Democratizing Text Analysis? A Critical Evaluation of Google's Gemini Pro for Sentiment Classification

Anthony Eryan, Kyle Rothwell, Savva Petrov

Department of Humanities, Arts and Social Sciences, Stevens Institute of Technology

HSS 317: Quantitative Social Science

Professor Paul Connor

May 10, 2024

Democratizing Text Analysis? A Critical Evaluation of Google's Gemini Pro for Sentiment Classification

The conduction of text analysis/sentiment analysis using large language machine models (LLMs) such as Google's Gemini Pro to compare accuracy across sentiments of over 1,000 tweets can determine whether LLMs are trustworthy in the interconnected field of text analysis, such as psychology, sociology, political science, and computer science. Such a topic is an essential interdisciplinary field to study as historically, a significant portion of human thought, ideas, and interactions are expressed textually, ranging from modern mediums such as emails and social media posts (in this case, tweets) to classics such as news articles, legal documents, books, and scripts that encapsulate our thoughts and feelings as humans. Furthermore, understanding emerging social trends of slang, cultural shifts, values, and the formation of online communities has boomed thanks to the massive strides in the computer science field with Natural Language Processing (NLP), which is the interdisciplinary field in which artificial intelligence is used to decipher the way machines and humans interact with human languages. However recent studies like that of Perkins et. al's (2023) research, discovered that the capabilities of AI detection should be questioned as revealed that the LLM model ChatGPT-4, Turnitin Artificial Intelligence (AI) detection tool heavily overmarked submissions to claim to be AI-generated content, as only 54.8% of the content was actually AI-generated rather than 91%, revealing that AI detection tools often struggle to pinpoint exact portions of AI-generated content, especially with rigorous prompting skills from the user. Therefore, we have seen LLMs such as GPT-4 significantly advancing in producing more fluent, detailed, and highly natural tones in language. Accordingly, recent advances in LLMs, like GPT-4, might enable them to perform sentiment analysis similarly to crowdworkers.

Therefore, in the contemporary explosion of the popularity of NLPs with LLMs, we are interested in whether these LLMs can effectively perform sentiment analysis from text. Could we replace crowdworkers, resource-heavy machine learning models, and classical dictionary analyses with LLMs? If this is possible, it could allude to massive shifts in methodology and accessibility within the scientific community.

Messeri and Crockett's paper (2024) further investigates the idea of AI and the scientific community via dividing AI into four different visions: Oracle, Surrogate, Quant, and Arbiter, to further understand how exactly AI can be utilized in the scientific field. AI, as an oracle, is able to digest, summarize, and even generate unique hypotheses from the body of scientific literature. With AI as an oracle, challenges occur with having too much quantity to process, inconsistent quality, and shallow or biased understanding of what was analyzed. While AI as a surrogate is based on data collection, essentially creating synthetic datasets in both social and physical sciences, such as simulating large groups of people responding to surveys or datasets generating simulated cosmological structures, protein sequences, etc. Challenges for AI as a surrogate occur when the data itself is too complex to generate promptly or cost-effectively. AI as a quant is used to analyze larger and more complex datasets that are extremely difficult for humans to curate and analyze to extract new insights. Lastly, AI as an arbiter is used as a peer reviewer, determining or assisting human judgment not only on scientific merit and tools but also on replicability and entire scientific papers. This vision faces difficulties due to potential bias and the limits of AI's ability to fully grasp long and complex papers. Therefore, with the vision of AI as an oracle and quant, AI will be used as a large-scale data analysis tool (measuring sentiment) while having AI play a part as an oracle as each LLM is trained on a large text corpus that can spot language patterns with sentiment, essentially "digesting" it's pre-existing knowledge. However, the

