GPT is an effective tool for multilingual psychological text analysis

Steve Rathje[1]*, Dan-Mircea Mirea[2]*, Ilia Sucholutsky[3], Raja Marjieh[2], Claire E. Robertson[1], Jay J. Van Bavel[1]

*these authors contributed equally

[1]Department of Psychology, New York University
[2]Department of Psychology, Princeton University
[3]Department of Computer Science, Princeton University

Steve Rathje: 0000-0001-6727-571X
Dan-Mircea Mirea: 0000-0002-4349-7059
Ilia Sucholutsky: 0000-0003-4121-7479
Raja Marjieh: 0000-0001-8156-1333
Claire E. Robertson: 0000-0001-8403-6358
Jay J. Van Bavel: 0000-0002-2520-0442

Keywords: GPT, Text Analysis, Large Language Models, Machine Learning, Artificial Intelligence

Abstract

The social and behavioral sciences have been increasingly using automated text analysis to measure psychological constructs in text. We explore whether GPT, the large-language model underlying the artificial intelligence chatbot ChatGPT, can be used as a tool for automated psychological text analysis in various languages. Across 15 datasets (n = 31,789 manually annotated tweets and news headlines), we tested whether GPT-3.5 and GPT-4 can accurately detect psychological constructs (sentiment, discrete emotions, and offensiveness) across 12 languages (English, Arabic, Indonesian, and Turkish, as well as eight African languages including Swahili, Amharic, Yoruba and Kinyarwanda). We found that GPT performs much better than English-language dictionary-based text analysis (r = 0.66-0.75 for correlations between manual annotations and GPT-4, as opposed to r = 0.20-0.30 for correlations between manual annotations and dictionary methods). Further, GPT performs nearly as well as or better than several fine-tuned machine learning models, though GPT had poorer performance in African languages and in comparison to more recent fine-tuned models. Overall, GPT may be superior to many existing methods of automated text analysis, since it achieves relatively high accuracy across many languages, requires no training data, and is easy to use with simple prompts (e.g., “is this text negative?”) and little coding experience. We provide sample code for analyzing text with the GPT application programming interface. GPT and other large-language models may be the future of psychological text analysis, and may help facilitate more cross-linguistic research with understudied languages.

Corresponding Authors: Steve Rathje (srathje@alumni.stanford.edu) and Dan-Mircea Mirea (dmirea@princeton.edu).

GPT is an effective tool for multilingual psychological text analysis

Introduction

Automated text analysis, or the analysis of written language through computational methods, is a rapidly growing tool for social and behavioral scientists (1–3). Because of the increasing availability of text data on the internet (e.g. social media sites and digitized book text), as well as the development of advanced machine learning methods, text analysis has been an increasingly useful tool for testing psychological questions with large datasets. The current paper examines whether text analysis can be made more effective and efficient by taking advantage of recent advances in artificial intelligence.

The growing field of computational social science (4) has used automated text analysis for a variety of different purposes. For example, researchers have applied text analysis tools to examine societal trends (5–8), explore what goes “viral” on social media (9–11), and identify linguistic correlates of mental health conditions (12), ideology (13), and personality (14). Large text datasets are typically analyzed for the presence of various psychological constructs, such as sentiment (i.e. positivity versus negativity) (15–17), discrete emotions such as anger or sadness (18, 19), offensiveness (20), moral emotions (21, 22), out-party animosity (5, 9), or toxicity (23, 24).

Despite the promise and popularity of text analysis, existing text analysis methods have several major shortcomings. One class of methods are dictionary or word-count approaches, which consist of counting the words of a certain category that are present in a text (e.g., counting the number of negative words in a tweet). This method is widely used within psychological research (9, 10, 15, 25). However, dictionary-based methods are often not very accurate as they do not take into account the broader context of a sentence, and are often not validated against manually annotated text (1, 26). As such, there is a great need for more accurate methods.

Machine learning methods have shown promise at accurately detecting psychological constructs in text data. For instance, researchers have used supervised machine learning classifiers to detect positive and negative sentiment (16, 17), moral outrage (22), incivility (23), out-party hate vs. in-party love (27), and discrete emotions (18, 28). Recently, researchers have also started using large-language models (LLMs), or neural networks with many parameters that have been trained on large quantities of unsupervised texts, for psychological text analysis (29). However, most machine learning models are time-consuming and resource-intensive to create. Moreover, they often require high coding proficiency to design or implement and tens of thousands of manually-annotated texts to train (22).

A further shortcoming of both of these approaches is that they are not well-equipped to analyze multilingual data. While several dictionaries have been translated into other languages (26), this translation process is time-intensive, and socio-cultural constructs captured in dictionaries developed for one language may not transfer to another language and culture (30). Similarly, traditional machine learning models tend to only work in the language the model was trained on.

Cross-lingual models, which have been developed in recent years to address this limitation (31), suffer from the same cultural limitations as translated dictionaries. This makes it difficult to study the same constructs in multiple languages, which likely limits the generalizability of text analysis findings. Thus, like many other areas of the social and behavioral sciences which have been criticized for relying too heavily on Western, Educated, Industrialized, Rich and Democratic (or WEIRD) populations (32, 33), text analysis may similarly be focusing on a narrow set of languages and cultures.

We propose that the Generative Pre-trained Transformer (GPT) (34), the LLM developed by OpenAI that underlies the chatbot ChatGPT, has the potential to overcome the downsides present in both dictionary-methods and machine learning methods for automated text analysis. GPT is trained on massive datasets of internet text (such as Common Crawl or Wikipedia), which makes it particularly promising for completing text analysis tasks across multiple languages without any additional training (known as "zero-shot" learning) (35). GPT has been lauded for its ability to exhibit human-level performance on a variety of tasks (e.g., passing the Bar Exam or acing the SAT test), and better performance than existing LLMs (36). Researchers have also recently noted GPT's ability to help with computational social science tasks (37–40), detect misinformation (41), infer politicians' ideologies (42), write persuasive political arguments (43) or even replace human research participants (44, 45). Building on these findings, we examined GPT's potential as a psychological text analysis tool across languages.

Overview

We tested the ability of GPT (3.5 and 4) to accurately detect psychological constructs in text across 15 datasets (*n* = 31,789 annotated tweets and news headlines, ***Table 1***). Each of these datasets were manually annotated by human raters for the presence or absence of specific psychological constructs – sentiment, discrete emotions, and offensiveness. For each psychological construct, we first examined GPT's performance in English as well as a second, unrelated language (Arabic, Indonesian, or Turkish) using 6 publicly available datasets with categorical labels (*datasets 1-6*). Then, we analyzed a dataset of news headlines rated for sentiment and discrete emotions on a Likert scale to examine how GPT performs with psychological scale ratings (46), a different type of text, and a dataset that is not publicly available on the internet (*dataset 7*). Finally, to examine whether GPT performs equally well with less commonly-spoken or studied languages, we tested GPT's ability to detect sentiment in eight African languages, such as Swahili, Amharic, Yoruba and Kinyarwanda (*datasets 8-15*). For each dataset, we compare the performance of GPT 3.5 and 4 to other common methods of text analysis, such as dictionary methods and fine-tuned machine learning models.

