Comparing Self-Reported Subjective Wellbeing to Sentiment Analysis Approaches to Measuring Happiness

Richard Young
University of Illinois Urbana-Champaign
Urbana, Illinois, USA
young@illinois.edu

Dongheng Li
University of Illinois Urbana-Champaign
Urbana, Illinois, USA
dli@illinois.edu

Ritul Soni
University of Illinois Urbana-Champaign
Urbana, Illinois, USA
rsoni@illinois.edu

Mohit Gupta
University of Illinois Urbana-Champaign
Urbana, Illinois, USA
mgupta@illinois.edu

ABSTRACT

In the domain of happiness and wellness studies, it has now become important to measure how close the lexical models predict user reported happiness ratings based on cognitive appraisal models. In this study, we evaluate how close are these models in operationalizing happiness. We used data from a Reddit survey on short-term and long-term subjective well-being and compared them with computer models generated ratings from large language models and lexical models after which we compared Pearson correlation between them. Our study's result indicates very weak but significant relationship between LLM-generated scores and self-reported happiness scores. But the LLMs, and valence-dominance ratings were found to show high correlation which might indicate LLMs to be using valence-dominance ratings in their training data to compute happiness scores. We found all the computational models to be weakly correlated with the self-reported values. Our study highlights how close computational models are in predicting user-rated measures of happiness and provides guidance for future research.

CCS CONCEPTS

• **Computing methodologies** → **Natural language processing**; *Machine learning*; • **Human-centered computing** → *Collaborative and social computing*.

KEYWORDS

sentiment analysis, wellbeing analysis, large language models, lexical approach to sentiment analysis, user reported sentiments

1 INTRODUCTION

Emotions are notoriously difficult to measure [1, 2]. In his classic critique of psychology and anthropology Robert Levy pointed out the failure of both disciplines to clearly articulate what an emotion was [3]. This failure has impeded interdisciplinary conversations because conceptions of emotion are often misaligned. This issue has played out in the conceptualization of happiness too. While psychologists have unified around the methods developed by the Journal of Happiness Studies [4–6], as late as 2012 anthropologists have continued to offer improvised measures that vary wildly [7]. In 2024 computational social scientists face a similar challenge in that the discipline's various sentiment analysis techniques, have driven computational social scientists to operationalize happiness in new ways [8–11]. The variety of new approaches are understandable given the novel problems that computational social scientists face. Questions remain over the alignment between these new approaches and those more commonly used by other social scientists. Although many computational social scientists regularly use traditional cognitive-appraisal measures to study emotion [11, 12], many more gravitate toward lexical measures like sentiment analysis or topic modeling when cognitive-appraisal methods (e.g. surveys) are not practical [9, 10, 13].

How we model happiness has significant implications for social scientists, social media companies, and the public. For social scientists' happiness can be used to evaluate the impact of economic and social policies [14]. For social media companies happiness can help explain user retention and engagement [11]. Operationalizing happiness is also important for understanding the mental health impacts of social media use [15, 16]. The public can be affected in additional ways too. Many people are directly influenced by the "science of happiness" as it appears in the self-help and spirituality sections of bookstores, magazine racks, online news outlets, blogs, and social media [17]. Conceptions of happiness also manifest in the workplace, where employers often view wellbeing and morale as means to a more productive or less expensive workforce [18–20]. To the degree that the lexical and cognitive-appraisal methods used to conceptualize happiness do not align, researchers risk misevaluating their discoveries.

Research Question: Does the lexical model of emotion operationalize happiness in a way that is aligned with the cognitive-appraisal model of emotion?

Motivation: The lexical model of emotion and the cognitive-appraisal model of emotion have significantly different epistemological assumptions about what emotions are and how they might be measured. If we hope to bring these approaches into dialogue it would be helpful to gauge the degree to which they are aligned or misaligned.

Approach: Our proposal is to evaluate the alignment between the Affective Norms for English Words (ANEW) dataset with long-term and short-term subjective wellbeing measures by using an informal reddit survey as a measure of ground-truth.

Research Paper, 2024,
2025. ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

2 RELATED WORK

Lexical approaches to emotion are clearly useful. They include sentiment analysis techniques and topic-modeling techniques. Researchers have used these tools to predict depression [21, 22], identify hate-speech [23, 24], model gratitude [13], and measure happiness [8–11]. Despite the popularity of these approaches Wang et al. (2012) demonstrated that Facebook's Gross National Happiness index, one example of a lexical approach to happiness, is not a valid measure for mood or wellbeing [12]. Other researchers have demonstrated very limited alignment of lexical approaches to self-reported attitudes about specific texts [25]. Boukes et al. (2020) have shown that many sentiment analysis tools don't meaningfully align with one another [26]. Atteveldt et al. (2021) have suggested that cognitive assessments of sentiment still outperform most Sentiment Analysis techniques [27]. These findings suggest that lexical sentiment analysis techniques require additional validation.

We propose to test the validity of the Affective Norms for English Words (ANEW) dataset as a measure of happiness. The creators of this dataset cognized emotion as an affective-lexical property of common english words. They asked a series of survey respondents to rank the words along three affective dimensions. Although the dataset is often deployed as a unified stand in for positive or negative emotion the Arousal dimension in particular was cognized a "happy" or "unhappy" measure [8]. Does the Arousal dimension of ANEW meaningfully align with the more commonly used cognitive assessment methods?