distinction in our case of AI would be that there is no hypothesis generation with oracles, as our study is not interested in utilizing the LLM to generate new research questions or insights directly from the dataset. Regardless of the introspection of the AI vision in the scientific community, Messeri and Crockett's paper highlights the importance of acknowledging the potential of AI's illusion of objectivity. More specifically, in our field of research, it is important to consider AI's viewpoints or how it determines sentiments is not so black and white. In other words, it is critical to recognize that AI, or any specific LLM, does not represent every standpoint or a spotless slate in terms of ideology. Overall, we are interested in how AI is utilized as this oracle and quant hybrid in the scientific community. Rathje et al. (2023) reveal that the LLM ChatGPT's accuracy is quite impressive. GPT, on its free and premium model, outperforms dictionary analyses, and it even outperforms top-performing machine learning models or competes with them through popular languages such as English and Arabic, to "non-weird" languages which stands for non-Western, educated, industrialized, rich, and democratic countries. This notion of GPT's incredible performance in social intelligence is also backed up by the Sufyan et. al (2024) study that was conducted where LLMs were compared to their performance on the Social Intelligence Scale, a well-known benchmarking tool in the psychotherapy field. With a sample of 180 male psychologists from King Khalid University in Saudi Arabia, results revealed that GPT-4 surpassed each psychologist in the study, with Bing's LLM outperforming 90% of the bachelor-equipped psychologists, and only 50% with those of a doctorate degree. Noticeably, Google's LLM (formerly known as Bard), struggled the most with only being equivalent to bachelor-level psychologists and surpassing only 10% of doctoral holders. Consequently, we are interested in witnessing if we can replicate a similar performance towards Google's LLM, Gemini, with our 13 human-coded sentiments from our selected dataset.

Method

Our dataset sourced is from Appen (formerly known as CrowdFlower), uploaded via a public domain license on data.world, a dataset platform, on July 15th, 2016. Given that the dataset is crowdsourced, as aforementioned briefly on the platform, we compare the performance of the LLMs to the crowdsourced workers recorded on the Appen dataset. Unfortunately, there is no specific information regarding further information of the Appen crowdworkers (quantity of workers, their education level, country of resignation, etc), therefore we cannot infer insights about the demographics of how our data was accumulated. Nonetheless, we used Application Programming Interfaces (APIs) to query these tweets and perform text sentiment analysis through the programming language R with our integrated development environment being RStudio. It is important to note that we have also determined a threshold of at least 400 queries in case of discrepancies associated with the number of API calls we were allowed to perform on our local machines. Google's Gemini LLM, specifically Gemini 1.0 Pro, is self-described as "The best performing model with features for a wide variety of text and image reasoning tasks." (Google, 2024) for its second-largest model that Google has to offer in the market. Additionally, Google advertises this model to compete with or perform stronger than OpenAI's ChatGPT LLM, specifically version 3.5, via multiple benchmarks of general reasoning, math, code, and image processing, which such claim alluding to a point of interest in future research regarding text-analysis related benchmarks. Nonetheless, all our prompts held the same standard: "Which sentiment best describes this text? Answer only with a number: 1 if anger, 2 if boredom, 3 if empty, 4 if enthusiasm, 5 if fun, 6 if happiness, 7 if hate, 8 if love, 9 if neutral, 10 if relief, 11 if sadness, 12 if surprise, 13 if worry. Here is the text:" with the smallest temperature recorded possible to maintain a further baseline consistency. In the context of generative AI, the

"temperature" of the LLM signifies how deterministic the AI will output, with higher values indicating more randomness. Moreover, we then compared Gemini's sentiment classification with the ground truth labels provided by the crowdworkers, with the following RStudio packages for analysis and visualization: "dplyr", which provides verb-based syntax data cleaning, filtering, transforming, and summarizing, "tidyr" to further complement data in which sentiment distribution had to be converted to a long format to create a stacked sentiment barplot. On the topic of bar plots in general, we employed "ggplot" for our data visualizations on our bar plots, and "caret" for calculating the proportion of tweets for which Gemini's sentiment label cross-matched with respect to the crowdworker label.