Results

For each of the 15 datasets (see ***Table 1*** for descriptions), we used the GPT application programming interface (API) to repeatedly prompt GPT using R or Python code. We used simple prompts, such as "Is the sentiment of this text positive, neutral, or negative? Answer only with a number: 1 if positive, 2 if neutral, and 3 if negative. Here is the text: [tweet or news

headline text]" (see ***Table 2*** for prompt summary). Then, we examined how GPT's performance aligned with human annotations. We used two metrics that are traditionally used to measure the performance of machine learning models: accuracy and average *F1*. Accuracy is the number of correct ratings (i.e. the number of GPT outputs that matched the manual annotations) over the total number of ratings. Average *F1* is a more complex metric that takes into account the various types of errors made by GPT (false positives and false negatives) and is used frequently in the machine learning literature. See ***Methods*** for a detailed description of these performance metrics, and see our OSF for code and datasets: https://osf.io/6pnb2/. Finally, we also examined whether we could improve GPT's accuracy by providing it with a few examples (known as "few-shot learning") and comparing the results to those without any examples ("zero-shot learning"). Importantly, several of our datasets (datasets *6-15*, see ***Table 1***) were not publically available on the internet at the time GPT-3.5 and 4 were trained, meaning they could not have been a part of GPT's training dataset[1].

[1] GPT was only trained on data that was made publically available on the internet in the year 2021 and before, according to https://help.openai.com/en/articles/6783457-what-is-chatgpt.

Table 1. Description of datasets used

Dataset	Construct	Text type	Size of dataset	Labels	Language	Number of Speakers (millions)
Sentiment of English tweets (2017)	Sentiment	Tweets	12,283	Positive, Negative, Neutral	English	1,450
Sentiment of Arabic tweets (2017)	Sentiment	Tweets	6,100	Positive, Negative, Neutral	Arabic	630
Discrete emotions in English tweets (2020)	Discrete Emotions	Tweets	1,421	Anger, Joy, Sadness, Optimism	English	1,450
Discrete emotions in Indonesian tweets (2020)	Discrete Emotions	Tweets	440	Anger, Fear, Sadness, Love, Joy	Indonesian	300
Offensiveness in English tweets (2019)	Offensiveness	Tweets	860	Offensive, Not Offensive	English	1,450
Offensiveness in Turkish tweets (2020)	Offensiveness	Tweets	3,528	Offensive, Not Offensive	Turkish	88
Sentiment & discrete emotions in news headlines (2023)	Sentiment, Discrete emotions	News headlines	213	1 = very negative; 7 = very positive	English	1,450
Sentiment of African tweets (2023)	Sentiment	Tweets	748	Positive, Negative, Neutral	Swahili	220
	Sentiment	Tweets	1000	Positive, Negative, Neutral	Hausa	72
	Sentiment	Tweets	1000	Positive, Negative, Neutral	Amharic	57.5
	Sentiment	Tweets	1000	Positive, Negative, Neutral	Yoruba	55
	Sentiment	Tweets	1000	Positive, Negative, Neutral	Igbo	42
	Sentiment	Tweets	949	Positive, Negative, Neutral	Twi	17.5
	Sentiment	Tweets	1026	Positive, Negative, Neutral	Kinyarwanda	15
	Sentiment	Tweets	234	Positive, Negative, Neutral	Tsonga	7

We used 15 different datasets which contained 31,789 manually annotated tweets and news headlines in 12 languages from various language families, annotated for 3 different psychological constructs (sentiment, discrete emotions, and offensiveness). Datasets 7-15 were not publically available on the internet at the time GPT was trained in 2021, and thus could not have influenced the training dataset.

Table 2. Prompt table

Sentiment analysis (categorical)	Emotion detection (categorical)	Offensiveness	Sentiment analysis (Likert)	Emotion detection (Likert)
Is the sentiment of this (Arabic/Swahili/…) text positive, neutral, or negative? Answer only with a number: 1 if positive, 2 if neutral, and 3 if negative. Here is the text: *[Tweet text]*	Which of these four emotions - [list of emotions] - best represents the mental state of the person writing the following (Indonesian) text? Answer only with a number: 1 if [emotion1], 2 if [emotion2], [...]. Here is the text: *[Tweet text]*	Is the following (Turkish) post offensive? Answer only with a number: 1 if offensive, and 0 if not offensive. Here is the post: *[Tweet text]*	How negative or positive is this headline on a 1-7 scale? Answer only with a number, with 1 being 'very negative' and 7 being 'very positive.' Here is the headline: *[Headline text]*	How much [emotion] is present in this headline on a 1-7 scale? Answer only with a number, with 1 being 'no [emotion]' and 7 being 'a great deal of [emotion].' Here is the headline: *[Headline text]*

Shown are all the prompts used for each construct. Non-English prompts were derived from the English prompts by specifying the language the text was written in. Prompts in combination with the tweet or headline text were run for each text entry in the dataset using the GPT API.

1. Sentiment

We first examined GPT's ability to detect sentiment – or the overall positivity, negativity, or emotional neutrality expressed in text. To assess GPT's performance across languages, we used manually annotated datasets of tweets in both English and Arabic (47). Both datasets came from the 2017 iteration of SemEval, a competition for designing machine learning methods for text analysis (see ***Methods***). We found that GPT achieves good performance at predicting human ratings in both English (Accuracy = 0.673, *F1* = 0.683) and Arabic (Accuracy = 0.700, *F1* = 0.720), (***Table 3***). Moreover, GPT outperforms the best model from the SemEval competition in both languages (***Table 4***). We note that the original study is from 2017, meaning these models may be outdated, though GPT's performance is also slightly better than a more recent classifier of Arabic sentiment (*F1* = 0.704) (48). Overall, GPT appears to be effective at multilingual sentiment analysis, with performance comparable to top-performing machine learning models from previous studies.

Interestingly, GPT-3.5 performs slightly better than 4 on both tasks (English: GPT-3.5 *F1* = 0.683 vs. GPT-4 *F1*= 0.633; Arabic: GPT-3.5 *F1* = 0.720 vs. GPT-4 *F1* = 0.707). Examination of the confusion matrices (***Figure 1***) reveals a possible driver of this effect: GPT-4 is more likely to classify "neutral" tweets as either "positive" (31% of neutral tweets in both English and Arabic) or "negative" (36% of neutral tweets in English and 34% in Arabic) compared to GPT-3.5 (English neutral tweets: 16% labeled as positive and 25% labeled as negative; Arabic neutral tweets: 18% labeled as positive and 15% labeled as negative). This suggests GPT-4 might have a cross-linguistic bias towards overestimating sentiment in a given text compared to humans.