3 DATA

3.1 Data Collection

In 2020, Richard Young conducted an informal survey of reddit users. The survey was conducted over a five-month period on r/askreddit and consisted of two questions, posted separately. The first question was posted each day from 5/1/2020 to 7/1/2020 and the second was posted each day from 7/2/2020 to 9/30/2020. The two survey questions designed to gauge the short-term and long-term conceptions of Subjective Wellbeing:

- Q1: "On a scale of 1-10 how happy are you right now? What would it take for you to be at a 10?"
- Q2: "On a scale of 1-10 how satisfied are you with your life? What would it take for you to be at a 10?"

These questions were adapted from the Organization for Economic Co-operation and Development's (OECD) guidelines for measuring subjective wellbeing [5]. In total there were 676 unique responses to Q1 and 572 unique responses to Q2. For this project we used the Python Reddit API Wrapper (PRAW) to collect one year of prior posting history from users who responded to the informal survey in 2020. The resulting dataset includes text from subreddit submissions and comments, as well as the name of the subreddit that the submission or comment was posted to, and the sum of upvotes and downvotes that each post received. Altogether the dataset has 414,000 comments and submissions. The number of comments per user ranges from 1 to 1678. The average number of comments and submissions per user is 370. Given that the user responses to the questions in the comments were majorly a numerical digit followed by the user comment, it was necessary to separate them. Moreover, many times the users provided ranges like (8 to 8.5) as their ratings. We separated the numerical and textual responses in the data so that the dataset would have a separate column for numerical ratings for the two questions.

3.2 Descriptive Statistics

The initial dataset, prior to adding large language model and word embedding model's ratings, contained 299,542 rows with 10 columns that include information such as Username, Type (Submission or Comment), ContentTimestamp, Content, Upvotes, ContentSubreddit, Score, QuestionID, SurveyResponse, and SurveyTimestamp. The survey responses are divided between two primary questions: Q1, which asks respondents about their short-term happiness (subjective well-being), and Q2, which gauges long-term life satisfaction.

For Q1, there are 86,951 records, with 171 unique respondents, and for Q2, there are 212,591 records, with 482 unique respondents. Descriptive statistics show that the overall average score across the dataset is 5.35 on a scale from 1 to 10. Specifically, for Q1 (short-term happiness), the mean score is 4.99 with a standard deviation of 2.38, while for Q2 (long-term life satisfaction), the mean score is 5.48 with a standard deviation of 2.67. The distribution of scores indicates that most respondents rated their happiness and life satisfaction between 5 and 7, with fewer respondents choosing the extremes (1 or 10).

In addition to the survey scores, the dataset also captures Upvotes, which measure the popularity of each post. The mean number of upvotes is 34.9, but this varies widely, ranging from a minimum of -460 to a maximum of over 146,000. The average content length (number of words per comment or submission) is approximately 18.5 words, with significant variation.

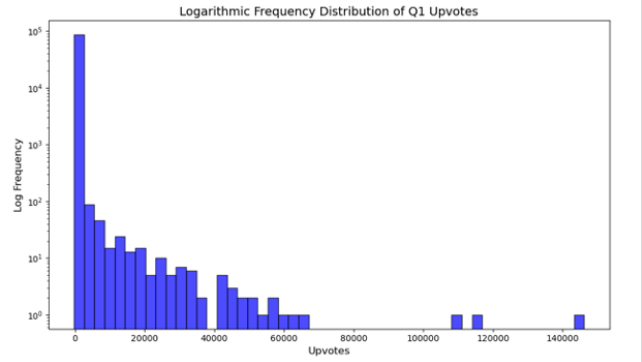
In the preliminary analysis, we found the following as top 10 most common subreddits: AskReddit (131,062), teenagers (12,883), memes (10,370), unpopularopinion (3,460), AMA (3,180), Showerthoughts (2,132), NoStupidQuestions (2,105), AskOuija (2,020), relationship_advice (1,889), and dankmemes (1,377).

The dataset structure is shown in Figure 1, which provides a snapshot of the actual data before pre-processing. The actual usernames have been replaced with dummy usernames for maintaining user privacy.

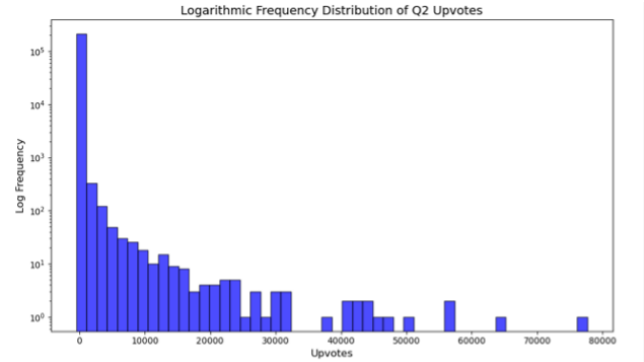
Username	Type	ContentTimestamp	Content	Upvotes	ContentSubreddit	Score	QuestionID	SurveyResponse	SurveyTimestamp
This-Username_	comment	1/10/21 16:54	They were stopped by Assassins	2	conspiracytheories	2	Q1	2 or a 3	6/30/20 20:27
This-Username_	comment	11/17/20 17:08	How many people did he kill	2	masskillers	2	Q1	2 or a 3	6/30/20 20:27
This-Username_	comment	11/15/20 18:50	I'm not dead	2	AskReddit	2	Q1	2 or a 3	6/30/20 20:27
This-Username_	comment	11/15/20 18:45	Staying inside	1	AskReddit	2	Q1	2 or a 3	6/30/20 20:27
This-Username_	comment	11/14/20 13:58	bread	1	AskReddit	2	Q1	2 or a 3	6/30/20 20:27

Figure 1: Snapshot of actual data before pre-processing. The actual username has been replaced with dummy username for maintaining user privacy.