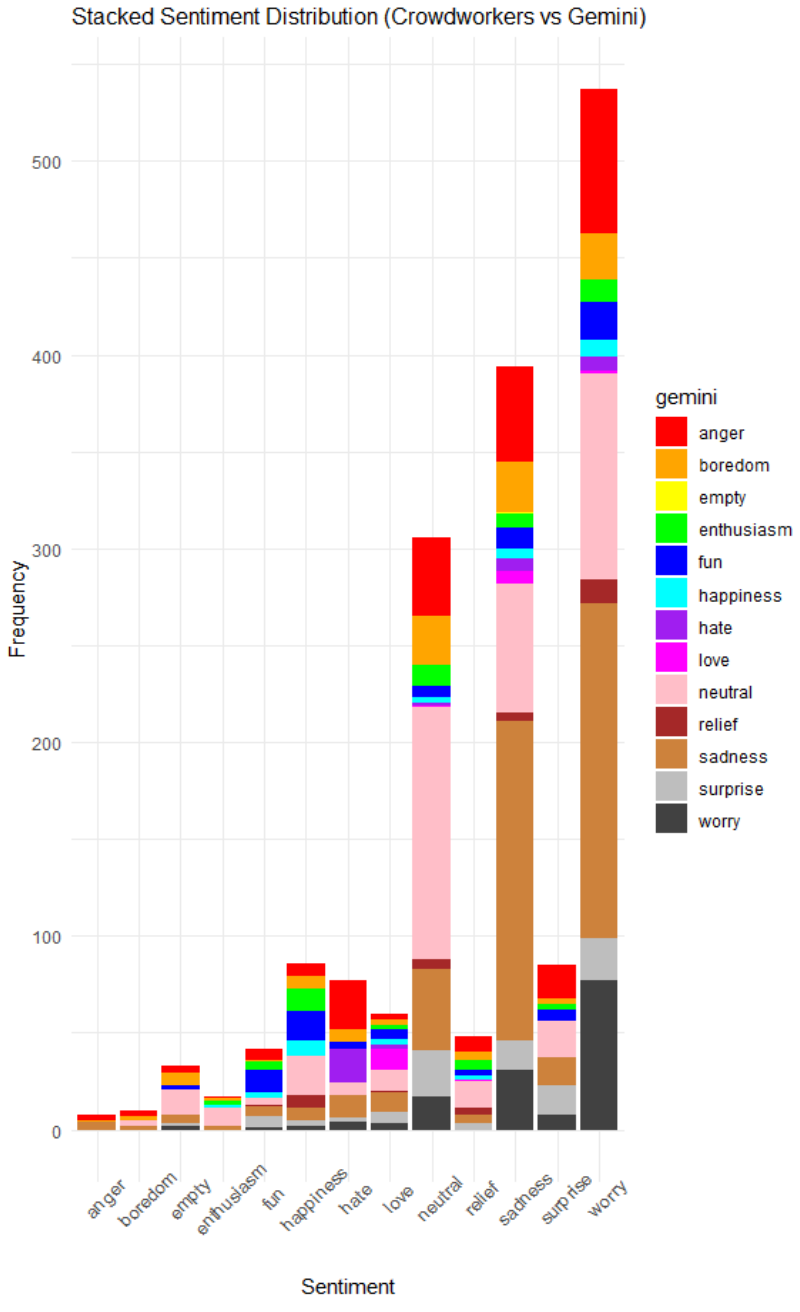
Results

Gemini's batch received was fairly generous at our best run, thus the only sampling method recorded was systematic sampling, meaning that we selected a random starting point (in this case, the first tweet) and queried through a fixed periodic interval, which in this case would be for every tenth tweet, meaning that tweets (1,11,21, … 17591) were queried from our 1703 batches out of 4,000. Given our generous batch, we utilized bar plots of the sentiments recorded from the crowdworkers and the LLM and used those data visualizations to further understand how a stacked bar plot becomes essential for an overall understanding of our performance. We chose the greatest batch from Google's "Generative Language API" (the official name of the API) and performed 1,703 tweets before unexpectedly stopping. However, Gemini refused to answer 54 tweets and was left blank, with many of these tweets notably including threats, violence, derogatory language, and lewd content. However, it would be disingenuous to say this is the root cause, as tweets that seemingly appear innocent such as tweets such as "Madly in love with The Row..wishing i had money" or the tweet "It was going to happen one day but I so feel

for the girl AND her mum" should seem harmless enough to query. This is not to be confused with the "empty" sentiment recorded, but rather Gemini refusing to answer, potentially malfunctioning, or safeguarding its results. Overall, given that our baseline is 400, this still exceeds our minimum requirement, and is quite promising from our querying method that only 3.17% of tweets were refused.