Table 3. GPT-3.5 vs. GPT-4 Results - Twitter data

		GPT-3.5		GPT-4	
Language	**Construct**	**Accuracy**	**F1**	**Accuracy**	**F1**
English	Sentiment	0.673	0.683	0.566	0.633
Arabic	Sentiment	0.700	0.720	0.655	0.707
English	Discrete emotions	0.738	0.714	0.816	0.779
Indonesian	Discrete emotions	0.686	0.686	0.741	0.740
English	Offensiveness	0.769	0.721	0.801	0.746
Turkish	Offensiveness	0.836	0.752	0.857	0.709
Swahili	Sentiment	0.596	0.560	0.492	0.488
Hausa	Sentiment	0.591	0.586	0.448	0.399
Amharic	Sentiment	0.206	0.238	0.737	0.609
Yoruba	Sentiment	0.542	0.506	0.607	0.579
Igbo	Sentiment	0.624	0.580	0.643	0.622
Twi	Sentiment	0.406	0.408	0.538	0.505
Kinyarwanda	Sentiment	0.574	0.574	0.622	0.624
Tsonga	Sentiment	0.291	0.258	0.311	0.302
Average	**-**	**0.588**	**0.571**	**0.631**	**0.603**

We report the ability of GPT-3.5 and GPT-4 to accurately detect three psychological constructs (sentiment, discrete emotions, and offensiveness) across 12 languages. We report two performance metrics commonly used in machine learning: accuracy (number of correct ratings over total number of ratings), and *F1*, a more complex measurement that takes into account different types of classification errors (see ***Methods*** for a detailed description of performance metrics). Green indicates instances where GPT-4 was better than GPT-3.5, whereas red indicates instances where GPT-4 was worse than GPT-3.5. As shown, GPT-4 is better than GPT-3.5 with lesser-spoken African languages and discrete emotions, whereas GPT-3.5 is better at sentiment in widely spoken languages. Because of cost limitations, we only analyzed the first 1000 texts for GPT-4 in three datasets: English sentiment, Arabic sentiment, and Turkish offensiveness.

Table 4. GPT vs. Top-Performing Machine Learning Models

Language	Construct	GPT-3.5 F1	GPT-4 F1	Top-performing model F1	Model type	Year of model
English	Sentiment	0.685	0.633	0.677	LSTM-CNN	2017
Arabic	Sentiment	0.720	0.707	0.610	Naive Bayes	2017
English	Discrete emotions	0.714	0.779	0.785	BERT	2020
Indonesian	Discrete emotions	0.686	0.740	0.795		2020
English	Offensiveness	0.721	0.746	0.829		2019
Turkish	Offensiveness	0.752	0.709	0.826	XLM-BERT	2020
Swahili	Sentiment	0.560	0.488	0.657	Fine-tuned XLM-R	2023
Hausa	Sentiment	0.586	0.399	0.826		
Amharic	Sentiment	0.238	0.609	0.640		
Yoruba	Sentiment	0.506	0.579	0.800		
Igbo	Sentiment	0.580	0.622	0.830		
Twi	Sentiment	0.408	0.505	0.675		
Kinyarwanda	Sentiment	0.574	0.624	0.726		
Tsonga	Sentiment	0.258	0.302	0.607		

We compare the performance of GPT-3.5 and GPT-4 to the performance of the top machine learning models reported in the papers from which we retrieved the tested datasets. GPT-3.5 outperforms top performing models for detecting English and Arabic sentiment. GPT-4 comes close to but does not outperform the discrete emotion models. GPT-3.5 and GPT-4 are not as good as more recent fine-tuned models for detecting sentiment in African languages. It should be noted though that these models have been extensively fine-tuned with labeled training data, and are based on LLMs that are similar to GPT. The abbreviations are as follows: LSTM, Long Short Term Memory; CNN, Convolutional Neural Network; BERT, Bidirectional Encoder Representations from Transformers; XLM, Cross-Lingual Model; XLM-R, XLM combined with RoBERTa (a variant of BERT with more extensive pretraining).