In Figures 2, we show the logarithmic frequency distribution of the upvotes that the posts have received for Q1 and Q2. The distribution is skewed to the left with most submissions or comments receiving fewer upvotes. The upvotes might be reflecting the agreement of users who might not have participated in responding but to save time, might have done so by liking the most relevant post to their experience. Thus, it might be another interesting question to answer the relation between the upvotes, the related comment, and the score.

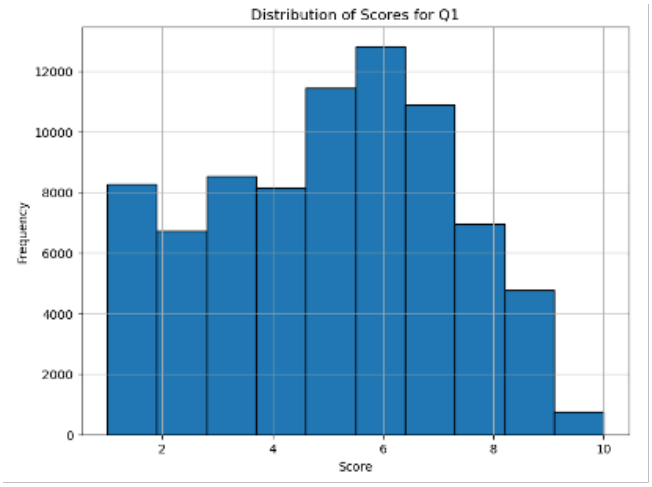


(a) Q1 upvotes distribution

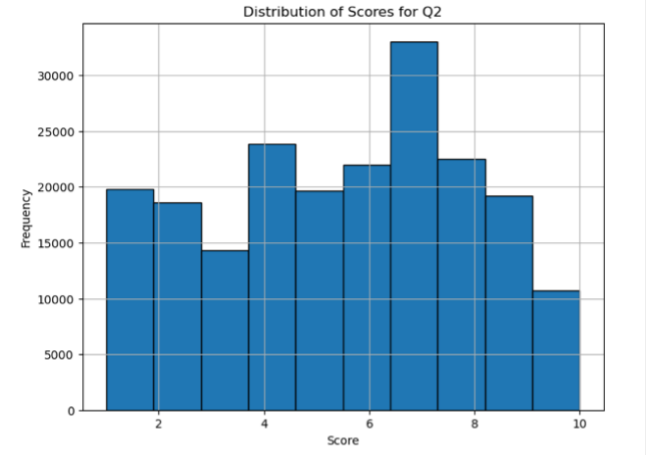


(b) Q2 upvotes distribution

Figure 2: Logarithmic frequency distribution of upvotes for (a) Q1 and (b) Q2



(a) Q1 score distribution



(b) Q2 score distribution

Figure 3: Distribution of scores for (a) short-term happiness question Q1 and (b) long-term happiness question Q2

In Figure 3, we visualize the distribution of scores of Q1 and Q2. Moreover, we created the percentage distribution of scores for Q1 and Q2, shown in Figure 4, to evaluate the distribution of scores based on proportions. For both Q1 and Q2, the middle range of scores (5-7) holds the highest percentages, indicating that most respondents rate their happiness and life satisfaction as moderate. However, Q2 shows a higher percentage of extreme scores—both at the low end (1-2) and high end (9-10)—compared to Q1. The slight lag between the two lines in the middle may give us an interpretation that in the short term, the subjects may not be happy, but they do feel overall satisfied in the longer run.

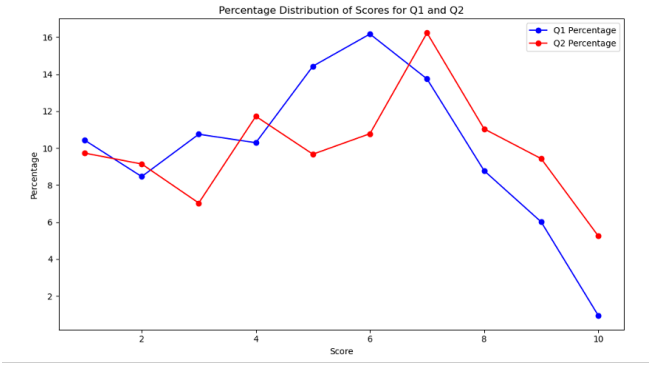


Figure 4: Percentage distribution of scores for Q1 and Q2

4 METHODS

In this section, we describe our usage of an existing LLM and a lexical-based approach to generate sentiment scores for each of the user posts in our dataset. The aim of this approach was to generate additional sentiment rating through computational approaches to enable us to compare them with user provided ratings which would help us evaluate how close are computational methods in gauging a writer's reported sentiment.

4.1 LLM Based Sentiment Scores

In this step, we asked OpenAI's GPT model 4o to rate the happiness score based on the content of the text. We created a function called `process_and_update_reddit_data` to define our batch size and provide our data and API key. We then sent each text entry to the language model with a prompt, *'You are an assistant that rates the happiness expressed in a given text on a scale from 1 to 10, where 1 is very unhappy and 10 is very happy. Only provide the numerical score'* to evoke happiness ratings from the model. We utilized batch processing and appended output to a new csv with 'LlmScore' column which has numerical scores for each text.