Figure 1

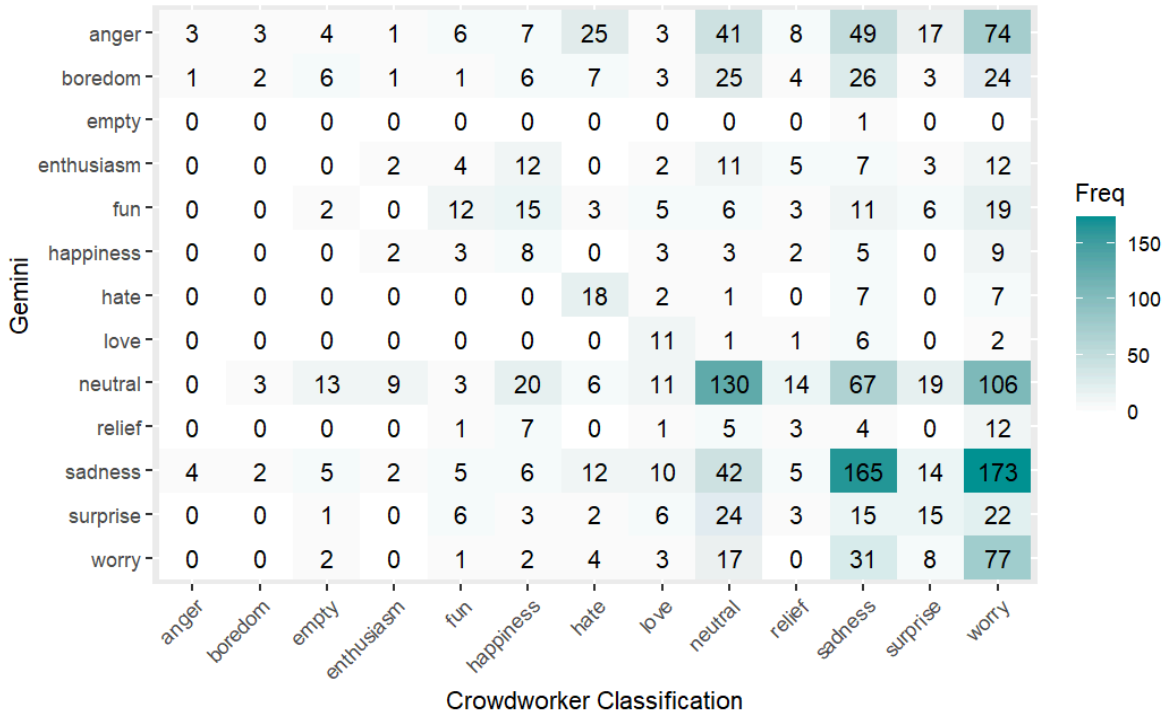
Stacked Crowdworker Sentiment Distribution (Crowdworkers vs Gemini Classification)



As we see in Figure 1, Gemini's performance holds strong variability throughout its performance. For instance, in our alphabetically first sentiment, anger, we see that Gemini leaned the majority of the time to sadness, with a slight deviation towards the emptiness sentiment, and slightly less than half correctly predicted the anger sentiment. The sentiment recorded in 1,703 tweets also shows us that even through systematic sampling, we only managed to query 8 tweets under the anger sentiment, which may explain the limited amount of sentiments recorded due to the underrepresentation through our sampling method. Furthermore, each sentiment outside of the first alphabetic few holds a considerable spread of sentiments, such as the sadness sentiment. Despite the sadness sentiment retrieving all 13 sentiments, it managed to identify the majority of the time the correct sentiment, while being the second largest tweet sentiment section. Additionally, Gemini sees further relative success in the sentiment of neutral, despite predicting 11 sentiments, and the significantly lesser of the greater two, fun, with 10 sentiments recorded.