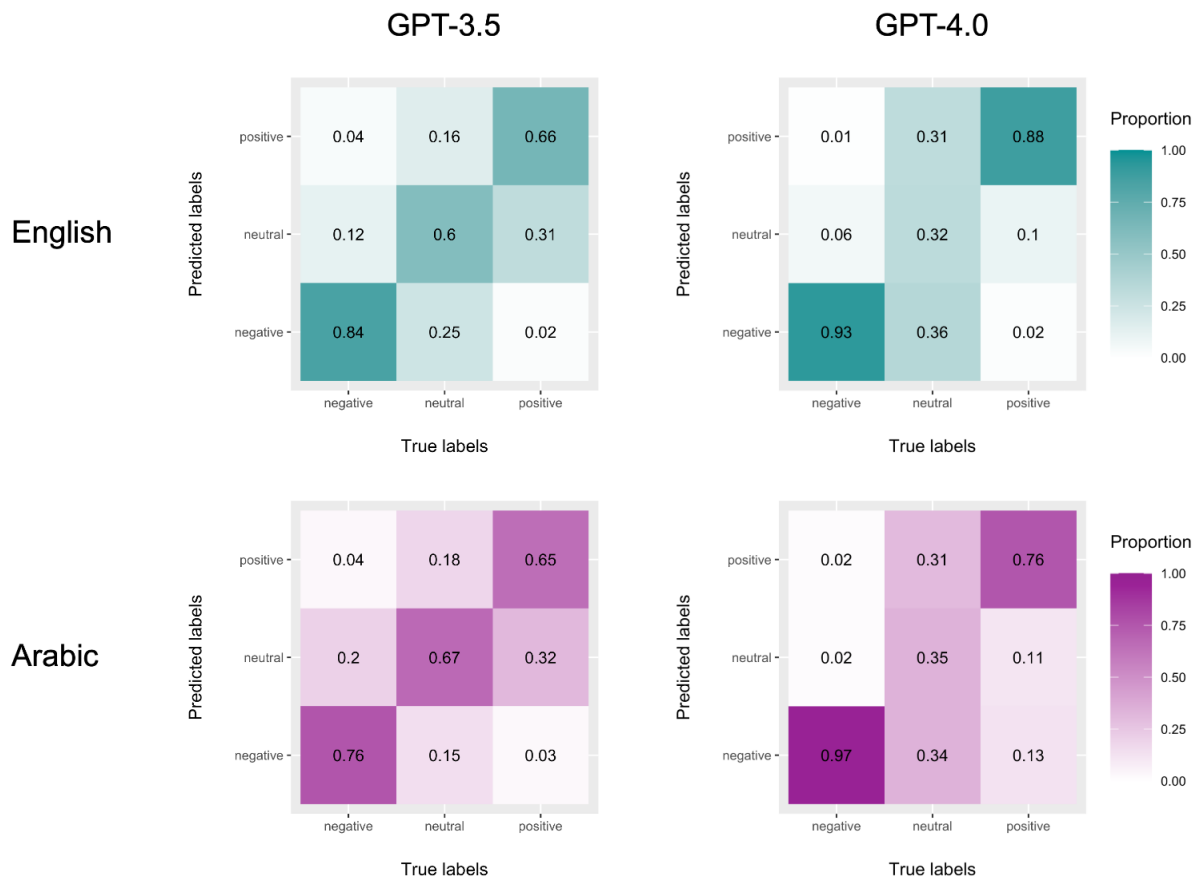

Figure 1. Confusion matrices of GPT-labeled ('predicted') sentiment vs. human-labeled ('true') sentiment. Numbers show the proportion of tweets labeled by humans with a particular label that received a particular GPT label (numbers sum up to 1 in each column). Darker colors reflect higher proportions. Colors reflect language: turquoise for English and pink for Arabic. Heatmaps on the left are from GPT-3.5 predictions, and heatmaps on the right are from GPT-4 predictions. As shown, GPT-4 appears to have a bias toward labeling neutral tweets as positive or negative. For GPT-4, because of cost limitations, only the first 1000 tweets were analyzed.

2. Discrete emotions

Next, we examined GPT's ability to accurately detect more complex discrete emotions, such as anger, joy, fear, and sadness. To assess the model's multilingual performance, and to show results generalize beyond English and Arabic, we compared English with another unrelated, lesser-studied language – Indonesian, once again using two existing datasets. We found that both versions of GPT showed high agreement with humans in both English (GPT-3.5 *F1* = 0.720, GPT-4 *F1* = 0.779) and Indonesian (GPT-3.5 *F1* = 0.678, GPT-4 *F1* = 0.740) (***Table 3***). GPT-4 showed an improvement over 3.5 in both languages, reaching an *F1* score that was roughly equivalent to the top-performing state-of-the-art LLM in English (***Table 4***). For Indonesian, GPT-4 with few-shot learning (i.e., adding one example for each class within the prompt) outperformed 4 with zero-shot learning, and reached the performance of the top existing model, a fine-tuned BERT model (***Supplementary Table S1***). Full confusion matrices can be found in ***Supplementary Figure S1*****.**

3. Offensiveness

We then examined GPT's ability to detect a different psychological construct, offensiveness, in both English and Turkish (20, 49). Offensive text was defined as text that "includes insults, threats, and posts containing any form of untargeted profanity" (20). Although we again found decent agreement between GPT and human ratings (*F1* > 0.7 in both languages), the performance did not reach that of the top performing models from their respective studies (***Table 4***). Moreover, GPT-3.5 outperformed GPT-4 for the Turkish dataset for the *F1* statistic. The lower performance of GPT-4 may have resulted from a lower likelihood of labeling tweets as offensive (***Supplementary Figure S2***), once again suggesting that the different versions of GPT can have slight biases when solving text analysis tasks.

4. Sentiment and discrete emotions measured on a continuous scale

We have so far seen that GPT is capable of accurately detecting psychological constructs in written language, with performance comparable to several top-performing machine learning models. However, it is unclear whether this performance generalizes to other types of text data besides tweets. Moreover, it is unclear whether GPT performs similarly with other types of ratings, such as Likert scales (e.g., 1 = strongly disagree, 7 = strongly agree), which are commonly used in psychology and the social sciences. Finally, since all of the datasets used so far are publicly available on the internet, it is possible that they were part of GPT's training set. This calls for testing the model on a new dataset that it could have not been possibly exposed to.

To address these considerations, we analyzed a recent dataset of news headlines annotated for sentiment and four discrete emotions using 1-7 Likert scales, which was not publicly available on the internet at the time GPT was trained (15). The prompts for Likert scales were slightly different (e.g., "How negative or positive is this headline on a 1-7 scale?"; see ***Table 2*** for prompts). We found very high correlations (between $r = 0.56$ and $r = 0.74$) between GPT-3.5 and human ratings, and even higher correlations for GPT-4 (between $r = 0.66$ and $r = 0.75$). (see ***Figure 2*** and ***Table 5***). This suggests that GPT is capable of accurately detecting psychological constructs in text regardless of the format of the ratings or the type of text.

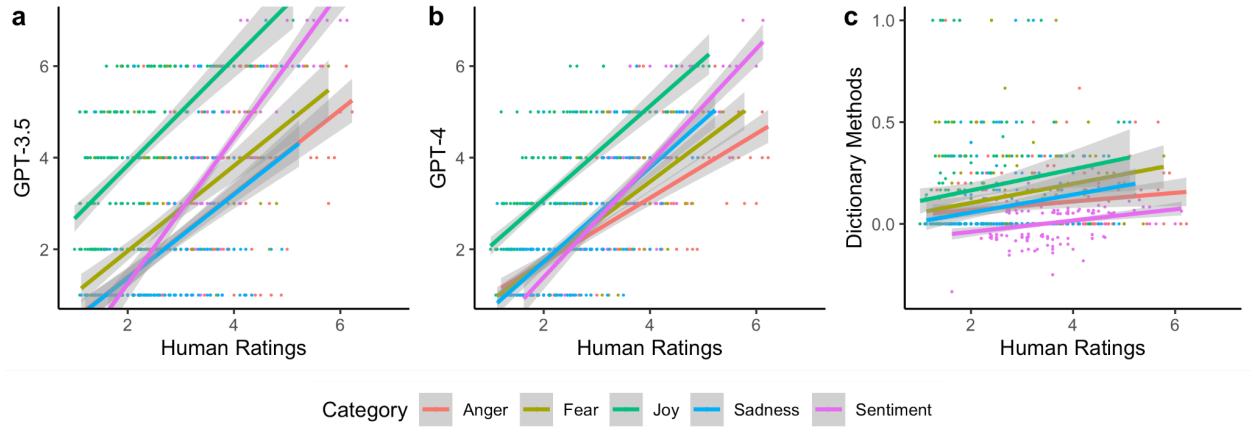

Figure 2. Scatterplots showing correlations between human ratings and ratings predicted by different text analysis methods. a) GPT-3.5 ratings; b) GPT-4 ratings; c) ratings computed using dictionary methods (LIWC and NRC dictionaries with negation handling). Data is from 213 manually annotated headlines (measured on a Likert scale from 1-7) and the corresponding dictionary method calculations are taken from (15).

