatical person and numbers, contribute differently to closeness tasks. Specifically, first-person pronouns contribute in all settings while second-person pronouns are more significant in settings involving private conversations, and third-person pronouns are mainly relevant only in perceived closeness in private conversations. Singular pronouns solely contribute to perceived closeness, while plural pronouns do not exert more influence on closeness than random tokens. Interestingly, pronouns in spelling variation form, which are typically considered as noise, make substantial contributions to predictions in perceived closeness. These findings are even more pronounced in respect tasks, where second-person, singular, and non-standard-written pronouns consistently outperform the random baseline across all settings.

Regarding sentence-ending particles, the findings highlight disparities between two particle subtypes: socially-rated and non-socially-rated. The SHAP values clearly reveal that the model relies on socially-related particles as cues for closeness, while not doing so for the latter subtype. Furthermore, we observe that particles with non-standard spelling influence the model's predictions of closeness and respect more than the random baseline across all three settings, with SHAP values of 1.33, 7.63, and 1.11 for closeness tasks, and 1.54, 2.14, and 0.98 for respect tasks. However, these values are still lower than the SHAP values of pronouns and pronouns with non-standard spellings in four out of six settings.

Spelling variations, on the other hand, do not exhibit high SHAP values across all settings. Its contributions from subtypes of the variations, however, become more pronounced. Morphophonemic variations, for instance, demonstrate SHAP values per token of 1.26, 5.37 and 0.95 in the closeness tasks, and 1.52. 1.90 and 0.86 in the respect tasks. Like pronouns, these values exceed the random baseline in 5 out of 6 settings. Importantly, in those 5 settings, its total contribution even surpasses that of pronouns and sentence-final particles by a considerable margin due to the higher frequency of spelling variations compared to pronouns and particles. This finding underscores the important role of spelling variations, especially in public conversations.

Our observations align closely with the findings obtained from the original work which presents the corpus and analyses it through statistical analysis (Nakwijit et al., 2024). This correspondence may provide evidence that the model leverages analogous linguistic cues to predict the target labels. We,

Lexical Features	Setting 1 Private-Self		Setting 2 Public-Perceived		Setting 3 Private-Perceived	
	Per token	Total	Per token	Total	Per token	Total
Reference						
Average per token	1.08	125.36	4.07	147.01	0.85	97.91
Pronoun						
All pronoun	1.13	4.05	4.52	9.47	1.60	5.65
» 1st person pronoun	1.25	2.85	5.15	7.73	1.14	2.56
» 2nd person pronoun	1.30	3.29	4.33	7.68	2.04	5.11
» 3rd person pronoun	0.71	1.31	3.47	5.61	1.71	3.14
» Singular pronoun	1.13	4.04	4.52	9.40	1.60	5.65
» Plural pronoun	1.07	1.07	4.30	5.73	0.49	0.49
» Pronoun in non-standard spelling	0.74	1.58	7.62	10.02	1.23	2.44
Sentence-final Particles						
All particles	1.75	8.81	4.16	7.54	0.93	4.68
» Socially-related particles	3.24	10.03	5.08	7.27	1.31	4.08
» Non-socially-related particles	0.85	2.97	3.47	5.45	0.69	2.43
» Particle in non-standard spelling	1.33	1.86	7.63	8.41	1.11	1.56
Spelling Variation						
All spelling variation	1.10	14.48	4.39	19.46	0.86	11.28
» Common misspelt words	0.83	1.29	3.80	5.24	0.80	1.24
» Morphophonemic variation	1.26	10.49	5.37	15.10	0.95	7.91
» Simplified variation	0.90	5.81	3.63	10.79	0.74	4.77
» Repeated characters	0.85	1.82	3.41	4.47	0.54	1.15

Table 2: The average of absolute SHAP values of three lexical features in **closeness tasks** across 3 conversational settings from **fine-tuned PhayaThaiBERT**. The values highlighted in grey denote values exceeding the SHAP values of their respective random baseline

however, obtained different results when applying the same method to fine-tuned WangChanBERTa and XLM-R. The detailed SHAP values for these two models are reported in appendix E.