4.2 Lexical Model Based Sentiment Scores

In order for our us to use lexical model to assign sentiment scores we conducted a pre-processing step before the main analysis. We removed any URLs, special characters, and numbers to ensure that the textual data was clean. In order to maintain the uniformity across the text, we converted the data to lowercase, tokenized the text into words, and further filtered the data using NLTK's list of stopwords [28]. So as to maintain semantic consistency, we used WordNet's Lemmatizer to lemmatize the word tokens. Lemmatization is a process to convert a word into its base form like the base form of 'swimming' is 'swim'. These steps ensured that our dataset was ready for further processing steps.

In order to evaluate the semantic relationship between the words in our reddit data, we used a Word2Vec model [29] using Gensim. In the model, we used a skip-gram architecture (sg=1) with a vector size of 300 dimensions and the context window size of 10 words. We included the words in the vocabulary which had a minimum frequency of five so that we could filter out rare words. We performed the training over 15 epochs to make sure of optimal learning of word embeddings. We thus derived a foundational vector representation of words in the dataset which we used in the future as input for sentiment propagation.

We used Warriner et al.'s (2013) [30] norming dictionary which had valence, arousal, and dominance rating of the 13,915 English lemma words collected in a survey. All these dimensions were rated on a scale of 1 to 9 where the valence represented ratings about unpleasant to pleasant, arousal represented calm to excited, and dominance represented participant's state of feeling controlled to feeling in-control of emotions upon listening the words. We then created seed words for our sentiment propagation model where we used percentile-based thresholds to identify words with high and low scores for each dimension. We selected the words which scored about the 90th percentile for high values and the words which were below the 10th percentile for the low values. We categorized the identified seed words and verified the counts for each category to make sure that sufficient seed words were present for the propagation task. The purpose of this step was to identify the seed words with extreme values for each of the dimension as they will serve as the foundation for our sentiment propagation model where their scores, both high and low, are propagated across the word embeddings. This step serves as a crucial perquisite for initializing the SentProp algorithm that we used later.

Now, for propagating sentiment scores across the vocabulary that we built from our reddit dataset, we used SentProp algorithm [31] in multiple steps. First, we extracted the vocabulary and the corresponding word embedding from the trained Word2Vec model as our foundation step for the propagation. We then built a kNN graph, with k = 10, using cosine similarity to make sure that we connected each word to its k most similar neighbor. We then converted the sparse kNN graph to a NetworkX graph in order to map each node to its respective word. In the next step, we normalized the graph's adjacency matrix row-wise to construct a transition matrix so as to enable random walk propagation followed by which we used the high and low seed words to initialize the seed vectors to assign positive scores to high seeds and negative scores to low seeds for the dimensions of valence, arousal, and dominance. In the final steps, we used a random walk algorithm with a restart probability ($\alpha = 0.85$) to propagate sentiment scores from the seed words throughout the graph after which we mapped the final scores back to the words to create a sentiment dictionary with valence, arousal, and dominance dimensions. The aim of this stage was for us to propagate sentiment scores across the reddit vocabulary in order for us to build a domain-specific sentiment lexicon for our reddit dataset.

Lastly, we used the generated sentiment dictionary to compute sentiment scores for each user's content. We first calculated the average valence, arousal, and dominance scores for every single post by mapping the matching words in the 'CleanContent' column with their corresponding values for every single dimension. We then normalized each of the three columns on a scale of 1 to 10 using the min-max normalization method to make it easier to compare it with outer ratings from the LLM stage and the survey stage. We also grouped the data by 'Username' to find the mean values for the three dimensions per user for our final analysis.

Thus, using LLM and lexical methods, we derived two additional kinds of scores based on the user posts. Overall, we built a robust dataset with user-reported ratings, LLM-rated sentiment ratings, and lexical-method based sentiment ratings for our comparison of how similar or converging are these ratings from each other.

5 RESULTS

We computed Pearson correlations and their respective p-values between the three emotion dimensions (valence, arousal, and dominance), the LLM generated scores, and the user provided scores. We found the LLM scores and self-reported scores to have a statistically significant weak positive correlation (0.1563, p-value < 0.0001) which suggests a weak but slightly reliable relationship between the two measures. For the valence dimension, we found a weak positive correlation with self-reported scores (0.0916, p-value = 0.007) while a stronger positive correlation with the LLM score (0.4923, p-value = 6.32E-54) which indicates that LLM prediction aligns well with the valence dimension. The arousal score shows a weak correlation with the self-reported scores (-0.0963, p-value = 4.61E-03) while a negligible correlation with the LLM scores which is not statistically significant (0.0067, p-value = 0.8446). This indicates that LLM scores do not align much with the arousal scores. Lastly, the dominance shows a weak correlation with the self-reported scores (0.0791, p-value = 2.00E-02) while a moderately strong and positive correlation with the LLM scores (0.4599, p-value = 1.90E-46) which

indicates that the LLM predictions aligns well with the dominance dimension ratings.