Figure 2

Heatmap: Accuracy Between Gemini and Crowdworker Classification



An examination of the confusion matrix, a tool that further encapsulates the performance of Gemini's classification by the true sentiment, the crowdworker classification, in which it was revealed that the overall accuracy is approximately 26.19%, indicating that the model correctly predicted the sentiment for about a quarter of the tweets sampled. With the inclusion of No Information Rate (NIR), we can observe the accuracy of Gemini if it were to always guess the most common sentiment in our dataset, to assess whether Gemini's sentiment analysis is providing any significant insight compared to a model simply guessing the sentiment under the most frequent sentiment. With NIR being 26.13%, slightly less than the accuracy (26.19%), we can conclude Gemini's performance is slightly better than simply guessing the most frequent sentiment. Our Kappa score is represented as the adjusted for the accuracy that could occur by random chance, our Kappa score being 0.1437 ranges in slight agreement, echoing Gemini's unreliable performance overall. Therefore, although some success has been measured, there is still quite a long way to go with Gemini's model. With the inclusion of Figure 2, the most precise sentiments recorded were revealed to be: neutral (approximately 42.48% accurate), sadness (approximately 40.94% accurate), and worry (approximately 32.21% accurate), and with an overall accuracy of 26.19%, these findings suggest that Gemini requires significant improvement to achieve reliable sentiment classification.

Figure 3

Crowdworker Sentiment Classification: Distribution of Sentiments

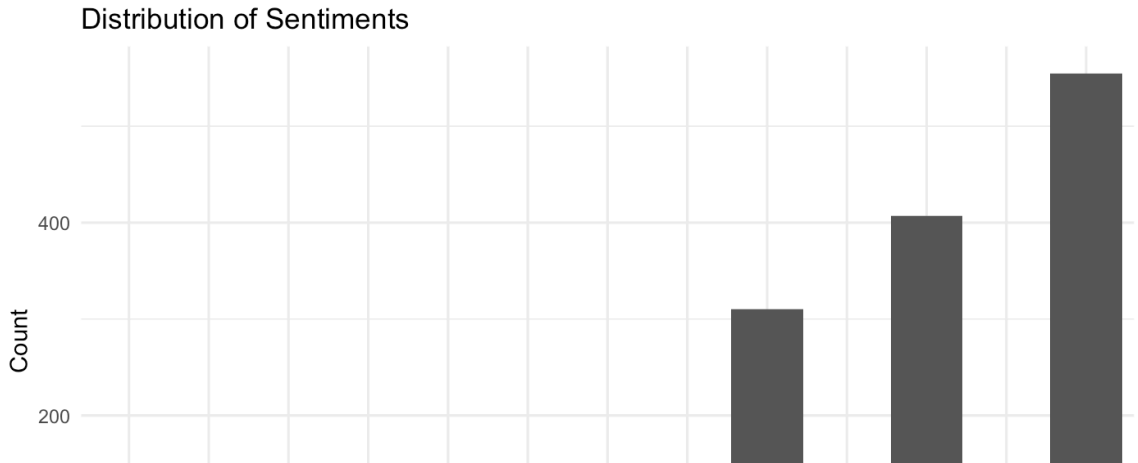


Figure 3 represents the count of each human-coded sentiment produced. Sentiments 9 (neutral), 11 (sadness), and 13 (surprise), corresponding to neutral, sadness, and worry, respectively, are the most prominently produced sentiments from the original testing method.

