Evaluating the utility of generative AI to categorize the sentiment of polarizing content on social media

Dhiraj Murthy[1], Jeremy Shermak[2], Shaurya Pathania[3], Kami Vinton[4], Kelsey Whipple[5], Lesley Willard[6]
[1]School of Journalism and Media, Moody College of Communication and Department of Sociology, [2]Columbia College Chicago, [3]Computational Media Lab, University of Texas, [4]School of Journalism and Media, Moody College of Communication, [5]College of Social land Behavioral Sciences, University of Massachusetts Amherst, [6]Emerson College

GOOD SYSTEMS
Ethical AI at UT Austin

The University of Texas at Austin
Research Development
Vice President for Research, Scholarship and Creative Endeavors

Columbia COLLEGE CHICAGO

Introduction

- Polarized political social media posts are complex for most machine learning models
- These posts cause difficulty for models and humans alike
- Human and Machine analysis have benefits and drawbacks
- Sentiment Analysis of polarized political social media content via machine learning has been established, but remains understudied in the context of generative AI

Research Goals

1. What sentiment do Chat GPT 3.5 and humans detect in polarized political discourse on Twitter?
2. To what degree do humans and Chat GPT 3.5 coding agree on the sentiment for polarized political discourse on Twitter?
3. What themes, topics and rhetorical strategies are employed in polarized political discourse on Twitter where human and machine sentiment scores disagree most frequently?

Methods

- Collected 6 millions polarizing tweets between 2016 and 2017 using Twitter's REST API
- Selected a random sample of 1,000 tweets from each year
- Coded each tweet with both humans and Chat GPT 3.5
- Constructed a Human versus Machine Spread table to measure the agreements between the coders
- Negative values represent more disagreement and positive represent more agreement

Levels of Human vs. Machine Spread (HMS) and Agreement Degree/Direction

HMS Value	Degree and Direction of Sentiment Agreement
4	Largest possible disagreement where human is more positive than machine
3	Large disagreement where human is more positive than machine
2	Significant disagreement where human is more positive than machine
1	Slight disagreement where human is more positive than machine
0	Agreement between human and machine
-1	Slight disagreement where machine is more positive than human
-2	Significant disagreement where machine is more positive than human
-3	Large disagreement where machine is more positive than human
-4	Largest possible disagreement where machine is more positive than human

Category	Description
@mentions	Number of @mentions that appear in the tweets
Hashtags	Number of hashtags that appear in the tweets
All Capital Letters	1 = Any word is in ALL CAPS 0 = No single word is in ALL CAPS Excludes typical acronyms such as "FBI" or "DACA"
Profanity	Defined broadly and included "hell," "damn," and "crap," as well as symbols standing in for words to get around online profanity filters (e.g., "f*ck"). Misspellings, abbreviations, or approximations were also counted. (Chen, Shermak, Brown, Reidl, Tenenboim 2017) 1 = Profanity present 0 = No profanity present

Results

Sentiment scores for machine vs. humans (as percentages)

Sentiment Score	Chat GPT	Human
0 Very negative	0.0%	3.8%
1 Negative	58.5%	41.9%
2 Neutral	24.1%	46.0%
3 Positive	17.3%	7.3%
4 Very positive	0.0%	1.0%

Comparisons of Human vs ChatGPT coding

Description	Human Sentiment	ChatGPT Sentiment
RT @zerohedge: USDJPY Spooked Below 113.00 On Trump-Abe Comments Regarding China https://t.co/RbUxKqz8Z1	2	2
RT @_Siya_Ngubo: @_Siya_Ngubo & people want to deny the fact that Trump inherently won the election because of his racism & bigotry.	1	2
RT @Crosscut: .@cmkshama said there was renewed urgency under Trump to protect private citizens. https://t.co/3N3JXtjQt3	3	3

Human sentiment correlation to tweet attributes (Spearman's rho)

	Sentiment Machine	Sentiment	mentions	hashtags	all caps	profanity	misspelling	retweet	sarcasm	expunct	links	abbreviations
Correlation Coefficient	.033	1.00	.038	-.010	-.063	-.220 **	-.145 **	-.002	-.200 **	.057	.162 **	-.115 **
Sig.	.318	-	.256	.763	.061	0.00	0.00	.956	0.00	.091	0.00	.001
N	891	892	892	892	892	892	892	892	892	892	892	892

** - Correlation is significant at the .01 * - Correlation is significant at the .05

Results

RQ1 - Analyzing the dominant sentiment of human coders and ChatGPT

- GPT does not fully think like a human, even if it scores are based on what it learned from millions of human inputs
- Human Coder Average Sentiment Score = 1.56
- Human Median Score = 2 & Human Mean Score = 2
- Chat GPT 3.5 Average Sentiment Score = 1.59
- GPT Median Score = 1 & GPT Mean Score = 1

RQ2 - Understanding how well machines and humans agree or disagree with each other

- Humans to GPT had 30.8% (Cohen's kappa) agreement
- 2 Humans got a score of 32.8% (Cohen's kappa) agreement
- Both scores are very low and ChatGPT is not far behind humans in this complex task

RQ3 - Determining what themes, topics, and items cause the most disagreement between the two

- Sarcasm, profanity, misspellings, and abbreviations are negatively correlated for humans
- URLs/links are positively correlated for humans
- Reasons GPT could not evaluate some of the tweets due to short tweets lengths (<10 characters), URLs/links, and non-keyboard characters

Conclusion

- Generative AI methods do not fully understand the complexities of polarized political content without human input
- Human Sentiment Analysis is more nuanced than ChatGPT as humans can better understand complexity
- Our work urges caution in solely using ChatGPT-based sentiment classification for polarized political social media
- Increased generative AI and human comparisons can better AI and make them more useful in this use case in the future
- Humans should are not the gold standard either as humans also face challenges evaluating polarized political social media content
- Generative AI needs to improve further before it can be relied on to interpret the sentiment of polarized political social media content