Table 5. GPT vs. Dictionary Methods (LIWC and NRC Discrete Emotions)

Psychological Construct	**Spearman Correlation with Manual Annotators's Ratings**			**Spearman Correlation between GPT-3.5 and GPT-4 output**
Method	**GPT-3.5**	**GPT-4**	**Dictionary Methods**	
Sentiment	0.74	0.75	0.30	0.79
Anger	0.64	0.69	0.22	0.72
Fear	0.56	0.66	0.29	0.72
Joy	0.64	0.70	0.30	0.79
Sadness	0.56	0.67	0.30	0.67

We show the Spearman correlation between the ratings from GPT-3.5 and GPT-4 and the ratings of manual annotators for sentiment and discrete emotions. We compare this to the correlation between dictionary methods (LIWC and NRC dictionaries with negation handling) and the ratings of manual annotators. Data is from 213 manually annotated headlines (measured on a Likert scale from 1-7) taken from (15).

We also compared GPT's performance to the performance of two popular dictionary methods used in the study the dataset was taken from: the Linguistic Inquiry and Word Count (LIWC) method of measuring sentiment (26), and the NRC Emotion Lexicon (50) method of measuring discrete emotions. Dictionary scores also included negation handling, which takes into account words like "not. " For details of how these scores were calculated, see (15). The correlations between these dictionary-based methods and manual annotations were much smaller (between $r = 0.22$ and $r = 0.30$) than the correlations between manual annotations and both versions of GPT. Z-tests found that all of the correlations between manual annotations and GPT output were significantly different from the correlations between manual annotations and dictionary methods (all $ps < 0.001$). Thus, GPT appears to be far more effective at measuring sentiment and discrete emotions than common dictionary-based methods that are very popular in psychology and the social sciences.

One potential limitation of using GPT is that it continues to evolve over time and may provide very different estimates from one version to the next. To rule out this problem, we tested the Spearman correlation between GPT-3.5 and GPT-4 for this dataset (see ***Table 5***). We found very high correlations between versions of GPT (between $r = 0.67$ and $r = 0.79$), indicating that different versions of GPT provide very similar (albeit not exact) output for text analysis problems.

5. Sentiment in lesser-resourced African languages

Our analyses so far have focused on widely-spoken languages that are relatively highly represented in CommonCrawl – a large repository of internet text data which is the main source of training data for GPT (51). To test GPT's ability on lesser-resourced languages, we took advantage of a recent collection of tweets manually annotated for sentiment in multiple African languages (52). We chose eight of these languages, some of which had less than 20 million speakers (***Table 1***) and relatively little representation in CommonCrawl.

We found that GPT-3.5 and GPT-4.0 generally performed worse in these languages than in English and Arabic (***Table 3***). GPT's performance was generally better than chance (which is 0.33 accuracy and *F1* for 3-label classification), with the exception of Tsonga (GPT-3.5: $F1 = 0.258$; GPT-4: $F1 = 0.302$) and Amharic (GPT-3.5: $F1 = 0.238$). Tsonga was the least-spoken language in our datasets (7 million speakers) and so the lower performance was expected, whereas Amharic was surprising given its higher number of speakers (57.5 million speakers). Upon further examination, we found that GPT-3.5 severely mistranslated Amharic sentences, which may explain the low performance.

This issue seems to have been solved in GPT-4, which achieved relatively high performance in Amharic ($F1 = 0.609$). More broadly, GPT-4 outperformed GPT-3.5 in all languages but the two most spoken ones, Swahili and Hausa (***Table 3***). Similarly to English and Arabic, GPT-4 appeared to be more likely than GPT-3.5 to label "neutral" Swahili tweets as either "positive" or "negative," which may have led to the lower performance (***Supplementary Figure 3***). The pattern in Hausa was less consistent. Despite the general improvement in performance from GPT-3.5 to GPT-4.0, these models did not come close to the performance of the state-of-the-art model, Afro-XLMR, an LLM trained on all these African languages and with further fine-tuning

(***Table 4***). Given that GPT had no fine-tuning on manually annotated datasets, this lower performance is not entirely surprising.

While classification accuracy and F1 were lower for African languages, they still tended to be within the range of many existing sentiment analysis tools. For instance, one evaluation of several popular sentiment analysis tools (such as SentiStrenth) found an accuracy between 0.53 and 0.72 (53). At least one version of GPT was able to reach this accuracy range in all languages except Tsonga, suggesting that GPT is reasonably accurate with under-resourced African languages – albeit less accurate than fine-tuned, top-performing models.

Discussion

We tested whether recent advances in artificial intelligence – specifically, the popular large language model GPT – could help make automated text analysis more effective and efficient. Across 12 different datasets, we found that the two most recent versions of GPT (3.5 and 4) could accurately detect various psychological constructs (sentiment, discrete emotions, and offensiveness) in different types of text (tweets and news headlines) and across 12 languages, including under-resourced African languages (54). GPT performs much better than English-language dictionary methods at both sentiment analysis and discrete emotion detection. In many cases, GPT performed close to (and sometimes better than) fine-tuned machine learning models. However, the performance of GPT was lower than the performance of more recent fine-tuned models based on LLMs. GPT also performed worse in African languages, (particularly lesser-spoken ones such as Tsonga or Twi), although this performance was still within the range of many commonly-used English sentiment analysis tools for all but one language (53). Overall, GPT shows promise as a multilingual text analysis tool.

We make the case that GPT is superior to many existing automated text analysis methods. While dictionary-based text analysis methods are often used because of their user-friendliness, GPT is also very easy to use and achieves much higher accuracy. Thus, there may be little need for popular dictionary methods such as LIWC. In some cases, GPT may also be a better choice than fine-tuned machine learning models. While machine learning classifiers require large amounts of manually annotated text to train and high coding proficiency, GPT does not require training data, and is intuitive to use with little coding experience, since it works via prompting with minimal programming. We provide sample code for analyzing text data with GPT on our OSF: https://osf.io/6pnb2/. However, researchers may still wish to fine-tune LLM-based machine learning models if they want to surpass the accuracy of GPT.

Given its high-performance across languages, GPT could also facilitate more complex cross-linguistic research that takes into account under-resourced languages. This might help solve the issue of text analysis focusing too much on WEIRD populations and English-language datasets. While GPT's performance was worse for lesser-resourced languages (such as Tsonga, which has 7 million global speakers), GPT-4 showed an improvement over GPT-3.5 for those languages. This provides hope that GPT and other LLMs will continue to get better at text

analysis tasks for lesser-resourced languages. Future research should continue to explore the accuracy of GPT and other LLMs across different languages and cultures.