Figure 4

Gemini Sentiment Classification: Distribution of Sentiments

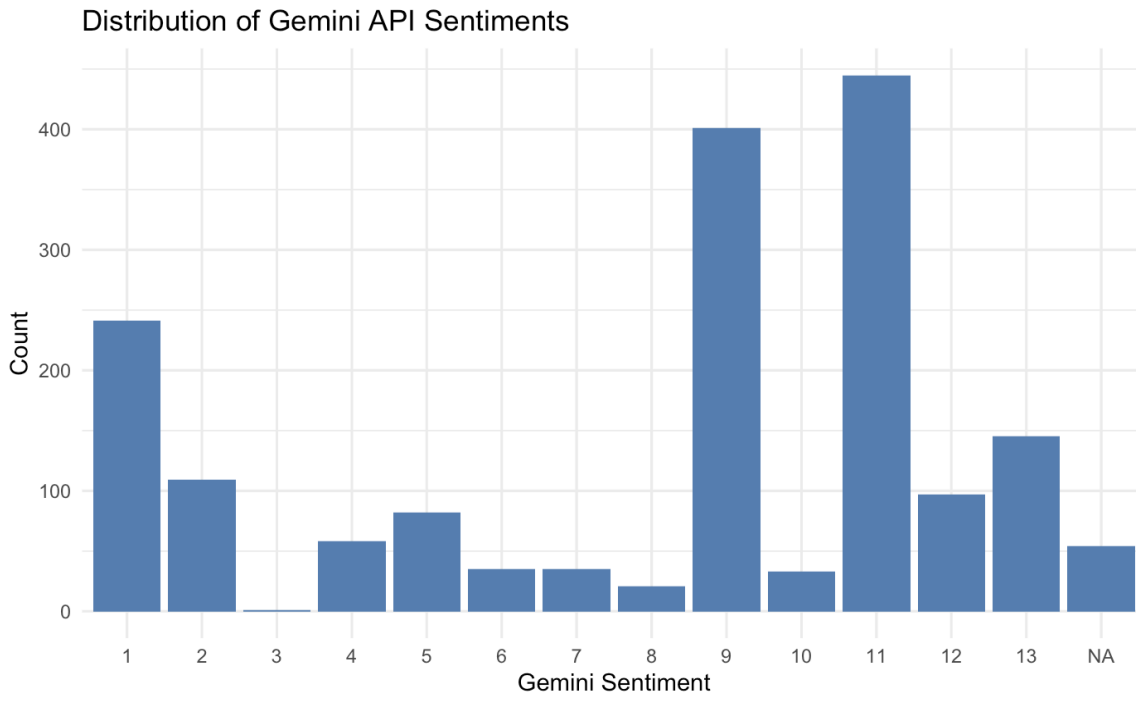


Figure 4 represents the count of each sentiment that the Gemini LLM produced when prompted by the question asking to produce a sentiment based on the tweet, with the inclusion of the 54 refused tweets by Gemini. Sentiments 1, 9, and 11, corresponding to anger, neutral, and sadness, were the most prominent.