5.3 Validation with Human Scores

To assess the validity of the explanation, we asked the participation of 13 native Thai-speaking teenagers aged between 18 and 20 years. Each participant was presented with a set of 1000 words selected based on their highest SHAP values and was asked to select one level of closeness/respect that was most closely associated with the given words. These relationship levels were then quantified using numerical values ranging from -2 to 2. Subsequently, we identified the most frequently selected levels among the participants as the final score corresponding to each word. Finally, we calculated the correlation between the human-assigned score and its SHAP value. The results are presented in Table 4. It is important to note that we excluded *Setting 1: Private-Self* because the principle of self-reported labels does not align with our validation methodology.

The findings are presented in Table 4. Our results reveal that, overall, there exists a weak correlation (r=0.20-0.32) between SHAP values and human scores in all tasks, except the perceived closeness in public conversation *(Setting 2: Public-Perceived)* which aligned with the low f1 in the same task found in Table 1.

Notably, pronouns demonstrate a consistent correlation across all settings, in contrast to sentence-final particles and spelling variations, which do not. Specifically, sentence-final particles only show a correlation in the respect tasks within public conversations, while spelling variations correlate in all settings except that task. The absence of correlation in certain instances remains unclear; this may be attributed to insufficient data or potential discrepancies between human perceptions and machine interpretations.

6 Effect of Spelling Variation on Embedding Structure

To build a further understanding of how the model represents social meaning, we adopt an analysis ap-

Lexical Features	Setting 1 Private-Self		Setting 2 Public-Perceived		Setting 3 Private-Perceived	
	Pertoken	Total	Pertoken	Total	Pertoken	Total
Reference						
Average per token	1.24	143.37	1.95	72.22	0.75	86.78
Pronoun						
All pronoun	1.88	6.75	2.93	6.77	1.71	6.27
» 1st person pronoun	1.74	3.98	1.90	2.98	1.62	3.78
» 2nd person pronoun	2.17	5.48	4.04	7.51	1.80	4.60
» 3rd person pronoun	1.88	3.49	1.95	3.48	0.78	1.44
» Singular pronoun	1.88	6.74	2.95	6.78	1.72	6.27
» Plural pronoun	1.14	1.14	1.09	1.34	0.26	0.26
» Pronoun in non-standard spelling	1.81	3.77	2.88	4.15	1.73	3.86
Sentence-final Particles						
All particles	1.16	5.89	1.87	3.60	0.65	3.27
» Socially-related particles	1.35	4.19	2.85	4.12	0.74	2.29
» Non-socially-related particles	1.05	3.69	1.23	2.11	0.60	2.09
» Particle in non-standard spelling	1.54	2.16	2.14	2.52	0.98	1.37
Spelling Variation						
All spelling variation	1.39	18.31	1.71	7.84	0.77	10.10
» Common misspelt words	1.37	2.13	1.74	2.40	0.88	1.36
» Morphophonemic variation	1.52	12.68	1.90	5.62	0.86	7.16
» Simplified variation	1.21	7.84	1.45	4.50	0.65	4.19
» Repeated characters	0.92	1.97	0.72	0.95	0.88	1.88

Table 3: The average of absolute SHAP values of three lexical features in **respect tasks** across three conversational settings from **fine-tuned PhayaThaiBERT**. The values highlighted in grey denote values exceeding the SHAP values of their respective random baseline

Lexical Features	Closeness		Respect	
	Setting 2 Private	Setting 3 Public	Setting 2 Private	Setting 3 Public
Overall	0.059	0.203*	0.315*	0.240
Pronoun	0.238	0.349*	0.498	0.355
Sentence-final Particles	0.037	0.022	0.017	0.442*
Spelling Variation	0.182*	0.299*	0.215*	0.045

Table 4: The correlations between *PhayaThaiBERT*'s SHAP values and human scores for words from three lexical features and its association with closeness/respect. Values with a p-value less than 0.05 are indicated by an asterisk (*).

proach proposed by Kozlowski et al. (2019). The core idea is to observe how closeness/respect are encoded by the model and how the representation changes when there are changes in linguistic features which we presented by the introduction of spelling variations on pronouns.

6.1 Methodology

The analysis consists of 3 steps: calculating the social dimension, projecting word embeddings onto the dimension and observing the social orientation of the words.

Step 1: Calculating the Social Dimension

Each conversation was represented as the average of the hidden embeddings for each token from the last layer of the fine-tuned PhayaThaiBERT. To represent two extreme groups, the embeddings were separated into two opposite groups based on their annotated labels: *Intimate* and *Dislike* for closeness, and *Highly Respectful* and *Disrespectful* for respect. The embeddings were subsequently averaged, and the vector differences from each pair were utilized as social dimensions for closeness and respect, respectively.