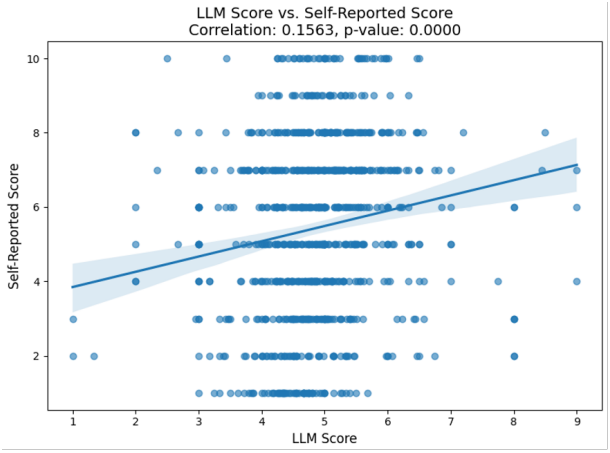


Figure 5: Positive correlation between the self-reported scores and the LLM scores

Table 1: Correlation between emotion dimensions, LLM scores and user provided scores

Dimension	Correlation	P-value
LLM Score vs Self-Reported Score	0.1563	< 0.0001
Valence vs Self-Reported Score	0.0916	7.02E-03
Valence vs LLM Score	0.4923	6.32E-54
Arousal vs Self-Reported Score	-0.0963	4.61E-03
Arousal vs LLM Score	0.0067	0.8446
Dominance vs Self-Reported Score	0.0791	2.00E-02
Dominance vs LLM Score	0.4599	1.90E-46

The correlation results are visualized in the scatter plots. As shown in Figure 6, the valence dimension shows positive correlations with both self-reported and LLM scores, with a stronger relationship to LLM scores. Figure 7 demonstrates similar patterns for the dominance dimension. In contrast, Figure 8 shows the negative correlation patterns for arousal scores.

Furthermore, to evaluate how these dimensions relate to short-term happiness (Q1) and long-term happiness (Q2), we produced scatter plots with regression lines for each of them along with their Pearson correlations. The detailed breakdown is shown in Table 2.

For short-term happiness, we found valence to have a weak positive correlation with the self-reported scores (0.1575, p-value < 0.001) but a slighter stronger correlation with LLM scores (0.2096, p-value < 0.001). The arousal score showed a weak negative correlation with the self-reported score (-0.094, p-value < 0.001) and with the LLM score (-0.054, p-value < 0.001) suggesting an inverse relationship. While the dominance consistently showed a positive relationship with self-reported (0.1996, p-value < 0.001) and LLM scores (0.1998, p-value < 0.001). For the long-term happiness, we found valence to be comparatively strongly correlated with

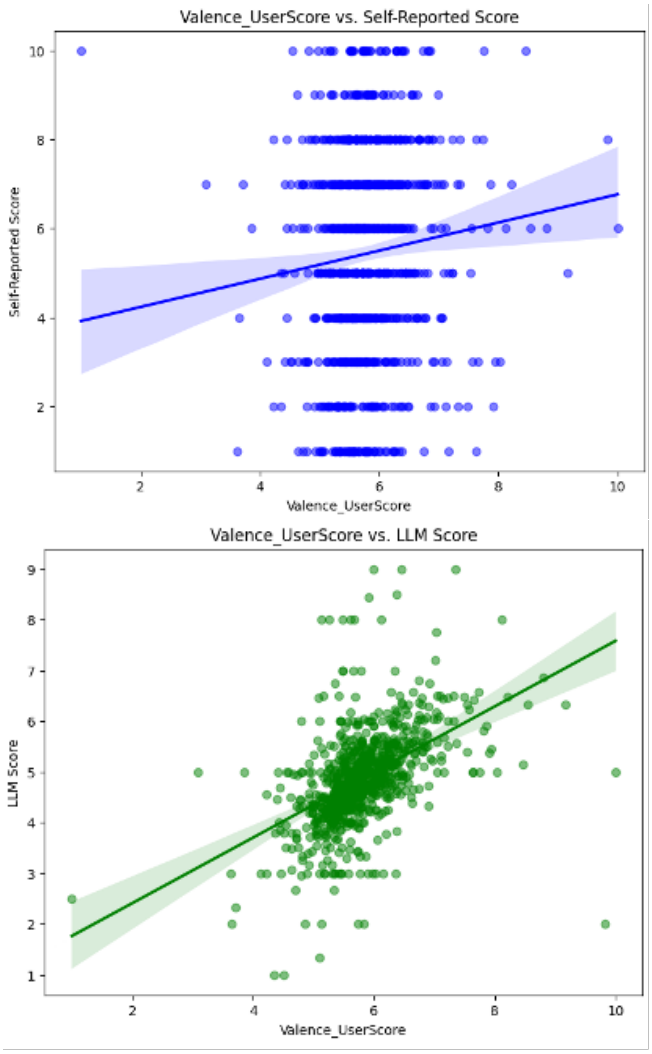


Figure 6: Positive correlation between the valence score with self-reported scores (a.) and LLM scores (b.)

self-reported scores (0.2191, p-value < 0.001) and with LLM scores (0.1714, p-value < 0.001). Like for Q1, we also found a persistent inverse trend between the arousal scores and self-reported (-0.1332, p-value < 0.001), and with LLM scores (-0.06, p-value < 0.001). Lastly, like for Q1, we also found a positive correlation between the dominance and self-reported scores (0.2038, p-value < 0.001), and with LLM scores (0.1582, p-value < 0.001). Our findings indicate valence and dominance as strong predictors of happiness while the arousal showed a persistently weak and negative relationship. These results show that the LLM predictions align moderately well with the self-reported scores, and particularly for valence and dominance scores.

