Original Paper

A Comparison of ChatGPT and Fine-Tuned Open Pre-Trained Transformers (OPT) Against Widely Used Sentiment Analysis Tools: Sentiment Analysis of COVID-19 Survey Data

Juan Antonio Lossio-Ventura[1], PhD; Rachel Weger[2], BA; Angela Y Lee[3], MA; Emily P Guinee[1], BA; Joyce Chung[1], MD; Lauren Atlas[4], PhD; Eleni Linos[5], MD; Francisco Pereira[1], PhD

[1]National Institute of Mental Health, National Institutes of Health, Bethesda, MD, United States

[2]School of Medicine, University of Pittsburgh, Pittsburgh, PA, United States

[3]Department of Communication, Stanford University, Stanford, CA, United States

[4]National Center For Complementary and Alternative Medicine, National Institutes of Health, Bethesda, MD, United States

[5]School of Medicine, Stanford University, Stanford, CA, United States

Corresponding Author:
Juan Antonio Lossio-Ventura, PhD
National Institute of Mental Health
National Institutes of Health
3D41
10 Center Dr
Bethesda, MD, 20814
United States
Phone: 1 3018272632
Email: juan.lossio@nih.gov

Abstract

Background: Health care providers and health-related researchers face significant challenges when applying sentiment analysis tools to health-related free-text survey data. Most state-of-the-art applications were developed in domains such as social media, and their performance in the health care context remains relatively unknown. Moreover, existing studies indicate that these tools often lack accuracy and produce inconsistent results.

Objective: This study aims to address the lack of comparative analysis on sentiment analysis tools applied to health-related free-text survey data in the context of COVID-19. The objective was to automatically predict sentence sentiment for 2 independent COVID-19 survey data sets from the National Institutes of Health and Stanford University.

Methods: Gold standard labels were created for a subset of each data set using a panel of human raters. We compared 8 state-of-the-art sentiment analysis tools on both data sets to evaluate variability and disagreement across tools. In addition, few-shot learning was explored by fine-tuning Open Pre-Trained Transformers (OPT; a large language model [LLM] with publicly available weights) using a small annotated subset and zero-shot learning using ChatGPT (an LLM without available weights).

Results: The comparison of sentiment analysis tools revealed high variability and disagreement across the evaluated tools when applied to health-related survey data. OPT and ChatGPT demonstrated superior performance, outperforming all other sentiment analysis tools. Moreover, ChatGPT outperformed OPT, exhibited higher accuracy by 6% and higher *F*-measure by 4% to 7%.

Conclusions: This study demonstrates the effectiveness of LLMs, particularly the few-shot learning and zero-shot learning approaches, in the sentiment analysis of health-related survey data. These results have implications for saving human labor and improving efficiency in sentiment analysis tasks, contributing to advancements in the field of automated sentiment analysis.

(JMIR Ment Health 2024;11:e50150) doi: 10.2196/50150

KEYWORDS

sentiment analysis; COVID-19 survey; large language model; few-shot learning; zero-shot learning; ChatGPT; COVID-19

Introduction

Background

Sentiment analysis is a field within natural language processing (NLP) that aims to extract sentiments and opinions from text related to specific entities and topics [1], such as people, organizations, events, and places [2]. Specifically, we consider the task of classifying texts as positive, neutral, or negative. Research in this area can occur at different levels of granularity, ranging from a single sentiment for an entire document or for each sentence within it to exploring various aspects associated with each entity, which can be associated with different sentiments [1,3].

Recently, we have witnessed an increase in the use of sentiment analysis to computationally evaluate the attitudes, perceptions, and emotions of social media users regarding the COVID-19 pandemic [4,5]. Most of these works study content from social media platforms such as Twitter, Reddit, and Facebook [6], as social media has been a main platform to express opinions related to COVID-19 in a public manner. Simultaneously, surveys, which refer to data collected from a group of people regarding their opinions, behavior, or knowledge through specifically designed questions, have also been used to investigate the impact of the COVID-19 pandemic. In particular, surveys conducted during the lockdown period in 2020 examined the effects on people’s lives, behaviors, and mental health, among other topics [7-9]. Web-based surveys are often semistructured, that is, composed of closed-answer components (eg, different clinical questionnaires) and open-ended questions that allow a free-text answer. Sentiment analysis tools have been applied to the latter to help monitor the attitudes, sentiments, and perceptions of the participants during the pandemic to assist health decision-making [10].

The application of sentiment analysis tools on free-text data obtained from surveys poses challenges for health care providers and researchers in the health domain. This is partly attributed to the fact that most state-of-the-art applications are designed for different domains, such as social media, and there is limited knowledge regarding their performance in survey data. In addition, recent studies have applied the most well-known sentiment analysis tools, including TextBlob [11], VADER (Valence Aware Dictionary and Sentiment Reasoner) [12], and Stanza [13], to analyze health-related content on social media platforms [14-16] and, more recently, in the context of COVID-19 [6,17]. These studies highlighted the need for a more comprehensive evaluation of sentiment analysis tools, as the initial results exhibited a lack of accuracy and yielded inconsistent outcomes [15,16]. The main reason for this discrepancy was the disparity in data sets and the potential sensitivity of the tools to the composition of the data set [16]. Consequently, researchers trained new algorithms tailored to their specific data set.

Two COVID-19 survey data sets were used in this study, both collected by teams from the National Institutes of Health (NIH) and Stanford University. The collected data were used to assess the general topics experienced by the participants during the pandemic lockdown. Researchers from both institutions aimed to comprehend the general sentiment patterns over time and identify an overall sentiment for events during that period, such as vaccines and the 2020 presidential elections. In both data sets, it was often the case that a complete response contained multiple topics, with many sentences referring to distinct subjects. Thus, this study is focused on the analysis of sentiment at the sentence level. By assessing each sentence independently, subtle shifts in sentiment could be captured, which could potentially be neglected at the document level. Moreover, we thought that an analysis based on sentence level, rather than aspect-based level, was more appropriate, given that our focus was not on the granularity of the various aspects of an entity. For instance, when evaluating different features of an intensive care unit, aspects might encompass ventilators, rooms, staff, nurses, and others. Therefore, the decision to focus on sentence-level sentiment analysis is influenced by practical considerations, our research objectives, and the nature of the survey responses.