Step 2: Projecting Word Embeddings

In this analysis, our focus was specifically on pronouns, given their notable outcomes thus far. We manually chose pronouns with spelling variants as an illustrative example of how the model changes its representation to align with spelling changes and their associated social meanings. The last hidden embeddings of the selected pronouns from all conversations, were projected onto the constructed dimension using cosine similarity.

Step 3: Observing the Social Orientation of the Words

Finally, we examined the social orientation of pronouns by plotting the distribution of projected values. The resulting plots are presented in Figure 1. Additionally, we conducted the Mann-Whitney U test over the mean value to ascertain whether the values in one group are different from those in the other group and reported the corresponding p-values.

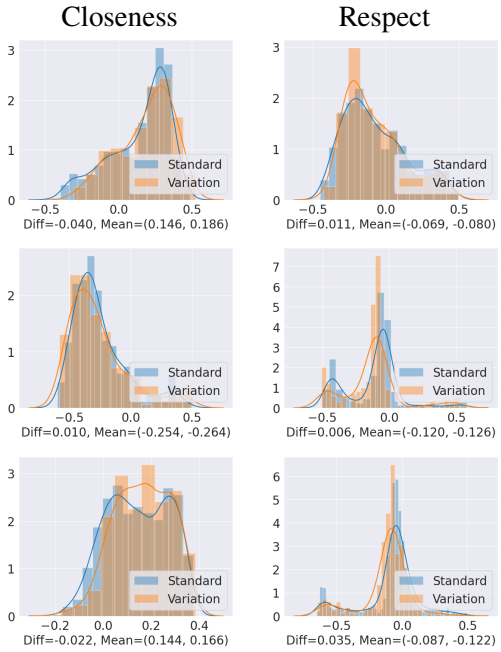

Figure 1: The distribution of social orientation values (cosine similarity) for word embeddings of pronouns and their spelling variations, projected onto dimensions representing closeness and respect from three settings: Private-Self (top), Public-Perceived (middle) and Private-Perceived (bottom)

6.2 Results and Discussion

The Figure 1 shows that, in general, the model represents a pronoun with an embedding that leans toward a closer relationship in private conversation with the average social orientation values of 0.146, and 0.144 for *Setting 1: Private-Self* and *Setting 3: Private-Perceived*. While leaning against a closer relationship in the public one with the average social orientation value of -0.254 for *Setting 2: Public-Perceived*. It, however, consistently leans toward disrespectful relationships across all three settings with the average social orientation values of -0.069, -0.120 and -0.087 for *Setting 1: Private-Self*, *Setting 2: Public-Perceived* and *Setting 3: Private-Perceived* respectively.

Expectedly, our results also suggested that the model represents pronouns and their variants in a similar distribution shape. However, we observed a slight shift in the distribution. The introduction of spelling variation generally makes the model shift toward greater closeness and lesser respect with the differences in mean between the two groups being -0.040*, 0.010, and -0.022* in closeness tasks and 0.011, 0.006*, 0.035* in respect tasks where * indicates when it has p-value less than 0.05. This further confirms that the model can represent social nuance quite nicely.

7 Conclusion

In summary, this research provides valuable insights into the mechanisms guiding encoder-only language models in identifying social relationships from text data. Through a series examination of both behavioural and structural aspects, we illustrated the critical roles played by three lexical features, including pronouns, sentence-final particles, and spelling variation, in shaping model predictions across three conversational settings. By using SHAP, we uncovered nuanced relationships between these lexical features and the behaviour of model predictions. For instance, pronouns of different grammatical persons and numbers contribute differently to tasks involving closeness: first-person pronouns are influential across all settings; second-person pronouns are particularly significant in private conversations; and third-person pronouns mainly affect the perception of closeness in private contexts. Additionally, our results emphasize the importance of spelling variations, often overlooked as linguistic noise, including non-standard forms of pronouns and sentence-final particles, as well as other words written in morphophonemic variations. Lastly, our embedding projection study shows that the models typically represent pronouns as signals of increased closeness and decreased respect. Its embeddings also retain a consistent distribution pattern even when spelling variations are introduced, albeit with a minor shift towards more closeness and less respect suggesting that spelling variation functions as an intensifier of the social meaning. Collectively, these results affirm that encoder-only language models effectively encode and use linguistic information, especially sociolinguistic clues in the lexical features, to a considerable extent.

