Human or LLM annotators? A quantitative evaluation of manual and automatic APPRAISAL analyses of ATTITUDE in the genre-specific corpus of TED talks

Mirela Imamovic, Silvana Deilen, Dylan Glynn & Ekaterina Lapshinova-Koltunski

(University of Hildesheim and University of Paris 8; University of Hildesheim; University of Paris 8; & University of Hildesheim)

Abstract

APPRAISAL theory (Martin and White 2005) is a proposal for the linguistic description of evaluation. One of its primary aims is to account for the implicit and explicit expression of ATTITUDE. Despite wide application and success as a manual tagset to date, there have been few attempts at automating its annotation. The two biggest hurdles are identifying evaluative items in text and then, in turn, tagging them using the original APPRAISAL scheme which includes many subtle, yet important, features. Previous research into modifying APPRAISAL theory has found that the automatic distinction of ATTITUDE types (JUDGEMENT and APPRECIATION) is difficult because of shared linguistic patterns (Bednarek 2009). Imamovic et al. (2024) showed that ChatGPT fails in correctly classifying sub-categories of ATTITUDE using the original APPRAISAL scheme. In this study, the primary aim is to identify which APPRAISAL features systematically cause the most difficulties for both human and machine identification. This will allow future research to focus on which categories are most problematic and warrant revision. Further, it will enable the development of a gold standard for the categories where agreement is more likely (Read and Caroll 2012). Following Fuoli's (2018) step-by-step method, we modify the sub-categories of the original APPRAISAL scheme and design our APPRAISAL guidelines for the human and machine annotators. The data employed are controlled for domain and gender variations and consist of 5 English transcripts of TED talks, each of them annotated by 2 coders. The study uses multiple coders – 10 undergraduate university students with a linguistic background and completed training in APPRAISAL Theory whose annotation is then replicated with the use of LLM ChatGPT Turbo 4. The coding results of the human annotators (gold standard) and the automatic annotation by machine are then compared. Using Cohen's Kappa (Cohen 1968) and regression modelling, we determine which features result in the highest levels of disagreement. We then examine the disagreement at different levels of granularity and report on the most problematic set of categories. We expect to achieve quantified and repeatable results that will enable a clear and substantial set of modifications for the APPRAISAL coding scheme.

Keywords

APPRAISAL annotation, Cohen's kappa, LLM, Regression, TED talks

References

Bednarek, M. 2009. Language patterns and ATTITUDE. *Functions of Language*, 16 (2), 165-192.

Cohen, J. 1968. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70, 213-220.

Fuoli, M. 2018. A stepwise method for annotating appraisal. *Functions of Language,* 25 (2), 229-258.

Imamovic, M., Deilen, S., Glynn, D., Lapshinova-Koltunski, E. 2024. Using ChatGPT for Annotation of Attitude within the Appraisal Theory: Lessons Learned. In *Proceedings of The 18th Linguistic Annotation Workshop (LAW-XVIII),* 112–123, St. Julians, Malta. Association for Computational Linguistics.

Martin, J., White, P. R. 2005. *The Language of Evaluation.* New York: Palgrave Macmillan.

Read, J., Carroll, J. 2012. Annotating expressions of Appraisal in English. *Language Resources and Evaluation*, 46 (3), 421–447.