In this study, as the first contribution, we analyzed 2 independent survey data sets containing free-text data collected during the lockdown period of the COVID-19 pandemic, with accompanying ground-truth sentiment labels generated by human raters for hundreds of responses. The second contribution involves a comparison of 8 widely used state-of-the-art sentiment analysis tools, which have been frequently and recently used in the health domain [16], on COVID-19 surveys at the sentence level. We demonstrate that performance across tools varies and that there is a complex correlation structure between their predicted polarity scores. The third contribution of this paper is to investigate whether the polarity prediction performance can be improved through few-shot learning on a small labeled data set or zero-shot learning with ChatGPT [18].

Related Work

There are 2 main approaches to performing sentiment analysis: lexicon based and machine learning based. Initial lexicon methods are the simplest rule-based methods and seek to classify the sentiment of a sentence as a score function of the word polarities existing in a dictionary [19-23]. Lexicon-based techniques use mostly adjectives and adverbs to compute the overall sentiment score of a text, for instance, Linguistic Inquiry and Word Count (LIWC) [24], Affective Norms for English Words [25], and SentiWordNet [26]. Dictionaries of lexicons are created either manually or automatically [27,28]. First, a list is generated from a specific domain. Then synonyms and antonyms are added from other existing dictionaries such as WordNet [29]. More sophisticated lexicon-based methods focus on complex rules, such as regular expressions [30,31], instead of simply computing a sentiment score based on word polarities.

Machine learning–based techniques use statistical methods to compute sentiment polarity. The process involves training a classifier on a labeled data set, such as movie reviews or social media posts, and then using the model to predict the sentiment of new, unlabeled data. Obtaining labeled data to train the classifiers is a time-consuming task. Machine learning–based methods often face challenges when processing negative and intensifying statements and can have low performance when applied to different domains, as they rely mainly on the data set size. The rules proposed in the lexicon-based approaches have also been used to extract relevant features and used as input to machine learning algorithms (eg, naive Bayes, *k*-nearest neighbors, decision tree, and logistic regression) to predict the sentiment [32-37]. Other machine learning methods are based on deep neural networks (DNNs). DNNs have been successfully used for sentiment analysis, as described in detail by Birjali et al [3], Zhang et al [38], and Yadav and Vishwakarma [39], having achieved state-of-the-art performance on several benchmarks. DNN architectures used include recurrent neural networks [40,41], long short-term memory networks [42,43], and convolutional neural networks [44-46].

More recently, transformers [47] (deep learning architectures) and large language models (LLMs) have gained popularity due to their ability to perform NLP tasks, including sentiment analysis, with remarkable performance. These LLMs have been pretrained on large text corpora using transformers, such as Bidirectional Encoder Representations from Transformers (BERT) [48], Robustly optimized BERT approach (RoBERTa) [49], Embeddings from Language Models (ELMo) [50], Generative Pre-trained Transformers (GPT) [51], and Pathways Language Model (PaLM) [52]. LLMs in sentiment analysis can handle several data types and domains as well as identify patterns and relationships between the semantics of words and phrases that are indicative of sentiment. LLMs for sentiment analysis can also be fine-tuned to specific domains and applications, which usually lead to better results, as shown in previous studies [53-59]. Finally, ChatGPT (OpenAI) [18] has suddenly emerged to produce human-like responses to user inputs. The notable performance of LLMs has led to increased interest in few-shot and zero-shot learning methods using them. Few-shot learning algorithms enable a model to learn from only a few examples, whereas zero-shot learning algorithms can transfer knowledge from one task to another without additional labeled training examples. These approaches have demonstrated comparable or superior performance to prior state-of-the-art fine-tuning methods on various NLP tasks [60-62].

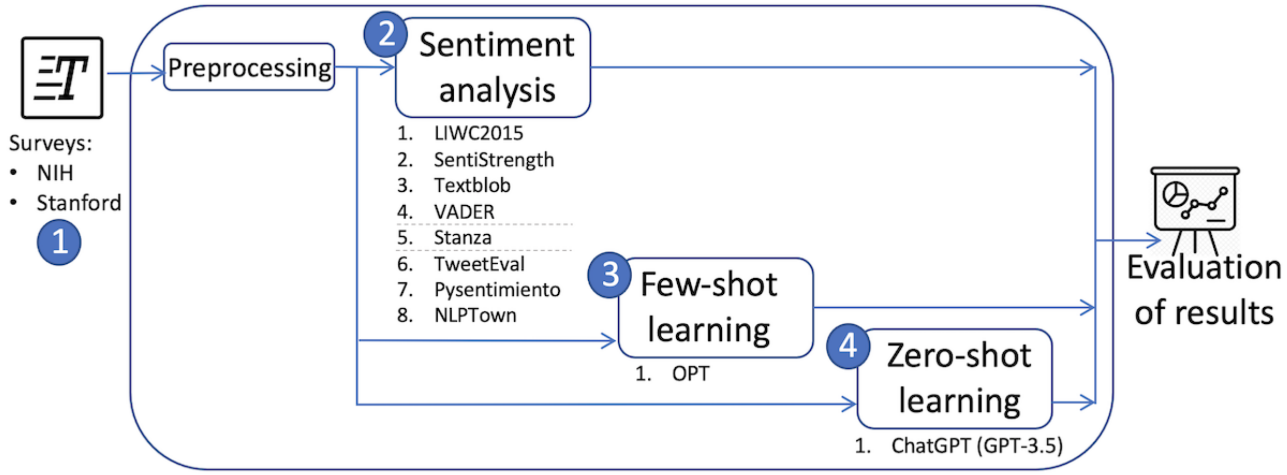
Sentiment analysis has become an increasingly popular technique in the health domain, as noted in the study by Rodríguez-Ibánez et al [63]. A recent study [64] also found that the main data source for studies on health is social media, such as Twitter and Facebook. This is attributed to advancements in mobile technology and their use as a source in health-related topics, such as finding treatments, sharing experiences and opinions, and addressing public health surveillance issues [65-67]. During the pandemic, we witnessed social media becoming the main forum to express opinions related to COVID-19, which helped authorities to understand and monitor sentiments toward topics related to the pandemic [68-73].