References

Yonatan Belinkov. 2022. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Joseph R Cooke et al. 1989. Thai sentence particles: forms, meanings and formal-semantic variations. In *Papers in Southeast Asian Linguistics No. 12: Thai sentence particles and other topics*. Pacific Linguistics.

Ian Covert, Scott M Lundberg, and Su-In Lee. 2020. Understanding global feature contributions with additive importance measures. *Advances in Neural Information Processing Systems*, 33:17212–17223.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

Gráinne M Fitzsimons and Aaron C Kay. 2004. Language and interpersonal cognition: Causal effects of variations in pronoun usage on perceptions of closeness. *Personality and Social Psychology Bulletin*, 30(5):547–557.

John J Godfrey, Edward C Holliman, and Jane McDaniel. 1992. Switchboard: Telephone speech corpus for research and development. In *Acoustics, speech, and signal processing, ieee international conference on*, volume 1, pages 517–520. IEEE Computer Society.

Michael Alexander Kirkwood Halliday. 1978. *Language as social semiotic: The social interpretation of language and meaning*, volume 42. Edward Arnold London.

Shirley Anugrah Hayati, Dongyeop Kang, and Lyle Ungar. 2021. Does BERT learn as humans perceive? understanding linguistic styles through lexica. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6323–6331, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yuphaphann Hoonchamlong. 1992. Some observations on phom and dichan: Male and female 1st person pronouns in thai. *Papers on Tai Languages, Linguistics, and Literatures: In Honor of William J. Gedney on his 77th Birthday*, 16:195–213.

Ewa Kacewicz, James W Pennebaker, Matthew Davis, Moongee Jeon, and Arthur C Graesser. 2014. Pronoun use reflects standings in social hierarchies. *Journal of Language and Social Psychology*, 33(2):125–143.

Thomas J Knutson, Rosechongporn Komolsevin, Pat Chatiketu, and Val R Smith. 2003. A cross-cultural comparison of thai and us american rhetorical sensitivity: Implications for intercultural communication effectiveness. *International Journal of intercultural relations*, 27(1):63–78.

Austin C Kozlowski, Matt Taddy, and James A Evans. 2019. The geometry of culture: Analyzing the meanings of class through word embeddings. *American Sociological Review*, 84(5):905–949.

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

Lalita Lowphansirikul, Charin Polpanumas, Nawat Jantrakulchai, and Sarana Nutanong. 2021. Wangchanberta: Pretraining transformer-based thai language models. *arXiv preprint arXiv:2101.09635*.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

Hosein Mohebbi, Ali Modarressi, and Mohammad Taher Pilehvar. 2021. Exploring the role of BERT token representations to explain sentence probing results. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 792–806, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Christoph Molnar. 2018. A guide for making black box models explainable. *URL: https://christophm. github. io/interpretable-ml-book*, 2(3):10.

Pakawat Nakwijit and Matthew Purver. 2022. Misspelling semantics in Thai. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 227–236, Marseille, France. European Language Resources Association.

Pakawat Nakwijit, Attapol T. Rutherford, and Matthew Purver. 2024. The language of closeness and respect in thai conversations: An analysis of lexical features and spelling variations. Unpublished.

Wannaphong Phatthiyaphaibun, Korakot Chaovavanich, Charin Polpanumas, Arthit Suriyawongkul, Lalita

Lowphansirikul, Pattarawat Chormai, Peerat Limkonchotiwat, Thanathip Suntorntip, and Can Udomcharoenchaikit. 2023. Pythainlp: Thai natural language processing in python. *arXiv preprint arXiv:2312.04649*.

Cate McKean Poynton. 1991. *Address and the semiotics of social relations: A systemic-functional account of address forms and practices in Australian English*. phdthesis, University of Sydney.

Marco Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 97–101, San Diego, California. Association for Computational Linguistics.

H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Lukasz Dziurzynski, Stephanie M Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin EP Seligman, et al. 2013. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one*, 8(9):e73791.

Lloyd S. Shapley. 1952. *A Value for N-Person Games*. RAND Corporation.

