# Sentiment and Politeness Analysis Tools on Developer Discussions Are Unreliable, but so Are People

Nasif Imtiaz, Justin Middleton, Peter Girouard, Emerson Murphy-Hill
{simtiaz, jamiddl2, plgiroua, ermurph3}@ncsu.edu

## Motivation

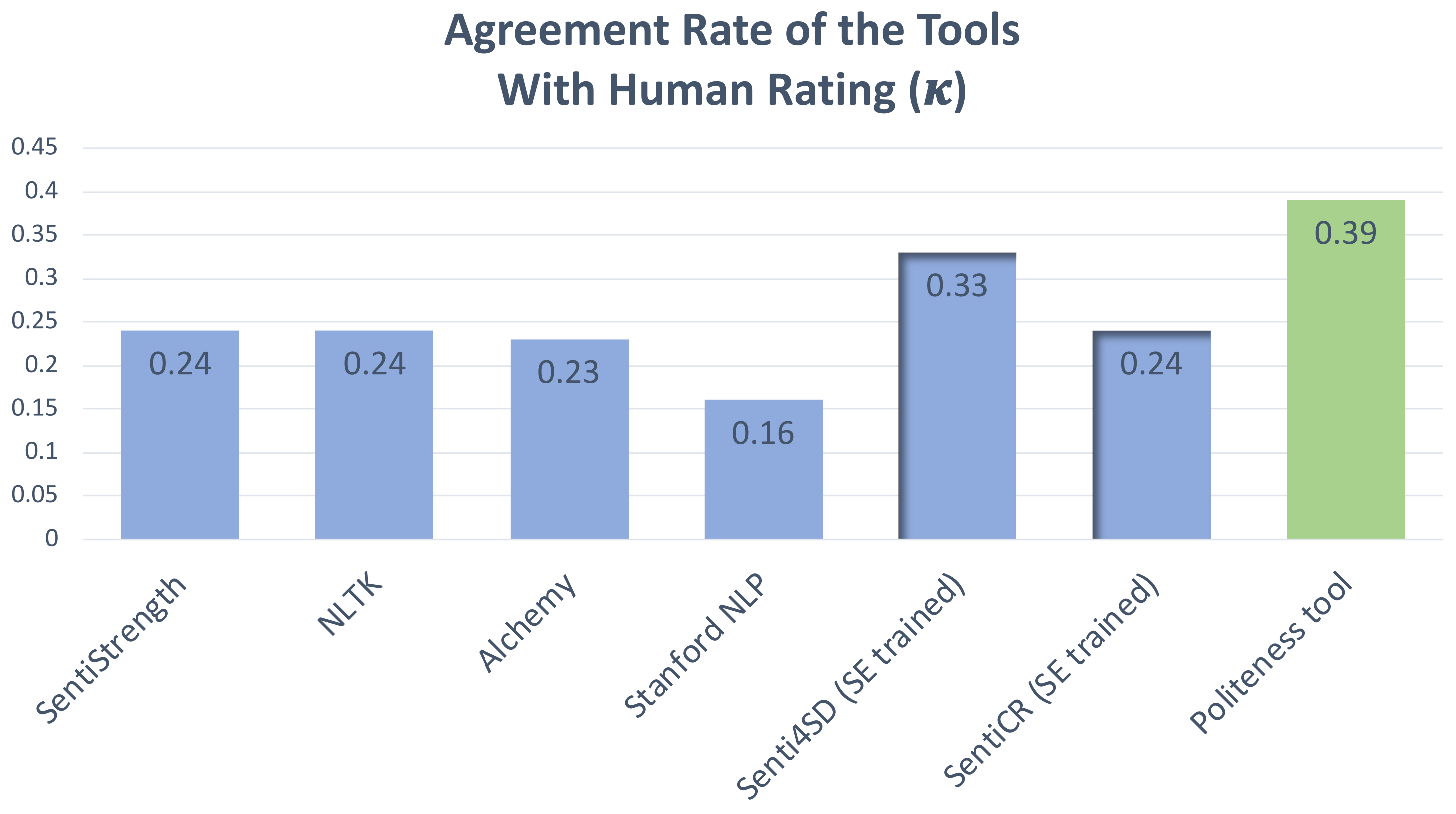
- Emotions affect developers' productivity[1]
- Developers interact over online platforms
- Researchers study emotion in the software engineering domain[2]
- Sentiment and Politeness analysis tools are common approaches[3]
- How reliable are these tools?
- How reliable are the study?