Various studies have proposed new sentiment analysis methods and compared existing tools (eg, TextBlob [74], VADER [12], and Stanza [13]) on topics related to COVID-19, mainly extracted from social media [6,16,17,75-78]. However, to the best of our knowledge, there are no studies that have compared several sentiment analysis tools on health-related surveys—a more structured type of text data than social media posts—that collected knowledge, beliefs, and habits during the COVID-19 pandemic [79-84]. The only study we are aware of that evaluates ChatGPT on various sentiment analysis tasks, comparing it with fine-tuned BERT, is the study by Wang at al [85]. The results demonstrated that ChatGPT exhibited promising zero-shot sentiment analysis ability, achieving performance on par with fine-tuned BERT and state-of-the-art models. However, it fell slightly behind domain-specific fully supervised state-of-the-art models.

Methods

This section presents the data sets used in this study along with our evaluation of sentence sentiment analysis methods, as illustrated in Figure 1. Specifically, we describe the (1) survey data sets, (2) state-of-the-art sentiment analysis tools, (3) few-shot learning with an LLM, and (4) zero-shot learning with ChatGPT.

Figure 1. Workflow of our study for evaluating sentence sentiment analysis using state-of-the-art sentiment analysis tools, few-shot learning with a large language model, and zero-shot learning with ChatGPT over health-related surveys. GPT: Generative Pre-trained Transformers; LIWC2015: Linguistic Inquiry and Word Count 2015; NIH: National Institutes of Health; OPT: Open Pre-Trained Transformers; VADER: Valence Aware Dictionary and Sentiment Reasoner.



Data

NIH Data Set

This data set was collected as part of a web-based survey assessing mental health during the pandemic, which started from April 2020 to May 2021. This was a sample of convenience, as participants were recruited from a pool of previous participants in the National Institute of Mental Health and National Center for Complementary and Alternative Medicine studies by advertising on social media and by flyers within the Washington metropolitan area. Participants who signed up completed various questionnaires at baseline, assessing demographics, clinical history, and psychological state [86]. The participants were then sent emails every 2 weeks for 6 months, inviting them to complete 3 of those questionnaires at that time. This latter survey consisted of 45 questions assessing various attitudes, behaviors, and impacts surrounding the pandemic and a single free-response question ("Is there anything else you would like to tell us that might be important that we did not ask about?"). There was a maximum of 13 potential survey (and free) responses per participant. Of the 3655 participants who enrolled in the study, 2497 (68.31%) responded at least once to the free-response item, yielding a total of 9738 item responses. These were composed of 26,411 sentences, which were the data used in this study. The semantic content of these responses (eg, main topics of concern over time) is available in the study by Weger et al [87].

Stanford Data Set

This data set was collected as part of a web-based survey conducted from March to September 2020 by a Stanford University team. The survey was conducted using a sample of convenience recruited through 3 social media platforms: Twitter, Facebook, and Nextdoor. They could participate by clicking on a survey link in the social media post upon seeing the recruitment materials. The survey comprised 21 questions including demographics and the impact of COVID-19 on individuals' lives [88]. In this study, we focus on the evaluation of 3 free-text responses to the following questions: (1) "Although this is a challenging time, can you tell us about any positive effects or 'silver linings' you have experienced during this crisis?" (2) "What are the reasons you are not self-isolating more?" and (3) "Have you experienced any difficulties due to the coronavirus crisis?." Of the 4582 participants recruited, 3349 (73.09%) responded to at least 1 of the 3 free-text questions, resulting in a total of 7182 item responses. These were composed of approximately 21,266 sentences, which were the data used in this study. The topics and sentiments in these responses are reported in the study by Lossio-Ventura et al [10]. Table 1 presents additional details regarding the NIH and Stanford data sets.

Table 1. Details of the National Institutes of Health (NIH) and Stanford data sets.

	NIH	Stanford
Start of the collection period	April 2020	March 2020
End of the collection period	May 2021	September 2020
Responders, n/N (%)	2497/3655 (68.31)	3349/4582 (73.09)
Response items, n	9738	7182
Sentences before processing, n	26,411	21,266
Sentences after processing, n/N (%)	26,188/26,411 (99.16)	21,035/21,266 (98.91)
Tokens after processing, n	462,518	299,735
Tokens per sentence, mean (SD)	17.66 (11.11)	14.25 (9.74)

Annotation

We created training and test sets for both the NIH and Stanford data sets. These sets were derived from the surveys after completing the preprocessing steps and were used for training, tuning, and the official evaluation.

Training Data Set

We randomly selected 260 sentences, with 130 sentences from each data set. Each subset of 130 sentences was annotated by a different annotator. The annotators were instructed to assign a polarity value of −1 (negative), 0 (neutral), or 1 (positive) to each sentence.

Test Data Set

A total of 1000 sentences were randomly chosen, with 500 sentences selected from each data set [89]. Each set was annotated by 3 separate and independent annotators: A.1, A.2, and A.3 for NIH and A.4, A.5, and A.6 for Stanford. The annotators were instructed to assess the polarity of each sentence on a scale of −1 (negative), 0 (neutral), or 1 (positive).

