Evaluating AI Sentiment Analysis

Aakriti Shah
Rollins College – Department of Mathematics and Computer Science

01. Introduction

The focus of this paper is an analysis of human and AI coding of qualitative transcripts for Crave, a leadership development program serving social entrepreneurs in Orlando. The program's focus is on spiritual formation, peer support, and developing leaders to support the non-profit space. The study compares categorical coding of statements taken from program discussion transcripts by an expert annotator (Dr. Myers), human annotators (3 undergraduate students), and two AI models: Claude (Anthropic) and Bing (Microsoft). The comparative analysis aims to evaluate AI sentiment analysis capabilities relative to human performance and identify areas for improving machine understanding.

02. Methods

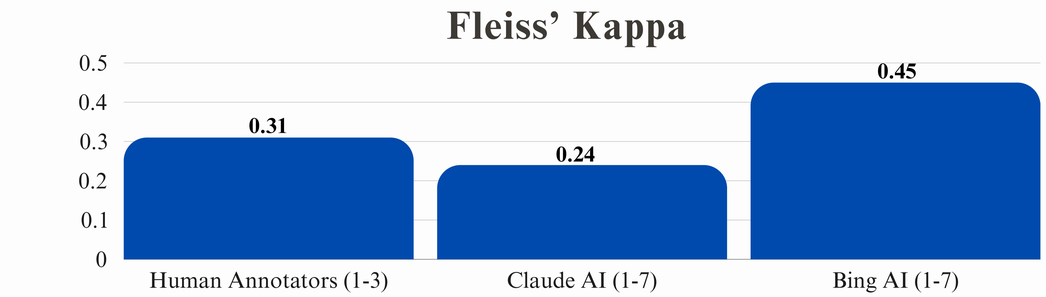
02.01. Data Collection

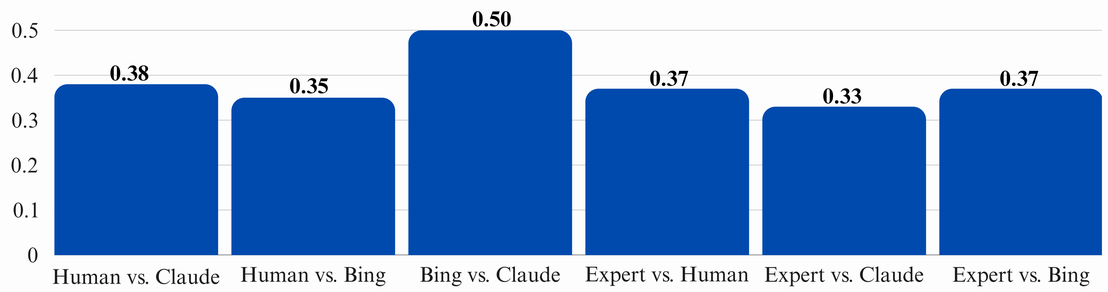
The text data was sourced from transcripts of group discussions and Ripple Effect Mapping (REM) sessions conducted to assess the impact of Crave, a nonprofit leadership development program for young professionals and social entrepreneurs engaged with spirituality and social change in Orlando. The REM sessions were facilitated by the Rollins College Impact Lab in 2021 and involved structured dialogue between Crave participants, alumni, staff and facilitators. For this analysis, 107 statements were extracted from the Crave session transcripts covering topics of community capitals including culture, society, politics, economics and spirituality in Central Florida.

02.02. Annotation Process

Three undergraduate students independently hand-coded each of the 107 extracted statements using the Community Capitals Framework, which has 8 categories: Natural, Cultural, Human, Social, Political, Financial, Built and None. The students were provided with the framework definitions and labeled each statement with one of the codes or None. The students then came to a consensus on the appropriate label for each statement to create a human "truth" benchmark. The students then gave the same statements, capital definitions, and coding instructions to the Claude and Bing AI systems. Each system generated predicted labels for the statements across multiple runs, resulting in 7 runs for Claude and 7 runs for Bing. This produced a human truth set, an expert annotated set, and modal predicted labels from the AIs for quantitative comparison.

03. Results





04. Analysis

The inter-annotator agreement and AI model performances demonstrate promising but imperfect capabilities in classifying capital statements compared to human judgments. While Claude edged out Bing in alignment with human labels, both models struggled significantly in identifying prevalent Societal (Social & Human Capital) examples, achieving only 50-70% accuracy despite the critical importance of these codes. Perfect Natural Capital classification meant little next to mediocre performance on the most societally impactful elements. The models aligned strongly only on negligible Natural cases while disagreeing substantially on essential Human Capital categorization, severely inhibiting reliability.

Classification confusion frequently occurred between similar codes like Social vs Human and Cultural vs Political Capital, which depend heavily on interpretative context. Even human coders disagreed in categorizing these prominent codes, reflected in the fair inter-annotator consensus. Still, improving disambiguation capabilities should remain a priority given the centrality of Societal codes on overall capital calculations and analysis.

While AI training across copious background data cannot replace specialized exposure needed to expertly classify niche conceptual interactions, the models' capabilities surpass human limitations in select cases. Evident human annotator oversights further demonstrate the value of AI to strengthen analysis by detecting overlooked signals. Optimally combining AI efficiencies with expert oversight balances productivity and quality for ideal analytical outcomes.

05. Takeaways

1. AI does not agree with its own performance, so single runs are not reliable, even through the aggregation of data (modal columns).
   a. Therefore, there is no sufficient reason to believe that the AI models are performing at a comparable level as the expert annotator.
2. AI struggled to classify Human vs. Social & Cultural vs. Political Capitals.
   a. "You don't have to present or code switch; you get to the real and get better at the real"
   b. "We felt like lab rats when the topic of racism came to the forefront"

06. Next Steps

1. Testing the AI models on larger and more diverse data may improve the accuracies.
2. Investigating if providing more context while prompting the AI models lead to higher agreement with human annotators – examples, emotional appeals, more information about the Community Capitals.

07. Acknowledgements

- Principle Investigator, Dr. Dan Myers
- Coresearchers, James McIntyre and James Temple
- Nonprofit Organization, Crave
- Rollins College Honors Program and Department of Math and Computer Science

08. References

McHugh, M. L. (2012). Interrater reliability: the kappa statistic. Biochemia Medica, 22(3), 276–282. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3900052/

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436 444. https://doi.org/10.1038/nature14539

Amershi, S., Begel, A., Bird, C., DeLine, R., Gall, H., Kamar, E., Nagappan, N., Nushi, B., & Zimmermann, T. (2019). Software engineering for machine learning: A case study. 2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP), 291–300. https://doi.org/10.1109/ICSE-SEIP.2019.00042

Haskell, M., Mehdinia, S., & Myers, D. S. (2021). Crave of CentralFlorida: A leadership development program for the "spiritually curious" [White paper]. Rollins College Community Impact Lab.

Crave. (n.d.). https://cravefla.org/