Panyut Sriwirote, Jalinee Thapiang, Vasan Timtong, and Attapol T Rutherford. 2023. Phayathaibert: Enhancing a pretrained thai language model with unassimilated loanwords. *arXiv preprint arXiv:2311.12475*.

Hanne Surkyn, Reinhild Vandekerckhove, and Dominiek Sandra. 2021. Social media data as a naturalistic test bed for studying sociolinguistic and psycholinguistic patterns in verb spelling errors. In *of the 8th Conference on Computer-Mediated Communication (CMC) and Social Media Corpora (CMC-Corpora 2021)*, volume 559, page 90.

Lucas Torroba Hennigen, Adina Williams, and Ryan Cotterell. 2020. Intrinsic probing through dimension selection. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 197–216, Online. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, et al. 2021. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*.

Tongshuang Wu, Marco Tulio Ribeiro, Jeffrey Heer, and Daniel Weld. 2021. Polyjuice: Generating counterfactuals for explaining, evaluating, and improving models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6707–6723, Online. Association for Computational Linguistics.

Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. 2024. Explainability for large language models: A survey. *ACM Transactions on Intelligent Systems and Technology*, 15(2):1–38.

A Conversation Corpus

The corpus was originally introduced by Nakwijit et al. (2024). It is designed to explore how lexical features interact with social relationships in private and public settings. The construction of the corpus is detailed in the following subsections.

A.1 Setting 1: Private-Self

The authors set up a messaging platform, and crowdsourced participants aged 18-30 to create a chat room and invite another participant for a conversation. Participants selected a seeding topic from the Switchboard corpus (Godfrey et al., 1992) and conversed on this topic for at least 30 turns. After the conversation, they privately filled out a form to identify their relationship in terms of closeness and respect, choosing from *Intimate*, *Close*, *Acquainted*, *Unfamiliar*, *Dislike*, and *Cannot describe* for closeness, and *Highly Respectful*, *Respectful*, *Normal*, *Disrespectful*, and *Cannot describe* for respect.

A.2 Setting 2: Public-Perceived

The authors collected tweets from X (formerly Twitter) based on 53 popular hashtags in 2022. Those tweets were filtered and selected with at least two replies. Each conversation was annotated by three recruited native Thai-speaking teenagers (16-18 years old), who assessed the degree of closeness and respect perceived in the conversation with the same set of labels presented in Setting 1. Each conversation was presented as a dialogue between an initiator (A) and a responder (B), withholding any identifying information about both individuals. Annotators were instructed to provide labels from the perspective of the responder (B). Only conversations with at least two annotators in agreement were retained; the rest were discarded.

A.3 Setting 3: Private-Perceived

The author re-annotated private conversations from Setting 1 by the annotators from Setting 2. The

same procedure and labeling scheme as in Setting 2 were applied during this re-annotation process.

B Lexical Features

Our analysis consider only three lexical features; pronouns, sentence-final particles and spelling variations. The selection of these features was guided by their prominence in sociolinguistic literature, particularly in relation to social factors such as gender, age, and social status in both English and/or Thai.

Pronoun was chosen as it is a well-studied lexical feature known for their social functionality across many languages (Hoonchamlong, 1992; Fitzsimons and Kay, 2004; Kacewicz et al., 2014). Their frequent use and significant role in communication make them a critical feature as a reference baseline.

Sentence-final particle was included because it represents a lesser-known social-related feature. These particles have limited studies due to their observation in a narrower range of languages, primarily East and Southeast Asian languages (Cooke et al., 1989).

Lastly, spelling variation was selected as it represents a recent linguistic pattern that has gained recognition for its potential semantic functions (Surkyn et al., 2021; Nakwijit and Purver, 2022). There are few studies on spelling variations, especially in Thai. Importantly, in this paper, spelling variation is specifically highlighted because of its increasing prevalence in modern conversations driven by the internet and social networks. By examining it, we aim to raise awareness of its importance in contemporary linguistic analysis.

C Social Relationship Models

Here is a detailed description of the input features for our models:

Naive Bayes Classifier: We used word count as input features, discarding terms with a frequency of less than five.