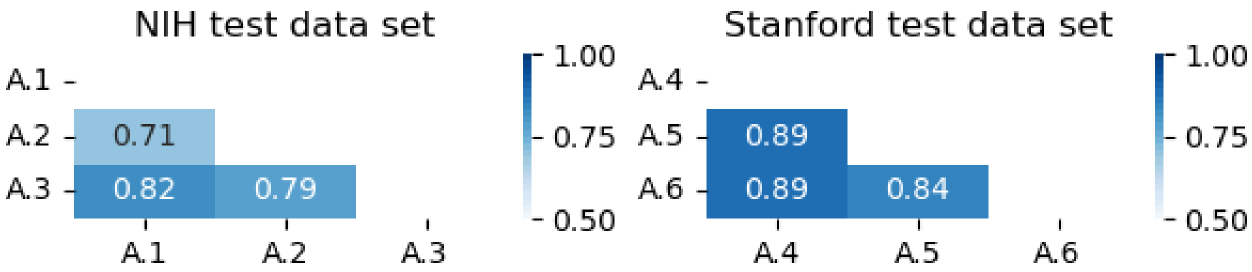
We used a 3-point scale to annotate the data. We then followed a 3-step procedure to determine the final labels, similar to that described in the studies by Nakov et al [90] and Rosenthal et al [91]. First, if all 3 annotators agreed on a label (full agreement), that label was accepted. Second, if 2 of the 3 agreed on a label (partial agreement), that label was also accepted. Third, if there was no agreement, the label was set as neutral (no agreement). Fleiss κ measure was calculated to assess the agreement between the 3 annotators of each test data set. The associated *P* values were computed to test if the agreement between annotators was substantially better than what would be expected by chance. Further details of the training and test data sets are provided in Table 2. Pearson correlation coefficients were also calculated to evaluate the degree of agreement between each pair of annotators, as shown in Figure 2.

Table 2. Details of the National Institutes of Health (NIH) and Stanford data sets.

	Training (n=130)		Test (n=500)	
	NIH	Stanford	NIH	Stanford
Sentences, n (%)	130 (100)	130 (100)	500 (100)	500 (100)
Negative sentences, n (%)	71 (54.6)	45 (34.6)	223 (44.6)	234 (46.8)
Neutral sentences, n (%)	51 (39.2)	41 (31.6)	232 (46.4)	117 (23.4)
Positive sentences, n (%)	8 (6.2)	44 (33.8)	45 (9)	149 (29.8)
Full agreement, n (%)	N/A[a]	N/A	340 (68)	385 (77)
Partial agreement, n (%)	N/A	N/A	159 (31.8)	112 (22.4)
No agreement, n (%)	N/A	N/A	1 (0.2)	3 (0.6)
Fleiss κ	N/A	N/A	0.6311	0.7572
P value	N/A	N/A	<.001	<.001

[a] N/A: not applicable.

Figure 2. Correlation of annotators on the National Institutes of Health (NIH) and Stanford test data sets. A.1, A.2, A.3 represent the 3 independent NIH annotators, while A.4, A.5, A.6 represent the Stanford annotators.



Preprocessing

The survey responses contained personal identifiable information and multiple sentences covering different themes, for example, 2020 presidential elections and COVID-19 vaccines. Therefore, preprocessing steps included splitting responses into sentences, replacing people's names, suppressing email addresses, and lemmatizing and converting text to lower case.

Sentiment Analysis Applications

We considered popular sentiment analysis applications available on the internet that use rules, machine learning, and fine-tuned LLMs.

Linguistic Inquiry and Word Count 2015

Linguistic Inquiry and Word Count 2015 (LIWC2015) [24,92,93] is a text analysis software that identifies and calculates the frequency of different categories of words in texts, such as pronouns, emotional words, cognitive words, and social words. LIWC2015 seeks to group words into categories that can be used to analyze psycholinguistic features in texts. Researchers in various fields, including psychology, sociology, and computer science, have used LIWC2015 to study a wide range of topics, such as personality, emotional expression, deception, and social interaction. LIWC2015 has also been used in various relevant studies on sentiment analysis. It provides with a summary variable "Tone" that combines positive and negative dimensions (*posemo* and *negemo*) into a single one. The higher the tone, the more positive it is. The tone ranges from 0 to 100. Numbers <50 indicate a more negative emotional tone. The default LIWC2015 Dictionary contains approximately 6400 words, word stems, and select emoticons.

SentiStrength

SentiStrength is a sentiment analysis tool that assigns scores to words and phrases based on their positive or negative sentiment [94-96]. It calculates an overall sentiment score for the text by combining these individual scores. This tool can provide dual-, binary-, trinary-, or single-scale results. In this study, a single scale ranging from −4 (extremely negative) to 4 (extremely positive) was chosen, with 0 indicating neutral sentiment. SentiStrength uses linguistic and lexicon-based methods. Linguistic methods involve rules and heuristics for identifying sentiment-bearing words and phrases, including cues such as repeated punctuation, emoticons, negations, and capital letters. The lexicon used consists of 2546 terms associated with polarity and intensity. Part of the lexicon was added from General Inquirer, including word roots such as "extrem*" to recognize variants. Training data sets included posts from various platforms such as BBC Forum, Twitter, YouTube, Digg.com, MySpace, and Runners World.

TextBlob

TextBlob is a Python library used in NLP tasks [11,74], such as part-of-speech tagging, sentiment analysis, and noun phrase extraction. TextBlob outputs a polarity score ranging from −1 to 1. A negative score signifies a negative sentiment, a positive score indicates a positive sentiment, and a score of 0 represents a neutral sentiment. TextBlob includes 2 analysis approaches: a rule-based model and a supervised machine learning naïve Bayes classifier model.