Figure 5

Gemini vs Crowdworker Sentiment Classification: Distribution of Sentiments

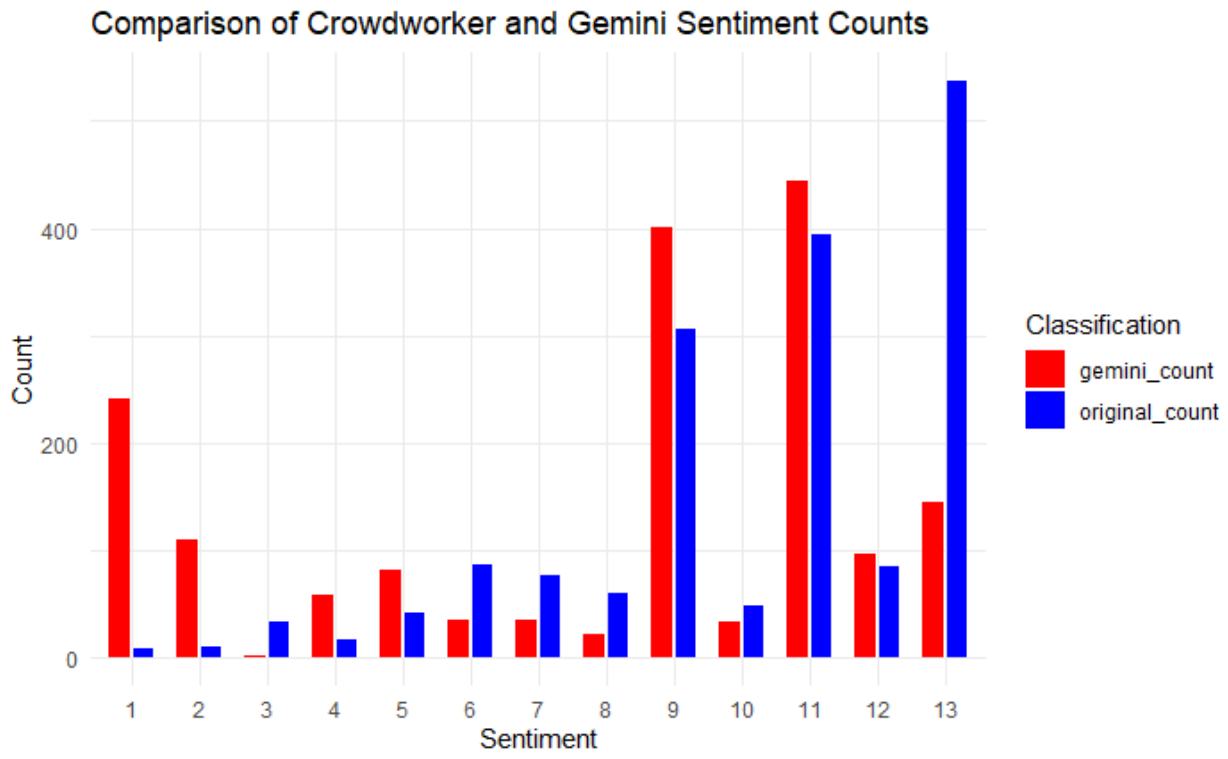


Figure 5 represents both Figure 3 and Figure 4 together. The graph shows that Gemini's count for sentiments 1 (anger), 2 (boredom), 4 (enthusiasm), 5 (fun), 9 (neutral), 11 (sadness), and 12 (surprise) was larger than the count produced by the original testing method, while sentiment counts produced by Gemini LLM for 3 (empty), 6 (happiness), 7 (hate), 8 (love), 10 (relief), and 13 (worry) were lower than the count produced by the original testing method. The most shocking disparities are for sentiments 1, 2, and 13 with Gemini's count being at 241, 109, and 145 respectively, while the count for the original testing method was 8, 10, and 555. For sentiment 1, the difference between Gemini's output and the original output is 233. For sentiment 2, the difference is 99. Finally, for sentiment 3, the difference is 410.

Chi-Squared Test

The X-squared ($X^2 = 870.63$) value is the test statistic that measures how much the observed counts deviate from the counts that would be expected if there were no associations between the variables in the contingency table. The degrees of freedom (df = 144) is calculated by taking the levels of Gemini subtracted by one and multiplying it by the levels of sentiment subtracted by one to achieve the value of 144. Although some disparity between Gemini's sentiment count and the original sentiment count was previously seen in Figure 4, the p-value evaluated at 2.2×10^{-16} indicates that the probability of observing such a strong association between Gemini and sentiment due to random chance alone is virtually nonexistent. Furthermore, the p-value is so small that it can be considered 0 in practical terms. Since the p-value is so low, we can strongly reject the null hypothesis that there is no association between Gemini and sentiment. This implies that Gemini's sentiment predictions are not independent of the true sentiments and that their distributions are associated in some way. However, these findings are discrepant with the findings from the heatmap shown in Figure 2, as the confusion matrix reveals the likelihood of a correct prediction made by Gemini is 26.19%. This low accuracy indicates that Gemini is not a reliable predictor of sentiment overall, despite the statistically significant association revealed by the chi-squared test. This is because a chi-squared Test is used to determine whether there's a significant association between two categorical variables (Gemini and sentiment), yet does not measure the strength or the predictiveness of the association. Consequently, this means that the low p-value from the chi-squared test is merely suggesting that some association exists between Gemini and sentiment, but does not actually quantify how effectively Gemini predicts sentiment. Furthermore, the accuracy and the Kappa score directly measure the predictive performance of

the model. This means that accuracy indicates the proportion of total correct predictions, while the Kappa score adjusts for the accuracy that could occur by random chance. Since these scores are low, it implies that despite there being a statistically significant association, the model is not effectively leveraging this association to make accurate predictions. This could also be due to the model having a hard time distinguishing between sadness and worry, as it seems to often misclassify these two as explicitly shown in Figure 2. Thus, this suggests the need for further analysis to understand the specific biases Gemini might have to improve its predictive capabilities.