6 DISCUSSION

We aimed to find out the alignment between the user reported happiness ratings with the self-reported happiness scores. Our

Table 2: Correlation between emotion dimensions, LLM scores and user provided scores at granular level

Happiness Type	Emotion Dimension	Correlation	P-value
Short-term	Valence vs Self-Reported Score	0.1575	< 0.001
Short-term	Valence vs LLM Score	0.2096	< 0.001
Short-term	Arousal vs Self-Reported Score	-0.094	< 0.001
Short-term	Arousal vs LLM Score	-0.054	< 0.001
Short-term	Dominance vs Self-Reported Score	0.1996	< 0.001
Short-term	Dominance vs LLM Score	0.1998	< 0.001
Long-term	Valence vs Self-Reported Score	0.2191	< 0.001
Long-term	Valence vs LLM Score	0.1714	< 0.001
Long-term	Arousal vs Self-Reported Score	-0.1332	< 0.001
Long-term	Arousal vs LLM Score	-0.06	< 0.001
Long-term	Dominance vs Self-Reported Score	0.2038	< 0.001
Long-term	Dominance vs LLM Score	0.1582	< 0.001

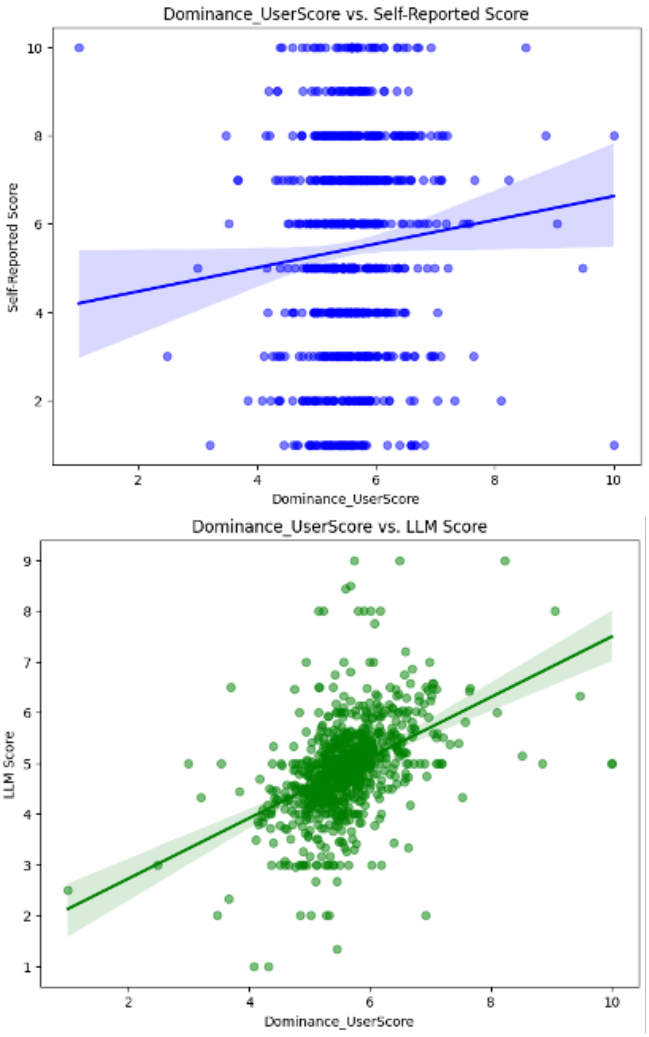


Figure 7: Positive correlation between the dominance score with self-reported scores (a.) and LLM scores (b.)

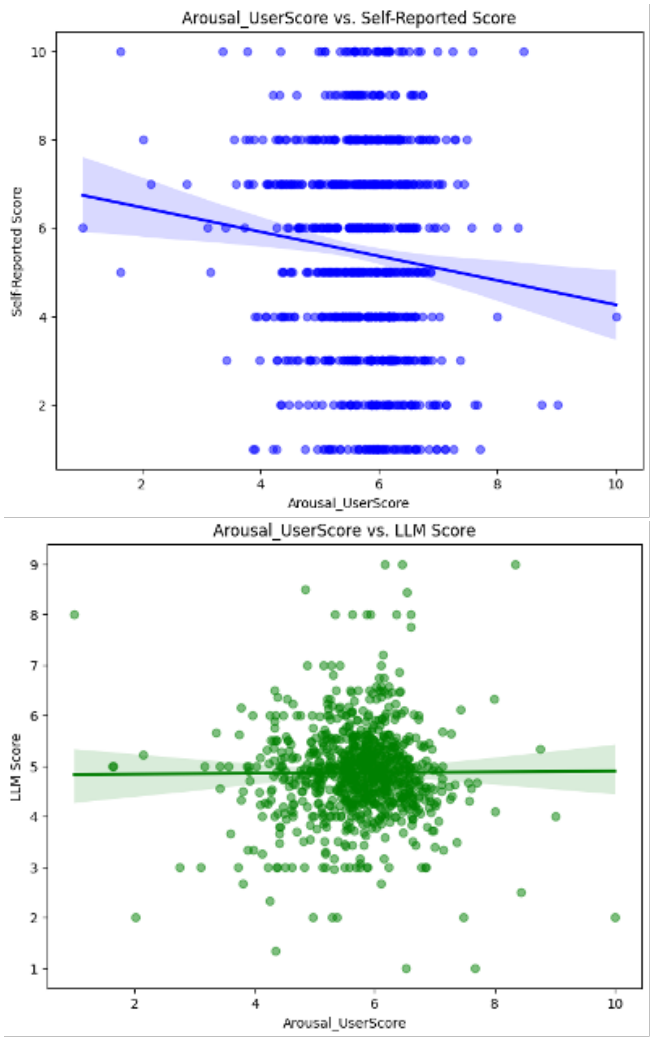


Figure 8: Negative correlation between the arousal score with self-reported scores (a.) and LLM scores (b.)