VADER

VADER [12,97] is a rule-based model designed for analyzing sentiment in social media text. It uses 5 rules based on grammatical and syntactical patterns to determine sentiment intensity. These rules involve punctuation, capitalization, degree modifiers, conjunctions such as “but,” and trigram evaluation to identify negations that can affect polarity. VADER was developed and validated using a gold standard list of lexical features, including LIWC, General Inquirer, and Affective Norms for English Words. The model was trained on various data sets, including tweets, New York Times opinions, movie reviews, and Amazon product reviews.

Stanza

Stanza is an open-source Python library that provides several methods for performing NLP tasks [13,98], including part-of-speech tagging, named entity recognition, dependency parsing, and sentiment analysis. Stanza’s sentiment analysis module assigns a positive, negative, or neutral sentiment score (0, 1, or 2, respectively) to each sentence in a given text. Stanza’s sentiment analysis tool is based on a convolutional neural network model using the vectors trained by Mikolov et al [99] on 100 billion words from Google News as well as a combination of lexical and syntactic features. It was trained on large data sets including movie reviews and the Stanford Sentiment Treebank. Unlike other methods, Stanza includes preprocessing of its own (sentence splitter and tokenizer).

TweetEval

TweetEval is a benchmarking platform for Twitter-specific classification tasks [100]. TweetEval consists of 7 NLP tasks: irony detection, offensive language detection, emoji prediction, emotion recognition, hate speech detection, stance detection, and sentiment analysis. Using TweetEval, a common set of evaluation metrics and data set, researchers and practitioners can compare the performance of different models on the same tasks and identify the most effective models for different NLP applications. TweetEval provides a leaderboard for ranking the performance of different models on the sentiment analysis task. The leaderboard is based on the F_1-score. TweetEval returns 3 labels (positive, negative, and neutral) associated with a weight. TweetEval sentiment analysis is based on the RoBERTa model, an LLM based on BERT (trained on 58M tweets), and fine-tuned on the SemEval 2017 sentiment analysis data set (approximately 40,000 tweets) [91].

Pysentimiento

Pysentimiento is an open-source Python library that includes models for sentiment analysis and social NLP tasks, such as hate speech detection, irony detection, emotion analysis, named entity recognition, and part-of-speech tagging, in several languages such as English, Spanish, Portuguese, and Italian [101,102]. The English model for sentiment analysis is based on BERTweet [103], a RoBERTa model, trained on English tweets and also fine-tuned on the SemEval 2017 sentiment analysis data set [91]. Pysentimiento returns 3 polarity labels per text associated with a weight.

NLPTown

NLPTown [104] is a sentiment analysis application based on a BERT-base-multilingual-uncased model, fine-tuned for sentiment analysis on product reviews for 6 languages (English, Dutch, German, French, Spanish, and Italian), and predicts the sentiment of the review as the number of stars (1-5).

Few-Shot Learning With Open Pre-Trained Transformers Language Models

As mentioned previously, few-shot learning seeks to address the challenge of sentiment analysis when only a small amount of labeled data is available for training. In traditional supervised learning, models are trained on large data sets with many labeled examples. However, in some applications such as sentiment analysis, labeled survey data are scarce or expensive to obtain, making it difficult to train accurate models. In this study, we used the Open Pre-Trained Transformers (OPT) [105], a suite of decoder-only pre-trained transformers ranging from 125M to 175B parameters created by Meta AI. OPT has been used in several applications but has never been applied to sentiment analysis. This model has shown to perform similarly to the GPT-3 [60] on several NLP tasks. The OPT model was built using a data set of 180B tokens. This represents approximately 23% (180B/780B) of the amount of data set tokens used for the Pathways Language Model [52]. The largest OPT model has comparable number of parameters to GPT-3 (175B parameters) [60], although we used all models except for the latter given graphics processing unit limitations. The novelty of OPT is its availability as open source (albeit only for academic research).

Zero-Shot Learning With ChatGPT

Zero-shot learning refers to the use of a model to perform a task for which it has not been explicitly trained. Thus, zero-shot learning for sentiment analysis recognizes and classifies sentiment in text without being explicitly provided with examples of sentiment labels. Instead, the model is trained on related tasks, such as language modeling or machine translation, which enables it to understand the underlying structure of the language and the context in which it is used. In this study, we used ChatGPT (based on GPT-3.5), which has significantly improved the performance of several NLP tasks. GPT-3.5 is a model with 175B parameters created by OpenAI and trained on a vast amount of text data sourced from the internet using both reinforcement and supervised learning techniques. For this paper, we generated a polarity score for each sentence *x* by asking ChatGPT “What is the sentiment of the following sentence ‘x.’”

Ethical Considerations

The NIH survey was approved by the Institutional Review Board of the NIH (reference number 20MN085), and all participants provided consent for the study. The Stanford survey was approved by Stanford’s Institutional Review Board (reference number 55436), and all participants provided consent for the study.

All survey data and responses in both the NIH and the Stanford data sets were anonymized and associated with a unique ID. Participants from both studies were not compensated for participating in the surveys.

Results

Evaluation Metrics

To assess the overall performance of the sentiment analysis tools, we evaluated the accuracy, macro *F*-measure, macro precision, and macro recall. Macro evaluation metrics were recommended in the NLP competition SemEval-2017 Task 4 [91].

