Understanding Annotator Perception: Modeling Psychological Inference from First- and Third-Person Annotations (Student Abstract)

Gerard Christopher Yeo[1,2], **Kokil Jaidka**[2,3]

[1] Institute for Data Science, National University of Singapore
[2] Centre for Trusted Internet & Community, National University of Singapore
[3] Department of Communications and New Media, National University of Singapore
e0545159@u.nus.edu

Abstract

Large language models (LLMs) are trained on vast amounts of publicly available text. However, the current training frameworks take for granted that these annotations are accurate reflections of the authors' true intents. This study questions that assumption by examining the gaps between writers' actual psychological states and the inferences made by third-party annotators. We explore how readers interpret psychological cues in text and demonstrate that third-person annotations often fail to align with first-person realities. By integrating both first- and third-person annotations, we develop computational models that reveal significant biases in how psychological states are perceived and the downstream effects these perceptions have on reader behavior. Our findings challenge the foundational assumptions of LLM training, suggesting that the reliance on potentially flawed third-person annotations could impact model accuracy and real-world applications.

Introduction

Large language models (LLMs) rely on vast, publicly sourced datasets where inferred third-person annotations of psychological states are assumed to be accurate representations of the authors' actual experiences. However, these assumptions often overlook the discrepancies between what writers truly experience (first-person experience) and how readers interpret those experiences from text (third-person inference). This gap is particularly concerning given the widespread use of such models in real-world applications where understanding human emotions and intentions is crucial (Hennig-Thurau, Walsh, and Walsh 2003).

This study examines how well readers' inferences align with writers' self-reported psychological states, by developing models that integrate both first- and third-person annotations. Autobiographical texts, which often include rich emotional and cognitive cues (Gygax, Garnham, and Oakhill 2004), serve as the basis for our analysis. We further assess how these inferences influence readers' future behavioral intentions, such as product purchases (Cheng and Ho 2015; Kim and Gupta 2012).

Our research highlights the discrepancies between inferred and actual psychological states, focusing on how third-party raters often misinterpret certain cognitive appraisals, and the implications this has for models leveraging these labels. This work underscores a broader observation in the field: while LLMs excel at tasks related to valence, they are challenged by tasks requiring deeper reasoning or understanding of complex psychological motivations.

Method

Data Collection

In the first part of this study, we collected first-person annotations, which we published as the Perception and Experience of Appraisals and Consumption Emotions in Reviews (PEACE-Reviews) dataset (Yeo and Jaidka 2023) with 1,400 self-annotated reviews of purchase experiences. Participants were screened based on whether they had recently purchased expensive products or services, ensuring that their responses would reflect emotionally significant experiences. Each participant provided detailed descriptions of their experiences with both positive and negative emotions, following structured prompts that elicited emotional intensity, appraisal dimensions, and behavioral intentions. These self-reported first-person annotations were used to establish ground-truth labels for the subsequent analysis.

In the second part, we have now supplemented the PEACE-Reviews dataset with third-person annotations. To gather these, we trained and recruited 365 participants from Amazon Mechanical Turk to rate the review texts along various psychological and emotional dimensions (cognitive appraisals, emotions and behavioral intentions), to obtain two independent sets of annotations on each of the 1,400 reviews. We also asked participants to provide ratings for consumer-related variables such as "helpfulness" and "trustworthiness" of the review. After qualifying participants through a task designed to ensure high-quality annotations, we obtained two independent ratings per review. These third-person ratings were compared to the first-person ground truth to evaluate how well external raters could infer the psychological states of the original writers. Inter-rater reliability was calculated to ensure consistency between the annotators. For subsequent analyses involving third-person ratings, we averaged the ratings provided by the two raters.

Variable	BERT		RoBERTa		Best_{fp}
	Acc	F1	Acc	F1	Acc / F1
Appraisals					
Accountability-circumstances	57.6	0.46	55.4	0.48	
Accountability-other	46.8	0.43	54.0	0.50	
Accountability-self	56.1	0.51	53.9	0.50	
Attentional activity	41.0	0.33	43.2	0.36	
Certainty	57.5	0.48	56.1	0.52	
Control-circumstances	46.8	0.40	42.4	0.33	
Control-other	48.9	0.48	48.9	0.45	
Control-self	51.1	0.46	53.2	0.49	
Coping potential	51.1	0.47	49.6	0.44	
Difficulty	58.3	0.57	57.5	0.52	
Effort	56.1	0.55	56.1	0.51	
Expectedness	60.4	0.57	59.7	0.55	50.8 / 0.49
External normative significance	58.3	0.53	60.4	0.55	
Fairness	59.7	0.55	61.9	0.57	
Future expectancy	54.7	0.49	52.5	0.47	
Goal conduciveness	59.7	0.53	64.0	0.58	70.5 / 0.71
Goal relevance	62.6	0.52	60.4	0.53	68.3 / 0.68
Novelty	46.8	0.41	46.0	0.39	
Perceived obstacle	58.3	0.53	54.7	0.49	
Pleasantness	59.7	0.56	61.8	0.58	73.4 / 0.74
Emotions	-	0.61	-	0.66	72.1 / 0.71
Behavioral Intentions					
Intent to promote	60.4	0.56	58.3	0.54	72.4 / 0.66
Intent to repurchase	61.9	0.57	59.7	0.55	70.1 / 0.61
Consumer-related variables					
Helpfulness	60.4	0.54	58.3	0.45	
Trustworthiness	59.0	0.50	56.1	0.45	

Table 1: Results of baseline 3-way multi-class classification of third-person variables for BERT and RoBERTa models. The values in the last column was obtained in Yeo and Jaidka (2023) using the BERT model.