Discussion

Through our sentiment analysis, we have observed how one of ChatGPT's competitors, Gemini, faced 1,703 tweets through various sentiments in which our results revealed Gemini's underwhelming performance overall. Our findings contribute to Rathje et. al paper in the sense the opposite was retrieved: we do not argue that LLMs, or at least Google's current Gemini Pro, can democratize automated text analysis given its performance is synonymous with that of the classical dictionary methods recorded in Rathje et al. paper, which ranges from 20% to 30% accuracy. Notably, our results are somewhat comparable to that of Sufyan et. al's psychology test conducted with Google's LLM, as it was the weakest LLM out of GPT-4 and Bing for both bachelor and doctorate degrees via a significant margin as aforementioned at the beginning of this paper. However, the comparison between our study and Sufyan et al.'s (2024) should not be taken as exact. While both studies benchmark LLMs within the social sciences, our focus is on text analysis, whereas theirs is on psychological examination. These two areas, while related, utilize distinct interdisciplinary fields and methods.

Another sign of poor performance could potentially be from the significantly greater sentiments to select from this dataset compared to Rathje et al., as their sentiment analysis contained only positive, neutral, and negative, while their discrete emotions section only had anger, joy, sadness, fear, and love, giving the LLMs greater possibilities to achieve the correct sentiment. Moreover, we could have modified our prompt to ask Gemini to determine how accurate the human-coded sentiment is (e.g. In this tweet, on a scale of 1-5, how accurate is the sentiment "anger" recorded?) effectively shifting the task from direct classification to assessing the confidence of a classification, which might be a more streamlined way to evaluate Gemini's abilities given our relatively large amount of sentiments. Additionally, Rathje et al. also asserts the concerns of test-retest reliability for LLMs, given that even at the lowest temperature (set to 0), responses still held some variance when they compared reproducibility. Other limitations included the number of samples (1703 tweets, with 54 blank) being significantly less, with Rathje et al. 47,925 samples of multi-lingual tweets and news headlines and the number of models equipped with testing, as Rathje et al. tested GPT-4, GPT-4 Turbo, and GPT-3.5. Thus, due to our limited scope of only one LLM, we can neither disprove nor approve of the claim that GPT is an effective tool for performing sentiment analysis. Therefore, potential avenues for future research would be to observe if other large LLMs from other competitors or more refined Gemini models can replicate the impressive results synonymous with Rathje et al.

References

Google. (n.d.). Google for Developers. Google. https://ai.google.dev/

Rathje, S., Mirea, D., Sucholutsky, I., Marjieh, R., Robertson, C., & Van Bavel, J. J. (2023, May 19). GPT is an effective tool for multilingual psychological text analysis. https://doi.org/10.31234/osf.io/sekf5

Messeri, L., & Crockett, M. J. (2024, March 6). Artificial intelligence and illusions of understanding in scientific research. Nature News. https://www.nature.com/articles/s41586-024-07146-0

Perkins, M., Roe, J., Postma, D., McGaughran, J., & Hickerson, D. (2023, October 31). Detection of GPT-4 generated text in Higher Education: Combining Academic Judgement and software to identify Generative AI Tool Misuse - Journal of Academic Ethics. SpringerLink. https://link.springer.com/article/10.1007/s10805-023-09492-6

Sufyan, N. S., Fadhel, F. H., Alkhathami, S. S., & Mukhadi, J. Y. A. (2024, January 22). Artificial Intelligence and social intelligence: Preliminary comparison study between AI models and psychologists. Frontiers. https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2024.1353022/full