## Annotation Scheme

- Prior studies use different or no annotation schemes
- We use a simple sentiment annotation scheme (positive, neutral, negative) developed by Mohammad[4]
- We developed a scheme[5] on how to rate politeness (polite, neutral, impolite) in textual conversations
- Based on Brown & Levinson's Politeness and Culpeper's Impoliteness theory

## Results: Tools are Unreliable!

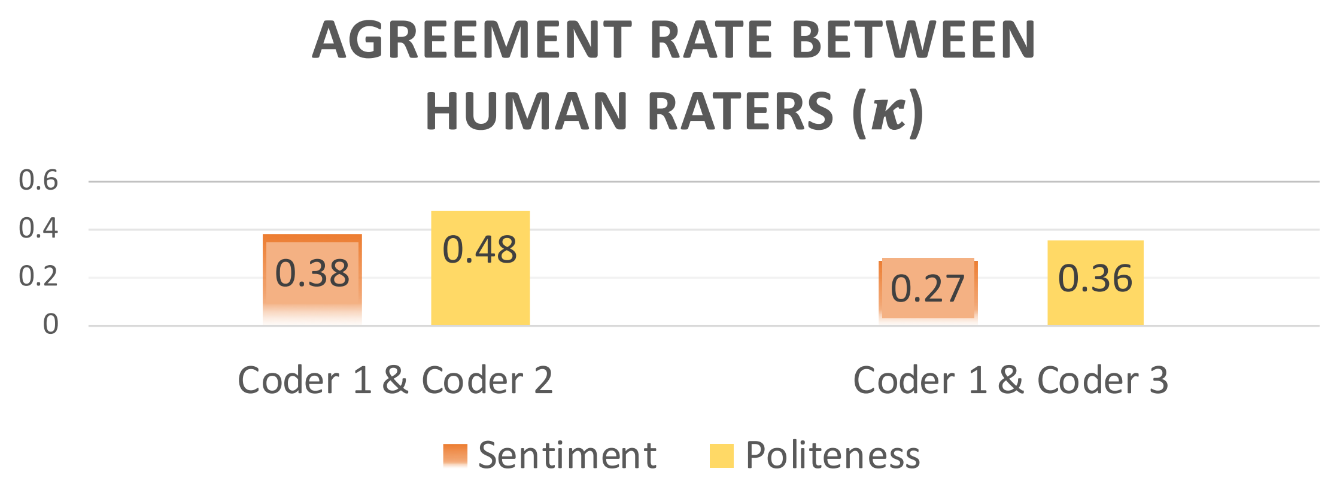
- 589 comments from GitHub
- 6 sentiment tools
- 1 politeness tool
- Tools compared against human rating
- Cohen's kappa to measure tools' agreement with humans



- All tools but one have fair agreement (0.2-0.4)
- "Not Reliable" as per the community standard
- Tools have varying precision and recall over different classes

## Results: Even Humans are Unreliable!

- Each comment rated individually by 2 coders
- Coders had fair and moderate agreements
- Unreliable consistency



- Disputed ratings resolved through discussion
- Need domain specific standardized annotation scheme for consistency

## Reference

1. Iftikhar Ahmed Khan, Willem-Paul Brinkman, and Robert M Hierons. 2011. Do moods affect programmersâĂŹ debug performance? Cognition, Technology & Work 13, 4 (2011), 245–258.
2. Robbert Jongeling, Proshanta Sarkar, Subhajit Datta, and Alexander Serebrenik. 2017. On negative results when using sentiment analysis tools for software engineering research. Empirical Software Engineering (2017), 1–42.
3. Marco Ortu, Bram Adams, Giuseppe Destefanis, Parastou Tourani, Michele Marchesi, and Roberto Tonelli. 2015. Are bullies more productive?: empirical study of affectiveness vs. issue fixing time. In Proceedings of the 12th Working Conference on Mining Software Repositories. IEEE Press, 303–313
4. Saif Mohammad. 2016. A Practical Guide to Sentiment Annotation: Challenges and Solutions.. In WASSA@ NAACL-HLT. 174–179.
5. https://git.io/vpVKO

Developer Liberation Front

NC STATE UNIVERSITY
Department of Computer Science

National Science Foundation