Experimental Setup

We conducted two sets of experiments: (1) modeling baseline third-person reasoning and comparing against the first-person models reported in Yeo and Jaidka (2023), and (2) modeling the influence of third-person ratings on predicting first-person variables based on theoretical formulations, following the multi-task framework in Yeo, Furniturewala, and Jaidka (2024). We primarily used the Bidirectional Encoder Representations from Transformers (BERT) model, fine-tuned for these tasks, so that it could be effectively compared against the baselines we reported in our analysis of the first-person annotations.

Results and Future Work

An analysis of the missing data suggests that while appraisals related to valence, such as pleasantness and goal conduciveness, are consistently observed with the lowest percentage of missing values, appraisals involving external agency, such as accountability to others or circumstances, exhibit higher percentages of missing data, suggesting that these appraisals are less clearly conveyed. Future expectancy also has a relatively high percentage of missing values, implying that readers struggle to infer future-oriented appraisals from the predominantly present-focused narratives.

This has implications for the degree of correspondence between first-person experiences and third-person inferences. We then compute the correlation between the third-person ratings and the first-person ratings. Appraisals related to pleasantness, goal conduciveness, and fairness showed medium-to-large correspondence, while dimensions like accountability-circumstances had the lowest correlation. For behavioral intentions, readers were able to infer future behaviors better than cognitive appraisals (r = .47). Altogether, The results suggest that the correspondence between first- and third-person ratings are at most moderate.

Multi-task models	Acc	F1	Best_{fp}
Text -> Appraisals -> Emotions (theoretical multi-task)	-	0.64	
Text -> Appraisals -> Intention to promote	63.3	0.58	72.1 / 0.65
Text -> Emotions -> Intention to promote	62.6	0.58	73.6 / 0.66
Text -> Appraisals + Emotions -> Intention to promote	64.0	0.58	73.4 / 0.71
Text -> Appraisals -> Intention to repurchase	61.9	0.57	69.3 / 0.58
Text -> Emotions -> Intention to repurchase	62.6	0.57	68.6 / 0.58
Text -> Appraisals + Emotions -> Intention to repurchase	62.6	0.57	69.3 / 0.58

Table 2: Results of multi-task models for predicting first-person emotions and behavioral intentions using third-person ratings. The values on the last column represent the model performance using first-person ratings to predict first-person variables in Yeo, Furniturewala, and Jaidka (2024). "Appraisals" refer to cognitive appraisals and "Emotions" refer to predicted emotional states.

Table 1 reports the predictive performance of using third-person annotations to train classifiers. We see that the BERT models trained on first-person annotations (reported in Yeo and Jaidka (2023)) consistently outperform the best third-person models, particularly in variables like goal conduciveness and intent to promote, where the F1 score difference is substantial while it is the least for emotion prediction, where the drop in accuracy ranges from 5% for RoBERTa to 11% for the BERT model. In pleasantness, the first-person model achieves a 0.74 F1 score compared to 0.56 and 0.58 for BERT and RoBERTa, respectively.

We also conducted multi-task models to use third-person annotations to predict first-person labels about their emotions and behavioral intentions. Table 2 shows that including cognitive appraisals and emotions in the models improved performance, with an F1 score of 0.64 for emotion prediction and higher accuracy for predicting intentions to promote and repurchase compared to the baseline. Yet, as reported in the last column, in comparison to the framework using first-person labels in Yeo, Furniturewala, and Jaidka (2024), we see an accuracy drop ranging from 6-11%.

Our results reveal that while annotators can moderately infer emotions and intentions, they struggle more with cognitive appraisals, especially those involving external circumstances. This has ramifications for the models we train, which tend to struggle particularly with nuanced appraisals like goal relevance and pleasantness. Attributes like Expectedness are unexpectedly better inferred through third-person than first person annotations, suggesting that they are simply difficult to express in text.

For future work, our findings suggest further investigating the reasons behind the discrepancies we observed, and examining other social media domains. We will also explore techniques to enhance third-person inference accuracy, such as leveraging multi-modal approaches and incorporating context-aware annotations. The code to replicate our experiments is included in the supplementary materials.

Acknowledgments

This work was supported by the Singapore Ministry of Education AcRF TIER 3 Grant (MOE-MOET32022-0001).

References

Cheng, Y.-H.; and Ho, H.-Y. 2015. Social influence's impact on reader perceptions of online reviews. *Journal of Business Research*, 68(4): 883–887.

Gygax, P.; Garnham, A.; and Oakhill, J. 2004. Inferring characters' emotional states: Can readers infer specific emotions? *Language and Cognitive Processes*, 19(5): 613–639.

Hennig-Thurau, T.; Walsh, G.; and Walsh, G. 2003. Electronic word-of-mouth: Motives for and consequences of reading customer articulations on the Internet. *International journal of electronic commerce*, 8(2): 51–74.

Kim, J.; and Gupta, P. 2012. Emotional expressions in online user reviews: How they influence consumers' product evaluations. *Journal of Business Research*, 65(7): 985–992.

Yeo, G.; Furniturewala, S.; and Jaidka, K. 2024. Beyond Text: Leveraging Multi-Task Learning and Cognitive Appraisal Theory for Post-Purchase Intention Analysis. In *Findings of the Association for Computational Linguistics ACL 2024*, 12353–12360.

Yeo, G.; and Jaidka, K. 2023. The PEACE-Reviews dataset: Modeling Cognitive Appraisals in Emotion Text Analysis. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 2822–2840.