Interestingly, GPT-4 seemed to show a bias towards detecting positive or negative sentiment in neutral texts, which made it perform slightly worse than GPT-3.5 at sentiment analysis in widely spoken languages. This striking difference between two consecutive versions of the same model might reflect a general shift on the internet towards more emotionally-charged language, which is in line with recent findings (6, 55). We also experimented with providing GPT-4 with examples ("few-shot" learning) in an attempt to improve its performance, finding this sometimes did increase but other times decreased performance (See ***Supplementary Appendix Table S1***). We encourage researchers to experiment with different GPT versions, prompts, and few-shot learning strategies for whatever construct they are measuring, and we offer example code to serve as a starting point. Ideally, researchers should validate GPT against manually annotated text if they are using it for a new construct and/or language to control for any potential biases that the model may inherit from its training datasets. For instance, GPT might have language-specific biases, such as a lower threshold toward labeling some tweets as negative in certain languages. If not taken into account, these biases could skew the conclusions of cross-cultural studies.

Using GPT for text analysis does have its limitations. First, GPT does not always surpass top-performing machine learning models, meaning that there will likely still be additional utility in fine-tuned models. Yet, GPT's performance is high enough that it can be used in many cases where machine learning classifiers are not available. Furthermore, the GPT API is behind a paywall, with more recent versions costing more (for instance, GPT-3.5 cost around $0.50 to analyze 1000 tweets, whereas GPT-4 cost over $5.00). These sums could quickly become unaffordable for researchers with fewer resources when analyzing datasets with hundreds of thousands or millions of pieces of text. GPT is also not open-source, meaning that some researchers may want to use other open-source large-language models to understand more of what is going on "under the hood." Finally, while we measured several psychological constructs across 12 languages, there are no manually annotated datasets available in many understudied languages, which prevents us from analyzing GPT's accuracy in less commonly spoken languages. Future research should continue to explore the accuracy of GPT across more languages and constructs.

Conclusions

Our results suggest that GPT is an effective tool for detecting various psychological constructs in text across several languages. GPT may have a number of benefits over existing text analysis methods, such as dictionary-based methods and fine-tuned machine learning models. It shows reasonable accuracy across languages, requires no training data, and is easy to use with little code and simple prompts. Therefore, we believe GPT and future LLMs may soon supplant existing automated text analysis approaches, and facilitate more cross-linguistic research with under-resourced languages.

Methods

Datasets

Sentiment of English tweets. We used the dataset of English tweets from SemEval-2017 Task 4: Sentiment Analysis on Twitter (47). Each tweet in this dataset was annotated by at least 5 human annotators from the crowdsourcing service CrowdFlower. We applied GPT-3.5 and GPT-4 to subtask A, which involved classifying the sentiment of each tweet into one of 3 classes: positive, negative or neutral. We used the designated test set for subtask A (*N* = 12,284). Because of cost limitations, for GPT-4, we only analyzed the first 1000 tweets.

Sentiment of Arabic tweets. We also used the Arabic dataset from SemEval-2017 Task 4, which was similarly annotated using CrowdFlower. For consistency with the English sentiment analysis task, we chose subtask A for the Arabic data as well, and tested the performance of GPT on the Arabic test set of subtask A (*N* = 6,100). Because of cost limitations, for GPT-4, we only analyzed the first 1000 tweets.

Discrete emotions in English tweets. To examine the performance of GPT at detecting discrete emotions in tweets, we applied it to a dataset from the TweetEval benchmark (56). This dataset was adapted from a previous one used in SemEval-2018 Task 1 (57), which was focused on emotion detection. The previous dataset contained tweets labeled with one or more of 12 emotion labels, following annotations by at least 7 CrowdFlower workers for each tweet. The TweetEval dataset was created from this dataset by removing tweets with multiple labels and only keeping the 4 most common labels: anger, joy, sadness and optimism. GPT-3.5 and GPT-4 were applied to the test portion of this dataset (*N* = 1,421).

Discrete emotions in Indonesian tweets. We used a dataset from the IndoNLU benchmark (58) to assess GPT's performance on detecting discrete emotions in a different language from English. This was a dataset of tweets labeled with one of five emotions - anger, joy, sadness, fear and love - by two annotators, taken from a previous study (59). We used the test portion of this dataset (*N* = *442*).

Offensiveness in English tweets. We used a dataset of English tweets from SemEval-2019 Task A: Offensive Language Identification (20). Each tweet was annotated by two people via the crowdsourcing platform Figure Eight. In the case of disagreement, a third annotator was used, and the annotation was decided by majority vote. Tweets were classified as either offensive or non-offensive. We used the test dataset (*N* = 860).

Offensiveness in Turkish tweets. We also used a dataset of Turkish tweets from SemEval-2020 Task 12: Multilingual Offensive Language Identification in Social Media (49). Most tweets were annotated by a single annotator. Tweets were classified as either offensive or non-offensive. We used the test dataset (*N* = 3,528). Because of cost limitations, for GPT-4, we only analyzed the first 1000 tweets in the dataset.

Sentiment and discrete emotions in news headlines. We used a dataset of 213 news headlines manually annotated for sentiment and discrete emotions (e.g., fear, joy, sadness, anger) (15). Manual annotations were made on a 1-7 scale by 8 annotators, and averaged for each construct. This dataset was created to evaluate two common approaches for measuring sentiment and emotions in text: the NRC emotion lexicon (50) and the Linguistic Inquiry and Word Count (LIWC) (26).

Sentiment analysis in African languages. Finally, we analyzed a recent collection of datasets of tweets in various African languages. The tweets were manually coded for sentiment and used to develop multilingual machine learning models within one of the tasks at SemEval-2023 - AfriSenti (52). Out of the 14 languages included, we excluded two Arabic dialects due to the overlap with our previous analysis of Arabic sentiment. We also excluded Mozambican Portuguese because it is a variety of Portuguese, meaning that GPT might perform better simply due to generalization from other varieties of Portuguese. Additionally, we excluded Nigerian Pidgin due to its lexical overlap with English, leading to the same potential generalization issue. Lastly, we excluded Tigrinya and Oromo, since the AfriSenti models were never trained on these languages (whereas GPT might have seen these languages in its training). For the remaining 8 languages, we applied GPT-3.5 and GPT-4 to their respective test sets. Due to cost constraints, we selected a random sample of 1000 tweets for the datasets which had significantly more than 1000 tweets.

GPT procedure

We used the OpenAI API to query GPT. The code for querying was written in R for GPT-3.5 and in Python for GPT-4. Analysis code was written in R. See https://osf.io/6pnb2/ for example code and data. We used a temperature of 0 to obtain the highest probability predictions of the models. This setting means that the GPT output would not differ if we ran our analysis a second time. For each task, we used tailored prompts that included the relevant question followed by an instruction to provide the answer as a number and an explanation of what the numbers meant (***Table 2***). The non-English versions were identical to the English versions, with the addition of the name of the respective language before the word ‘text’ or ‘post’. The prompts were identical for the 3.5 and 4 runs.