Preparation of Applications for Evaluation

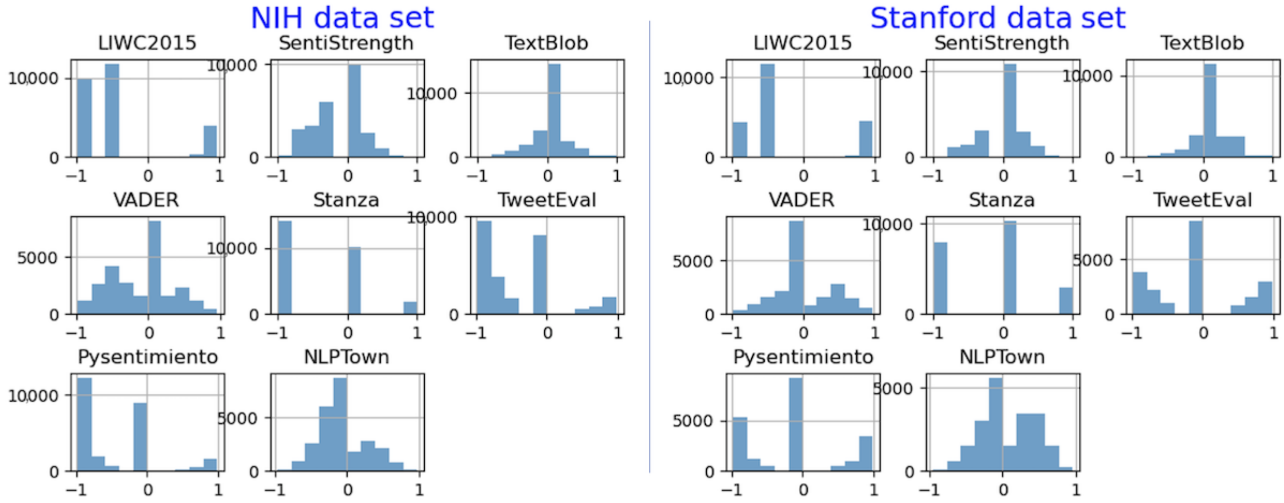
Harmonization of Applications' Outputs

The LIWC2015, Stanza, and SentiStrength applications produce outputs that are measured on distinct scales. LIWC2015 generates a continuous value ranging from 0 to 100; SentiStrength generates an integer score ranging from −4 to 4; and Stanza produces a discrete whole number score of 0, 1, or 2, which correspond to negative, neutral, and positive sentiments, respectively. Therefore, it is necessary to convert these scores to a common range of [−1, 1], as formally defined in equation 1.

$$\text{score'}(x) = 2 \times (\text{score}[x] - \text{score}[x]_{min}) / (\text{score}[x]_{max} - \text{score}[x]_{min}) - 1 \quad \textbf{(1)}$$

The distribution of sentiment scores across all tools is shown in Figure 3. We then classify all negative values as negative sentiment, all 0 values as neutral, and all positive values as positive sentiment. It is important to note that the VADER application uses a slightly different classification approach, considering a score ≤0.05 to be negative, a score between −0.05 and 0.05 to be neutral, and a score ≥0.05 to be positive.

Figure 3. Distribution of sentiment scores across all applications on the National Institutes of Health (NIH) and Stanford data sets. LIWC2015: Linguistic Inquiry and Word Count 2015; VADER: Valence Aware Dictionary and Sentiment Reasoner.



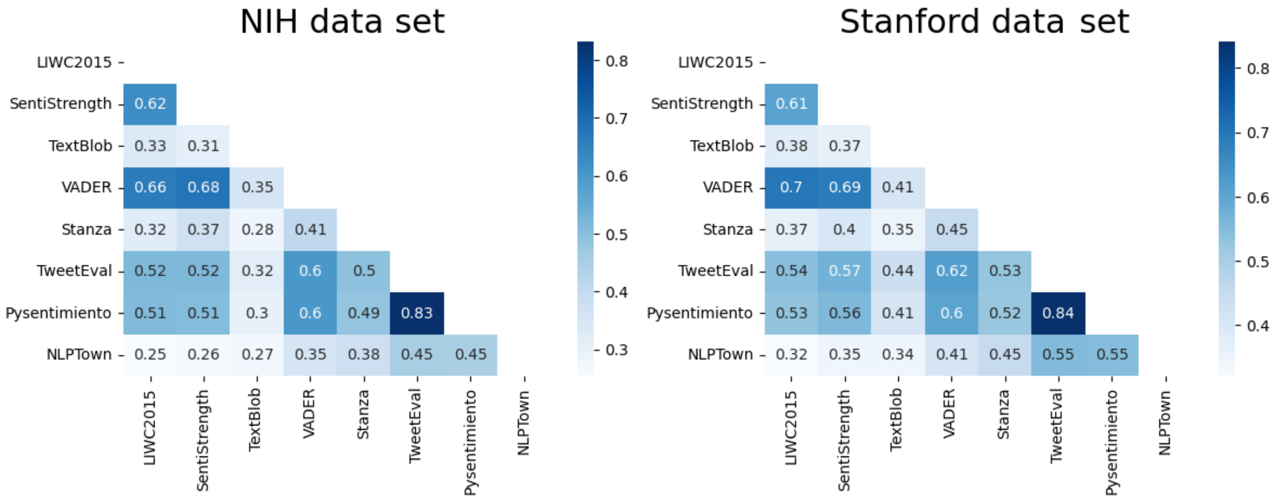
Fine-Tuning for Few-Shot Learning

We used few-shot learning using our small amount of training data to fine-tune the OPT models, rather than training them from scratch. For this experiment, the training data set was split into 85% (110/130) for feeding the model and 15% (20/130) for validation. Given the memory constraints, we considered only OPT 125M, 350M, 1.3B, and 2.7B. We performed a hyperparameter search to optimize the performance of the model on sentiment analysis. We considered learning rate=[3×10^{-4}, 1×10^{-4}, 3×10^{-5}, and 1×10^{-5}], batch size=[4, 8, 16, and 32], number of epochs from 1 to 7, and the AdamW optimizer. The models that performed the best were OPT-1.3B and OPT-2.7B, using a learning rate of 1×10^{-5}, a batch size of 32, and 5 epochs. These were the models used to obtain the test set results reported in next subsections.

Experiment 1: Correlation Between the Outputs of Applications

The objective was to evaluate the agreement level among various methods for predicting the sentiments of COVID-19 survey responses. Understanding the methods' agreement or divergence was crucial in determining the reliability and accuracy of predictions, allowing for accurate studies of the relationship between language use and mental health. The Pearson correlation coefficient was used to assess the reliability of the tools, as shown in Figure 4. Disagreement among the methods prompted us to evaluate few-shot learning to obtain high-quality predictions.