findings demonstrate that LLM-generated happiness scores are weakly but significantly correlated with the self-reported scores (0.1563, p-value < 0.0001) which indicates low but some reliability in LLM's prediction about an individual's happiness scores. Moreover, the emotion dimensions like valence showed some similarity with self-reported scores (0.0916) and strong similarity with LLM scores (0.4923). This may suggest that the LLMs may majorly use valence ratings to compute their happiness scores and that is why they show strong alignment with the valence scores rather than the self-reported scores. In the similar fashion, the dominance ratings also show moderately positive correlation with the LLM scores (0.4599) while they show weak correlation with the self-reported scores (0.0791). This again indicates that the LLM scores might be using valence and dominance data that they are trained on to compute the happiness scores because of which they are more closely correlated to valence and dominance rating generated by Word2Vec model but shows weak correlation with self-reported values. But lastly, arousal persistently showed weak correlation with the self-reported scores (-0.0963) and very negligible alignment with the LLM scores (0.0067). Even at granular level, the trend was similar for short-term happiness scores and long-term happiness scores. The LLMs were moderately and positively correlated with the valence and dominance scores while their correlation with the self-reported scores was negligible.

Our findings highlight how happiness scores are operationalized in the LLMs using valence and dominance ratings through high correlation between them while actively comparing them with the user rated values derived from cognitive appraisal model-based survey design. These findings establish strong theoretical proof for how language models might be evaluating happiness scores as demonstrated by their strong alignment with the valence and dominance scores but also establishes that the user's reported sentiment or wellness score is not simply a sum of the valence and dominance scores. For industry leaders, our findings highlight misalignment between user-reported scores and LLM regenerated scores which cautions them in their dependence on LLMs for sentiment analysis. As the computer generated and user-reported scores misalign, they raise concern about using LLMs for policy decisions and well-being evaluations. In ethical sense, this highlights LLM's lack of transparency about their limitations in wellness judgement and shows current irreplaceable state of humans in such judgements.

Our study has some limitations like the survey data that was collected in past has limited generalizability and a potential sampling bias. Moreover, even though the computational approaches give insights into sentiment, their dependence on training data restricts accuracy. While the study greatly relies on the collected self-reported survey data, such data may be prone to subjective ratings for which the controlled groups were not created and thus limiting equal comparison of apples with apples. The future work could address the limitations of mentioned here through a dataset where the confounding variables are controlled. Moreover, modeling the impact of text length may further improve the accuracy of results as in our dataset, the text length varied significantly.

7 CONCLUSION

In this study, we investigated how similar the computer-generated wellness scores are with the self-reported user scores of wellnesses and happiness whether measuring short-term or long-term values. Our results showed that LLM generated scores are moderately correlated with the valence and dominance scores but very weakly correlated with the self-reported scores. This might be indicative of LLM's usage of valence and dominance scores from their training data to compute wellness scores resulting in great deviation from user-reported wellness scores. Our potential key-takeaway is that the computational models are heavily reliant on their training data to measure more holistic scores like wellness or happiness scores. As we identified gaps in alignment, our study contributes to highlighting the areas where the current computational models need refinement in order to generate holistic sentiment ratings which aligns with human reported wellness ratings.

8 ACKNOWLEDGMENTS

We acknowledge the Reddit community members who participated in our informal survey, making this research possible. We also thank the reviewers for their valuable feedback.

