Abstract Review Ratings

Print Date: August 1, 2023

Abstract Information		
Abstract No.	**Abstract Name**	**Form Name**
337	Caregivers Attitude Detection From Clinical Notes	AMIA 2023 Annual Call for Submissions

Reviewer Summary	
Reviewer #1	
Comments to the author.	- This paper explores whether some of the Hugging Face models can be used for sentiment analysis in notes written by clinicians. There is a fair bit of novelty in this paper, and it would be nice to discuss at AMIA. My primary reservations are related to rationale for cohort selection and then the potential mis-interpretation of results in the tables (see below). - There were several missing words related to the "Section" at the end of the Introduction. - The background information related to using sentiment analysis techniques from social media provides a strong rationale for the importance of this study. - it's unclear why the authors selected neonatal patients who lived less than 1 year. - With the low agreement by human annotators, it's possible we're asking too much of language models - it was interesting just to see the level of disagreement among humans in this paper. - I'm glad to see BLOOM was assessed - we need more findings reported for this model in academia. - It's good to see your transparency about the differences in labeling options for the various models. - I don't understand why the authors reported greater performance of all models in Table 2 (few-shot) compared to Table 1 (zero shot). For example, micro-F1s in Table 1 were 0.7604 for ROBERTa while the highest in Table 2 was 0.7128 for MiniLM and BLOOM. I'd encourage the authors to look at their results again and/or clarify their statement.
Reviewer #2	
Comments to the author.	The abstract mentions the major topic areas where the submission has had an impact, but it does not provide specific details or results about the impact. Including some key findings or outcomes related to usability, efficiency, clinical decision support, interoperability, etc., would strengthen the impact aspect of the abstract. Overall, the abstract provides a clear overview of the study on detecting caregivers' attitudes from clinical notes.
Reviewer #3	
Comments to the author.	Important work which evaluates caregiver attitudes from clinical text data and differentiates it from routinely used sentiment analysis.Addresses important issues such as burnout amongst caregivers. Technically, well done using Transformers and appropriate libraries.
Reviewer #4	
Comments to the author.	Discussion will need to be engaging and at multiple audience levels since the writing is deeply computational. interesting study into clinician attitudes with penetrating, sharp, and smart research. Understanding and detecting the tolls of clinician burden /burnout early, is integral to ensuring the best possible patient care. Looking forward to watching the evolution of this project.
Reviewer #5	
SPC Comments to the Author	Utilizing Natural Language Processing (NLP) approaches to identify caregiver attitudes is relevant and important for the patient's care journey. This paper employed MIMIC clinical notes upon annotation. The Hugging Face platform has been used for sentiment analysis with the following models: DistilBERT, RoBERTa, MiniLM, and BLOOM. The results include zero-shot, few-shot and full-trained approaches with RoBERTa providing the best result. Although, paper is well written and organized, cohort and model selection and inter-rater agreement interpretation should be explained better.
Reviewer #6	

Abstract Review Ratings

Print Date: August 1, 2023

Reviewer Summary	
Comments to the author.	1. The paper stated that "Results indicate a moderate level of agreement among the five annotators. Fleiss' kappa, Cohen's kappa, and Krippendorff's alpha all produced agreement values in the range of 0.257-0.264, with an average agreement value of 0.260. This suggests that while the agreement between the annotators was not perfect, it was still reasonable and acceptable for the task." However, for Cohen's kappa, values = 0 as indicating no agreement and 0.01–0.20 as none to slight, 0.21–0.40 as fair, 0.41– 0.60 as moderate, 0.61–0.80 as substantial, and 0.81–1.00 as almost perfect agreement. (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3900052/). Not sure what the "moderate level of agreement" definition is used here, and this does not logically prove that "This suggests that while the agreement between the annotators was not perfect, it was still reasonable and acceptable for the task." 2. Model selection. The paper stated that "state-of-the-art language models from the Hugging Face platform", however, they failed to justify how the SOTA models was selected. There are models trained on clinical notes, such as ClinicalBERT, which may be proper for this study. Although the author mentioned that "In this work, we used the most popular sentiment analysis models, both pre-trained and fine-tuned, available on Hugging Face", they failed to prove that these selected models were SOTA.