Few-shot learning

We ran GPT-4 with few-shot learning on each of the first 6 datasets to test its ability to improve performance over the default, zero-shot approach. To achieve few-shot learning, we added one example of text and its corresponding label taken from the same dataset to the prompt, which we then excluded from the analysis. An example prompt used for few-shot learning in the English discrete emotion detection task is shown as an example in ***Supplementary Table 2*****.**

Performance evaluation metrics

Accuracy. The classification accuracy was computed in each Twitter task by calculating the number of tweets which were identically coded by humans and GPT and dividing that number by the total number of tweets in the dataset. This simple metric has the issue that it is biased

towards classes or labels with more data points (e.g. if a dataset has 90 positive tweets and 10 negative tweets, a classifier which labels all tweets as positive would have an accuracy of 90%).

Macro-averaged F1. We used the macro-averaged F1 score to quantify classification accuracy in a way that is less sensitive to imbalances in the datasets. The F1 score of a classification model for a specific class (e.g. for detecting negative tweets vs. all other tweets) represents the harmonic mean of the model's precision and recall.

$$F_1 = 2\frac{precision \cdot recall}{precision + recall}$$

The precision represents the proportion of data points labeled with the given class by the classifier that are truly of that class ('true positives') as opposed to falsely labeled ('false positives'). In the negative tweets example, precision would be the proportion of tweets labeled as negative by the classifier that are actually negative.

$$precision = \frac{true\ positives}{true\ positives + false\ positives}$$

The recall represents the ratio of true positives over the sum of true positives and false negatives (members of the class which are wrongly labeled by the classifier as not belonging to the class). In our example, recall is the proportion of tweets that are actually negative that are labeled by the classifier as negative.

$$recall = \frac{true\ positives}{true\ positives + false\ negatives}$$

In each Twitter task, the F1 score for each class was calculated (e.g. for negative tweets vs. all others, for positive tweets vs. all others, etc.) and the arithmetic mean of all F1 scores was computed to give the macro-averaged F1.

Spearman correlation. The results in the news headline task, which was coded on a 1-7 Likert scale, were evaluated by Spearman correlation between the GPT and human values for the different constructs (sentiment and the four basic emotions).

Acknowledgements: We are grateful for funding from an NSERC fellowship (567554-2022) to Ilia Sucholutsky, a Gates Cambridge Scholarship awarded to Steve Rathje (OPP1144), a Russell Sage Foundation grant awarded to Jay Van Bavel and Steve Rathje (G-2110-33990), and NIMH grant R01MH119511 awarded to Yael Niv and supporting Dan-Mircea Mirea. We thank Tom Griffiths for helpful discussions and Eric Shuman for help writing R code.

Author Contributions: S.R., D.M.M., I.S., R.M., and J.V.B. designed research, D.M.M., I.S. and S.R. performed research, D.M.M. and S.R. analyzed data, S.R. and D.M.M. drafted the paper and all authors provided critical edits.

Competing Interest Statement: The authors declare no competing interests.

References

1. J. Wilkerson, A. Casas, Large-scale computerized text analysis in political science: Opportunities and challenges. *Annual Review of Political Science* **20**, 529–544 (2017).
2. T. Al Baghal, L. Sloan, C. Jessop, M. L. Williams, P. Burnap, Linking Twitter and survey data: The impact of survey mode and demographics on consent rates across three UK studies. *Social Science Computer Review*, 0894439319828011 (2019).
3. J. C. Jackson, *et al.*, From text to thought: How analyzing language can advance psychological science. *Perspectives on Psychological Science* **17**, 805–826 (2022).
4. D. M. Lazer, *et al.*, Computational social science: Obstacles and opportunities. *Science* **369**, 1060–1062 (2020).
5. A. Simchon, W. J. Brady, J. J. Van Bavel, Troll and divide: The language of online polarization. *PNAS nexus* **1**, pgac019 (2022).
6. M. Scheffer, I. van de Leemput, E. Weinans, J. Bollen, The rise and fall of rationality in language. *Proceedings of the National Academy of Sciences* **118**, e2107848118 (2021).
7. J. Bollen, *et al.*, Historical language records reveal a surge of cognitive distortions in recent decades. *Proceedings of the National Academy of Sciences* **118**, e2102061118 (2021).
8. T. E. Charlesworth, A. Caliskan, M. R. Banaji, Historical representations of social groups across 200 years of word embeddings from Google Books. *Proceedings of the National Academy of Sciences* **119**, e2121798119 (2022).
9. S. Rathje, J. J. Van Bavel, S. van der Linden, Out-group animosity drives engagement on social media. *Proc Natl Acad Sci USA* **118**, e2024292118 (2021).
10. W. J. Brady, J. A. Wills, J. T. Jost, J. A. Tucker, J. J. Van Bavel, Emotion shapes the diffusion of moralized content in social networks. *Proceedings of the National Academy of Sciences* **114**, 7313–7318 (2017).
11. S. Rathje, C. Robertson, W. Brady, J. J. Van Bavel, People think that social media platforms do (but should not) amplify divisive content. *PsyArXiv* (2022).
12. J. C. Eichstaedt, *et al.*, Facebook language predicts depression in medical records. *Proceedings of the National Academy of Sciences* **115**, 11203–11208 (2018).
13. J. Sterling, J. T. Jost, R. Bonneau, Political psycholinguistics: A comprehensive analysis of the language habits of liberal and conservative social media users. *Journal of Personality and Social Psychology* (2020).
14. H. A. Schwartz, *et al.*, Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one* **8**, e73791 (2013).
15. C. E. Robertson, *et al.*, Negativity drives online news consumption. *Nature Human Behaviour*, 1–11 (2023).
16. J. Singh, G. Singh, R. Singh, Optimization of sentiment analysis using machine learning classifiers. *Human-centric Computing and information Sciences* **7**, 1–12 (2017).
17. D. Antypas, A. Preece, J. Camacho-Collados, Negativity spreads faster: A large-scale multilingual twitter analysis on the role of sentiment in political communication. *Online Social Networks and Media* **33**, 100242 (2023).
18. R. Fan, K. Xu, J. Zhao, Weak ties strengthen anger contagion in social media. *arXiv preprint arXiv:2005.01924* (2020).
19. S. Vosoughi, D. Roy, S. Aral, The spread of true and false news online. *Science* **359**, 1146–1151 (2018).
20. M. Zampieri, *et al.*, Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval). *arXiv preprint arXiv:1903.08983* (2019).
21. M. Mooijman, J. Hoover, Y. Lin, H. Ji, M. Dehghani, Moralization in social networks and the emergence of violence during protests. *Nature human behaviour* **2**, 389–396 (2018).
22. W. J. Brady, K. McLoughlin, T. N. Doan, M. Crockett, How social learning amplifies moral