Logistic Regression: We used 15 lexical features as input features, including the number of unique words, number of Thai words, number of long words (more than 7 characters), number of out-of-vocabulary words, number of 1st person pronouns, number of 2nd person pronouns, number of 3rd person pronouns, number of pronouns in non-standard spellings, number of socially-related particles, number of non-socially-related particles, number of sentence-final particles in non-standard spellings, number of common misspelt words, number of morphophonemic variations, number of abbreviations, and number of repeated characters,

For each conversation, we examined each word and identified its lexical type using a dictionary-based approach. The dictionaries for each lexical type were provided by the authors of the corpus. We counted the number of words corresponding to each lexical feature. Finally, the values for each lexical feature were normalized by the total number of words in the conversation.

D Human Validation

In our validation in section 5.3, we intentionally recruited participants aged 18-20. This decision was made to closely match the age range of the participants in the original corpus.

We acknowledge that this decision introduces a bias, potentially affecting the interpretation of results, as language usage can vary across different age groups. However, this age group was our target population because they have grown up with text-only communication technology and are familiar with internet slang and variations, making them ideal candidates for validating our experiments.

During the annotation process, each word was presented without context. The annotators were asked the following question: "ตอบในมุมมองของคนที่ใช้คำๆนี้ในบทสนทนา ถ้าเห็นเขาใช้คำๆนี้แล้ว คิดว่าเขามีความสัมพันธ์อย่างไรกับคนที่เขากำลังพูดด้วย " (translation: Answer from the perspective of the person using this word in the conversation. When you see them using this word, what do you think their relationship is with the person they are speaking to?).

E SHAP Value from LMs

The tables below present the average of absolute SHAP values across all tokens for three lexical features (pronoun, sentence-final particles, spelling variation) in three conversational settings. Values highlighted in grey indicate those exceeding 10% of their respective random baselines, which are calculated from the SHAP values of 100 randomly selected tokens.

E.1 Fine-tuned XLM-R

E.1.1 Closeness

Lexical Features	Setting 1 Private-Self		Setting 2 Public-Perceived		Setting 3 Private-Perceived	
	Per token	Total	Per token	Total	Per token	Total
Reference						
Average per token	1.07	123.94	1.80	65.14	0.58	67.67
Pronoun						
All pronoun	0.80	3.64	1.29	3.46	0.25	1.13
» 1st person pronoun	0.78	2.14	1.21	2.20	0.23	0.63
» 2nd person pronoun	0.87	2.78	1.23	2.67	0.28	0.88
» 3rd person pronoun	0.49	1.02	1.02	1.92	0.20	0.41
» Singular pronoun	0.80	3.63	1.27	3.37	0.25	1.13
» Plural pronoun	0.23	0.23	2.33	3.00	0.11	0.11
» Pronoun in non-standard spelling	0.47	0.96	1.33	2.01	0.23	0.45
Sentence-final Particles						
All particles	2.98	22.07	1.86	4.27	3.39	25.04
» Socially-related particles	7.11	29.19	2.68	4.40	8.35	34.35
» Non-socially-related particles	0.60	2.98	1.29	2.29	0.51	2.53
» Particle in non-standard spelling	0.85	1.43	2.44	2.81	0.92	1.54
Spelling Variation						
All spelling variation	1.27	23.45	1.65	9.98	0.55	10.13
» Common misspelt words	0.96	1.69	1.50	2.19	0.20	0.35
» Morphophonemic variation	1.70	18.67	2.09	7.44	0.79	8.63
» Simplified variation	0.68	6.22	1.30	5.24	0.24	2.21
» Repeated characters	0.53	1.14	1.50	1.97	0.15	0.32