Figure 4. Pearson correlation matrix of score applications on the National Institutes of Health (NIH) and Stanford data sets. LIWC2015: Linguistic Inquiry and Word Count 2015; VADER: Valence Aware Dictionary and Sentiment Reasoner.



Experiment 2: Prediction of Sentiment Scores

Tables 3 and 4 show the performance results obtained by all applications, few-shot learning, and zero-shot learning techniques on the NIH and Stanford test data sets, respectively. Both test sets comprised 500 sentences each, as detailed in the *Data* section. The top 2 performance results are italicized. Of note, a perfect classifier that accurately categorizes all items obtains a value of 1, whereas a perverse classifier that misclassifies all items achieves a value of 0. However, a trivial classifier that assigns all sentences to the same category (positive, negative, or neutral) and a random classifier both have a value of 0.3333.

ChatGPT achieved a significant improvement in sentiment analysis compared with other models through zero-shot learning. On the NIH data set, ChatGPT outperformed few-shot learning (OPT-1.3B and OPT-2.7B) by 6% in accuracy and 7% in *F*-measure. Similarly, on the Stanford data set, ChatGPT showed better results than the OPT-1.3B and OPT-2.7B models, with 6% higher accuracy and 4% higher *F*-measure.

Moreover, to further evaluate the sentiment analysis tools, we used Bayesian analysis, as recommended by Benavoli et al [106], to assess the statistical significance of the performance of the methods. Specifically, we applied the Bayesian signed-rank test [107] to compare the accuracies achieved across multiple data sets. This test quantifies the likelihood of observing the signed ranks of accuracy differences under both the null hypothesis (indicating no significant difference) and alternative hypothesis (indicating a significant difference). The Bayesian signed-rank test is designed to compare performance over multiple data sets (≥2); therefore, we further partitioned the independent Stanford and NIH data sets. Each data set was partitioned into 3 subsets, based on the sentiment label assigned to them, resulting in positive, neutral, and negative subsets for each data set.

This division was influenced by insights from our prior analysis, which highlighted inherent distinctions among sentences associated with positive, neutral, and negative labels. For instance, positive sentences exhibited a preponderance of positive adjectives, whereas negative sentences featured more negative adjectives, and neutral sentences tended to emphasize facts that are characteristic of the neutral category. Therefore, we assumed a degree of independence across subsets within each data set. The heat map diagram in Figure 5 shows the results of our Bayesian analysis, with cells corresponding to row *i* and column *j*. On the left side, “A higher than B” indicates the probability that method *i* performs better than classifier *j*. The center indicates the probability of practical equivalence between methods *i* and *j*. Similarly, on the right side, “B higher than A” indicates the probability that method *j* is better than classifier *i*. These experiments confirmed that ChatGPT performed better than all the other alternatives. The OPT models showed similar performance to methods other than ChatGPT and could be considered as a viable second option.

Table 3. Results on the National Institutes of Health (NIH) test data set.

Application	Precision	Recall	*F*-measure	Accuracy
LIWC2015[a]	0.2733	0.5226	0.3587	0.4540
SentiStrength	0.5732	0.6006	0.5814	0.6480
TextBlob	0.4505	0.4776	0.4053	0.4340
VADER[b]	0.6302	0.7036	0.6097	0.6580
Stanza	0.6178	0.5758	0.5886	0.6300
TweetEval	0.7818	*0.8318* [c]	0.7898	0.7840
Pysentimiento	0.7738	0.7780	0.7699	0.7760
NLPTown	0.4338	0.5173	0.4210	0.4520
OPT[d] 1.3B (few-shot)	0.8032	0.8000	0.7992	0.8000
OPT 2.7B (few-shot)	*0.8061*	0.8040	*0.8050*	*0.8040*
ChatGPT (zero-shot)	*0.8526*	*0.8926*	*0.8668*	*0.8600*
All negative	0.1487	0.3333	0.2056	0.4460
All neutral	0.1547	0.3333	0.2113	0.4640
All positive	0.0300	0.3333	0.0550	0.0900

[a] LIWC2015: Linguistic Inquiry and Word Count 2015.

[b] VADER: Valence Aware Dictionary and Sentiment Reasoner.

[c] Italicization represents the top 2 performance results.

[d] OPT: Open Pre-Trained Transformers.

Table 4. Results on Stanford test data set.

Application	Precision	Recall	*F*-measure	Accuracy
LIWC2015[a]	0.3752	0.4391	0.3890	0.5400
SentiStrength	0.5738	0.5561	0.5335	0.5420
TextBlob	0.4757	0.4872	0.4527	0.4600
VADER[b]	0.5875	0.5919	0.5755	0.5840
Stanza	0.5975	0.4987	0.4859	0.5040
TweetEval	0.7366	0.7178	0.7090	0.7200
Pysentimiento	0.6731	0.6362	0.6267	0.6440
NLPTown	0.5163	0.5192	0.5056	0.5420
OPT[c] 1.3B (few-shot)	*0.8323* [d]	*0.8160*	*0.8211*	*0.8160*
OPT 2.7B (few-shot)	0.8288	0.8100	0.8147	0.8100
ChatGPT (zero-shot)	*0.8632*	*0.8779*	*0.8662*	*0.8740*
All negative	0.1560	0.3333	0.2125	0.4680
All neutral	0.0780	0.3333	0.1264	0.2340
All positive	0.0993	0.3333	0.1531	0.2980