REFERENCES

[1] Ruut Veenhoven. 2000. The Four Qualities of Life: Ordering concepts and measures of the good life. *Journal of Happiness Studies*, 1(4), 74–100.
[2] Jan Plamper. 2012. Emotion Laboratories and Laboratory Emotions, or, The Birth of Psychological Conceptions of Emotion From the Experimental Spirit. *The History of Emotions: An Introduction*, 178-188.
[3] Robert I. Levy. 1984. Emotion, Knowing, and Culture. *Culture Theory: Essays on Mind, Self, and Emotion*, 214-237.
[4] Ed Diener. 1984. Subjective well-being. *Psychological Bulletin* 95, 3 (May 1984), 542–575.
[5] Organization for Economic Co-operation and Development (OECD). 2013. OECD Guidelines on Measuring Subjective Wellbeing.
[6] James H Fowler and Nicholas A Christakis. 2008. Dynamic spread of happiness in a large social network: longitudinal analysis over 20 years in the Framingham Heart Study. *BMJ* 337, (December 2008), a2338.
[7] Barbara Rose Johnston, Elizabeth Colson, Dean Falk, Graham St John, John H. Bodley, Bonnie J. McCay, Alaka Wali, Carolyn Nordstrom, and Susan Slyomovics. 2012. On Happiness. *American Anthropologist* 114, 1 (2012), 6–18.
[8] Margaret M. Bradley & Peter J. Lang. 1999. Affective Norms for English Words (ANEW): Instruction Manual and Affective Ratings. Technical Report C-1, The Center for Research in Psychophysiology, University of Florida.
[9] Peter Sheridan Dodds and Christopher M. Danforth. 2010. Measuring the happiness of large-scale written expression: Songs, blogs, and presidents. *Journal of Happiness Studies* 11, 4 (2010), 441–456.
[10] Chao Yang and Padmini Srinivasan. 2016. Life Satisfaction and the Pursuit of Happiness on Twitter. *PLoS ONE* 11, 3 (March 2016), 1–30.
[11] Michal Kosinski, David Stillwell, and Thore Graepel. 2013. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences of the United States of America* 110, 15 (April 2013), 5802–5805.
[12] N. Wang, M. Kosinski, J. Rust, and D.J. Stillwell. 2012. Can Well-Being be Measured Using Facebook Status Updates? Validation of Facebook's Gross National Happiness Index. *Social Indicators Research* (January 2012), 1–9.
[13] J. Bao, D. Jurgens, J. Wu, Y. Zhang, and E. Chandrasekharan. 2021. Conversations gone alright: Quantifying and predicting prosocial outcomes in online conversations. In *The Web Conference 2021 - Proceedings of the World Wide Web Conference, WWW 2021*, April 19, 2021. Association for Computing Machinery, Inc, 1134–1145.
[14] Panel on Measuring Subjective Well-Being in a Policy-Relevant Framework; Committee on National Statistics; Division on Behavioral and Social Sciences and Education; National Research Council; Sone AA, Mackie C, editors. 2013. Subjective Well-being: Measuring Happiness, Suffering and Other Dimensions of Experience.
[15] Twitter. 2018. Twitter health metrics proposal submission.
[16] Facebook Research. 2019. Instagram Request for Proposals for Well-being and Safety Research.

[17] Edgar Cabanas and Eva Illouz. 2019. *Manufacturing happy citizens: how the science and industry of happiness control our lives* (English edition. ed.). Polity Press, Cambridge, UK.
[18] Arlie Russel Hochschild. 1985. *The Managed Heart: Commercialization of Human Feeling*.
[19] Michael Barbaro. 2007. At Wal-Mart, Lessons in Self-Help. *New York Times*.
[20] William Davies. 2015. *The Happiness Industry: How the Government and Big Business Sold Us Well-Being*.
[21] Inna Pirina and Çağrı Çöltekin. 2019. Identifying Depression on Reddit: The Effect of Training Data. 2008 (2019), 9–12.
[22] Fionn Delahunty, Ian D. Wood, and Mihael Arcan. 2018. First insights on a passive major depressive disorder prediction system with incorporated conversational chatbot. *CEUR Workshop Proceedings* 2259, (2018), 327–338.
[23] E. Chandrasekharan, U. Pavalanathan, A. Srinivasan, J. Eisenstein, A. Glynn, and E. Gilbert. 2017. You can't stay here: The efficacy of Reddit's 2015 ban examined through hate speech. *Proceedings of the ACM on Human-Computer Interaction* 1, CSCW (November 2017).
[24] Brendan Kennedy, Mohammad Atari, Aida Mostafazadeh Davani, Leigh Yeh, Ali Omrani, Yehsong Kim, Kris Koombs, Shreya Havaldar, Gwenyth Portillo-Wightman, Elaine Gonzalez, Joe Hoover, Aida Azatian, Alyzeh Hussain, Austin Lara, Gabriel Olmos, Adam Omary, Christina Park, Clarisa Wang, Xin Wang, Yong Zhang, and Morteza Dehghani. 2020. The Gab Hate Corpus: A collection of 27k posts annotated for hate speech. (2020), 1–47.
[25] J. D. Featherstone and G. A. Barnett. 2020. Validating Sentiment Analysis on Opinion Mining Using Self-reported Attitude Scores. In *2020 Seventh International Conference on Social Networks Analysis, Management and Security (SNAMS)*, December 14, 2020. 1–4.
[26] Mark Boukes, Bob van de Velde, Theo Araujo, and Rens Vliegenthart. 2020. What's the Tone? Easy Doesn't Do It: Analyzing Performance and Agreement Between Off-the-Shelf Sentiment Analysis Tools. *Communication Methods and Measures* 14, 2 (April 2020), 83–104.
[27] Wouter van Atteveldt, Mariken A. C. G. van der Velden, and Mark Boukes. 2021. The Validity of Sentiment Analysis: Comparing Manual Annotation, Crowd-Coding, Dictionary Approaches, and Machine Learning Algorithms. *Communication Methods and Measures* 15, 2 (April 2021), 121–140.
[28] Bird, S., Klein, E. and Loper, E., 2009. Natural language processing with Python: analyzing text with the natural language toolkit. O'Reilly Media, Inc.
[29] Mikolov, T., 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 3781.
[30] Warriner, A.B., Kuperman, V. and Brysbaert, M., 2013. Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior research methods*, 45, pp.1191-1207.
[31] Hamilton, W.L., Clark, K., Leskovec, J. and Jurafsky, D., 2016, November. Inducing domain-specific sentiment lexicons from unlabeled corpora. In *Proceedings of the conference on empirical methods in natural language processing. conference on empirical methods in natural language processing* (Vol. 2016, p. 595). NIH Public Access.