E.1.2 Respect

Lexical Features	Setting 1 Private-Self		Setting 2 Public-Perceived		Setting 3 Private-Perceived	
	Per token	Total	Per token	Total	Per token	Total
Reference						
Average per token	0.22	25.50	1.37	50.69	0.30	34.33
Pronoun						
All pronoun	0.18	0.83	2.07	6.11	0.16	0.74
» 1st person pronoun	0.15	0.43	1.68	3.16	0.16	0.46
» 2nd person pronoun	0.20	0.64	2.72	6.25	0.17	0.56
» 3rd person pronoun	0.16	0.33	0.77	1.59	0.14	0.30
» Singular pronoun	0.18	0.83	2.10	6.10	0.16	0.74
» Plural pronoun	0.15	0.15	0.64	0.84	0.08	0.08
» Pronoun in non-standard spelling	0.17	0.34	0.75	1.22	0.12	0.26
Sentence-final Particles						
All particles	0.52	3.86	0.96	2.29	0.21	1.53
» Socially-related particles	1.12	4.61	1.40	2.27	0.24	1.00
» Non-socially-related particles	0.17	0.86	0.70	1.36	0.19	0.92
» Particle in non-standard spelling	0.19	0.31	1.03	1.22	0.20	0.34
Spelling Variation						
All spelling variation	0.20	3.63	0.91	5.80	0.19	3.53
» Common misspelt words	0.22	0.39	1.01	1.45	0.19	0.33
» Morphophonemic variation	0.21	2.33	1.04	3.91	0.21	2.32
» Simplified variation	0.19	1.75	0.82	3.48	0.18	1.64
» Repeated characters	0.19	0.40	0.28	0.37	0.19	0.40

E.2 Fine-tuned WangChanBERTa

E.2.1 Closeness

Lexical Features	Setting 1 Private-Self		Setting 2 Public-Perceived		Setting 3 Private-Perceived	
	Per token	Total	Per token	Total	Per token	Total
Reference						
Average per token	1.34	156.00	2.91	105.40	1.17	135.70
Pronoun						
All pronoun	1.51	5.42	3.68	7.72	1.92	6.78
» 1st person pronoun	1.61	3.69	4.51	6.76	1.67	3.77
» 2nd person pronoun	1.91	4.83	3.76	6.66	2.41	6.05
» 3rd person pronoun	0.94	1.74	2.21	3.57	1.87	3.44
» Singular pronoun	1.52	5.42	3.72	7.73	1.92	6.77
» Plural pronoun	0.46	0.46	1.32	1.76	0.24	0.24
» Pronoun in non-standard spelling	0.90	1.92	6.02	7.91	1.72	3.43
Sentence-final Particles						
All particles	2.87	14.46	3.30	5.99	1.51	7.64
» Socially-related particles	5.24	16.26	3.43	4.91	2.58	8.02
» Non-socially-related particles	1.43	5.00	3.21	5.03	0.86	3.03
» Particle in non-standard spelling	1.98	2.77	7.20	7.94	1.08	1.51
Spelling Variation						
All spelling variation	1.39	18.25	3.36	14.89	1.08	14.23
» Common misspelt words	1.09	1.69	3.14	4.33	1.22	1.89
» Morphophonemic variation	1.64	13.66	4.21	11.83	1.19	9.90
» Simplified variation	1.08	7.00	2.69	7.98	0.87	5.67
» Repeated characters	0.64	1.37	3.39	4.44	0.41	0.88

E.2.2 Respect

Lexical Features	Setting 1 Private-Self		Setting 2 Public-Perceived		Setting 3 Private-Perceived	
	Per token	Total	Per token	Total	Per token	Total
Reference						
Average per token	1.49	173.20	2.16	80.16	0.46	53.21
Pronoun						
All pronoun	3.57	12.84	2.64	6.11	1.11	4.06
» 1st person pronoun	3.76	8.59	1.86	2.92	1.24	2.89
» 2nd person pronoun	4.17	10.54	3.25	6.05	1.00	2.57
» 3rd person pronoun	3.28	6.10	1.92	3.43	0.44	0.82
» Singular pronoun	3.59	12.85	2.66	6.11	1.11	4.06
» Plural pronoun	0.55	0.55	1.09	1.34	0.22	0.22
» Pronoun in non-standard spelling	3.62	7.52	2.00	2.89	1.07	2.39
Sentence-final Particles						
All particles	1.53	7.77	2.26	4.35	0.49	2.47
» Socially-related particles	2.02	6.27	3.16	4.58	0.69	2.14
» Non-socially-related particles	1.24	4.35	1.67	2.86	0.37	1.29
» Particle in non-standard spelling	1.79	2.52	2.45	2.87	0.39	0.55
Spelling Variation						
All spelling variation	1.91	25.11	1.93	8.86	0.46	6.01
» Common misspelt words	1.53	2.37	2.26	3.12	0.49	0.76
» Morphophonemic variation	2.29	19.08	2.16	6.40	0.51	4.27
» Simplified variation	1.45	9.36	1.64	5.07	0.38	2.44
» Repeated characters	0.66	1.42	1.28	1.69	0.18	0.39