outrage expression in online social networks (2021).
23. J. A. Frimer, *et al.*, Incivility is rising among American politicians on Twitter. *Social Psychological and Personality Science*, 19485506221083812 (2022).
24. D. Kumar, J. Hancock, K. Thomas, Z. Durumeric, Understanding the Behaviors of Toxic Accounts on Reddit (2023).
25. A. Ashokkumar, J. W. Pennebaker, Social media conversations reveal large psychological shifts caused by COVID-19's onset across US cities. *Science advances* **7**, eabg7843 (2021).
26. Y. R. Tausczik, J. W. Pennebaker, The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of language and social psychology* **29**, 24–54 (2010).
27. X. Yu, M. Wojcieszak, A. Casas, Affective polarization on social media: In-party love among American politicians, greater engagement with out-party hate among ordinary users (2021).
28. P. Saha, *et al.*, On the rise of fear speech in online social media. *Proceedings of the National Academy of Sciences* **120**, e2212270120 (2023).
29. F. A. Acheampong, H. Nunoo-Mensah, W. Chen, Transformer models for text-based emotion detection: a review of BERT-based approaches. *Artif Intell Rev* **54**, 5789–5829 (2021).
30. B. Thompson, S. G. Roberts, G. Lupyan, Cultural influences on word meanings revealed through large-scale semantic alignment. *Nature Human Behaviour* **4**, 1029–1038 (2020).
31. A. Conneau, *et al.*, Unsupervised Cross-lingual Representation Learning at Scale (2020) https:/doi.org/10.48550/arXiv.1911.02116 (April 25, 2023).
32. J. Henrich, S. J. Heine, A. Norenzayan, The weirdest people in the world? *Behavioral and brain sciences* **33**, 61–83 (2010).
33. S. Ghai, It's time to reimagine sample diversity and retire the WEIRD dichotomy. *Nature Human Behaviour* **5**, 971–972 (2021).
34. T. B. Brown, *et al.*, Language Models are Few-Shot Learners (2020) https:/doi.org/10.48550/arXiv.2005.14165 (May 8, 2023).
35. J. Wei, *et al.*, Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652* (2021).
36. D. M. Katz, M. J. Bommarito, S. Gao, P. Arredondo, Gpt-4 passes the bar exam. *Available at SSRN 4389233* (2023).
37. C. Ziems, *et al.*, Can Large Language Models Transform Computational Social Science?
38. M. M. Amin, E. Cambria, B. W. Schuller, Will affective computing emerge from foundation models and general ai? A first evaluation on chatgpt. *arXiv preprint arXiv:2303.03186* (2023).
39. C. Bail, Can Generative AI Improve Social Science? *SocArchiv* (2023).
40. P. Törnberg, ChatGPT-4 Outperforms Experts and Crowd Workers in Annotating Political Twitter Messages with Zero-Shot Learning. *arXiv preprint arXiv:2304.06588* (2023).
41. E. Hoes, S. Altay, J. Bermeo, Using ChatGPT to Fight Misinformation: ChatGPT Nails 72% of 12,000 Verified Claims (2023).
42. P. Y. Wu, J. A. Tucker, J. Nagler, S. Messing, Large Language Models Can Be Used to Estimate the Ideologies of Politicians in a Zero-Shot Learning Setting. *arXiv preprint arXiv:2303.12057* (2023).
43. J. G. Voelkel, R. Willer, Artificial Intelligence Can Persuade Humans on Political Issues.
44. D. Dillon, N. Tandon, Y. Gu, K. Gray, Can AI language models replace human participants? *Trends in Cognitive Sciences* (2023).
45. M. Binz, E. Schulz, Using cognitive psychology to understand GPT-3. *Proceedings of the National Academy of Sciences* **120**, e2218523120 (2023).
46. R. Marjieh, I. Sucholutsky, P. van Rijn, N. Jacoby, T. L. Griffiths, What Language Reveals about Perception: Distilling Psychophysical Knowledge from Large Language Models. *arXiv*

preprint arXiv:2302.01308 (2023).
47. S. Rosenthal, N. Farra, P. Nakov, SemEval-2017 Task 4: Sentiment Analysis in Twitter in *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, (Association for Computational Linguistics, 2017), pp. 502–518.
48. K. Elshakankery, M. Fayek, M. Farouk, LASTD: A Manually Annotated and Tested Large Arabic Sentiment Tweets Dataset in *2021 the 5th International Conference on Information System and Data Mining*, (2021), pp. 62–66.
49. M. Zampieri, *et al.*, SemEval-2020 task 12: Multilingual offensive language identification in social media (OffensEval 2020). *arXiv preprint arXiv:2006.07235* (2020).
50. S. M. Mohammad, P. D. Turney, Nrc emotion lexicon. *National Research Council, Canada* **2**, 234 (2013).
51. , Statistics of Common Crawl Monthly Archives by commoncrawl (May 7, 2023).
52. S. H. Muhammad, *et al.*, SemEval-2023 Task 12: Sentiment Analysis for African Languages (AfriSenti-SemEval) (2023) (May 7, 2023).
53. M. A. Al-Shabi, Evaluating the performance of the most important Lexicons used to Sentiment analysis and opinions Mining. *IJCSNS* **20**, 1 (2020).
54. A. Magueresse, V. Carles, E. Heetderks, Low-resource languages: A review of past work and future challenges. *arXiv preprint arXiv:2006.07264* (2020).
55. D. Rozado, R. Hughes, J. Halberstadt, Longitudinal analysis of sentiment and emotion in news media headlines using automated labelling with Transformer language models. *Plos one* **17**, e0276367 (2022).
56. F. Barbieri, J. Camacho-Collados, L. Neves, L. Espinosa-Anke, TweetEval: Unified Benchmark and Comparative Evaluation for Tweet Classification (2020) (April 21, 2023).
57. S. Mohammad, F. Bravo-Marquez, M. Salameh, S. Kiritchenko, SemEval-2018 Task 1: Affect in Tweets in *Proceedings of the 12th International Workshop on Semantic Evaluation*, (Association for Computational Linguistics, 2018), pp. 1–17.
58. B. Wilie, *et al.*, IndoNLU: Benchmark and Resources for Evaluating Indonesian Natural Language Understanding in *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, (Association for Computational Linguistics, 2020), pp. 843–857.
59. M. S. Saputri, R. Mahendra, M. Adriani, Emotion Classification on Indonesian Twitter Dataset in *2018 International Conference on Asian Language Processing (IALP)*, (2018), pp. 90–95.