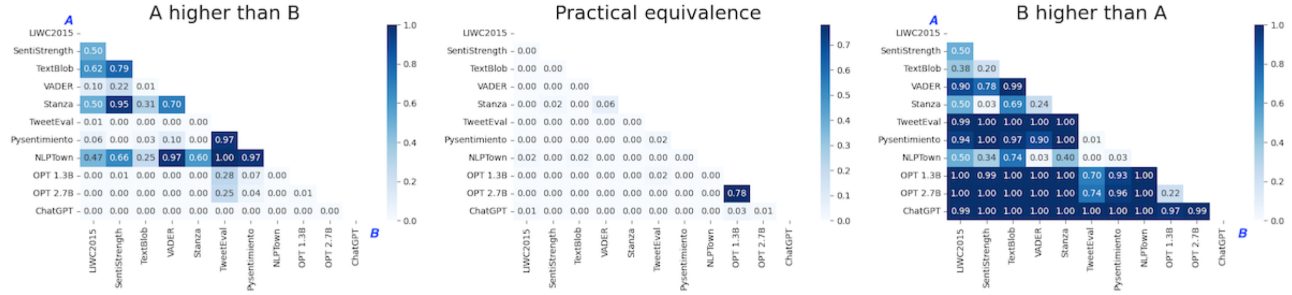
[a] LIWC2015: Linguistic Inquiry and Word Count 2015.

[b] VADER: Valence Aware Dictionary and Sentiment Reasoner.

[c] OPT: Open Pre-Trained Transformers.

[d] Italicization represents the top 2 performance results.

Figure 5. Bayesian analysis conducted on accuracy performances of 11 sentiment analysis methods across 6 different subsets. LIWC2015: Linguistic Inquiry and Word Count 2015; OPT: Open Pre-Trained Transformers; VADER: Valence Aware Dictionary and Sentiment Reasoner.



Discussion

Principal Findings

Our primary objective was to assess various sentiment analysis tools for the purposed of predicting the sentiments of survey responses during the COVID-19 pandemic. Obtaining a thorough understanding of the tools' degree of agreement, as shown in Figure 4, was crucial for determining whether they could be used as surrogates for human labeling. The disagreement between tools led us to try ensemble methods to produce more reliable ratings. Fine-tuned BERT models such as TweetEval and Pysentimiento outperformed the other baseline methods. Fine-tuned methods have the ability to learn domain-specific patterns from text, resulting in better performance than lexicon- and rule-based methods. However, these techniques often require large training data sets to achieve optimal performance, such as the 40k tweet data set used to train TweetEval and Pysentimiento.

As part of the process of determining agreement between tools, we labeled a small data set (260 sentences), which is what prompted us to consider the possibility of using few-shot and zero-shot learning techniques. We then investigated the performance of OPT, which is unexplored in sentiment analysis, for few-shot learning using a small training data set (260 sentences). The OPT-1.3B and OPT-2.7B models surpassed all the baseline methods as well as the fine-tuned BERT models. This highlighted the potential of few-shot learning in dealing with scarce annotated data and the effectiveness of few-shot learning. Although better results could have been achieved with a larger training set, these experiments primarily aimed to investigate the potential of OPT using limited annotated data. The potential is to be able to produce models tailored to specific research applications, with only a small time investment by domain experts. We believe that these models can significantly contribute to the sentiment analysis of health- and clinical-related surveys and can be further fine-tuned with additional data and optimized hyperparameters.

Our investigation also encompassed zero-shot learning with ChatGPT, which exhibited remarkable performance compared with all other models, including few-shot learning with OPT, as presented in Tables 3 and 4. Note that GPT-3.5—the model behind ChatGPT—is trained on related tasks, such as language modeling or machine translation. This enabled it to understand the underlying structure of sentiment-related language and the context in which it is used. Moreover, the necessity for manual text annotations in sentiment analysis tasks makes ChatGPT and other LLMs particularly attractive. As demonstrated by Ziems et al [108], LLMs can alleviate the workload of human annotators in a zero-shot manner, thereby enhancing the efficiency of social-science analysis. In addition, a study [109] found that ChatGPT outperformed crowd workers in various text annotation tasks, including assessing relevance, stance, topics, and frame detection. These findings suggest that there may be potential in using ChatGPT and other recent LLMs for annotation in clinical NLP and reserving human input for quality control. Sentiment analysis tools based on LLMs, such as ChatGPT, automatically identify relevant features, reducing the need for manual engineering, which is a common requirement in tools such as LIWC 2015 and VADER. In addition, LLMs enable fine-tuning, allowing for potential adaptation to different sentiment analysis tasks (eg, in new domains) without the need for complete retraining. LLM-based tools can also capture longer-range context for more accurate sentiment assessment.

Limitations

There exist several limitations and risks of ChatGPT and other non–open-source LLMs regarding protected health information (PHI). Non–open-source LLMs require sending information to an external server and do not provide transparency into how they handle PHI, making it difficult to assess how the model is processing and protecting sensitive information. They may also have security vulnerabilities that can be exploited to gain unauthorized access to PHI. Note also that LLMs are not specifically designed for sentiment analysis, which may sometimes lead to errors, for instance, subtle sarcasm such as "Oh yes, great job!," context-dependent negation as in "The vaccine was not as bad as I thought," and idiomatic expressions such as "It's a piece of cake." They may encounter difficulties with nuanced health-related terminology and concepts. Therefore, specialized health terminology may require additional adaptation beyond general text fine-tuning, for instance, medical abbreviations and acronyms such as "The patient teared up because of a significant increase in their CD4 count" and "So, my mom's HbA1c levels have improved after insulin therapy." In addition, although several outputs may sound plausible, they may occasionally be incorrect. In our view, the output of LLMs should not be used without a plan for human quality control (eg, via sampling) or mitigation (eg, repeated validation). This is crucial for ensuring the accuracy and reliability of the generated content, as LLMs may produce results that require refinement or correction before dissemination. Moreover, there are constraints on the ability to access ChatGPT via its application programming interface, and this may make it too costly or time-consuming to do so. Therefore, researchers and health care practitioners might also opt to use an open-source language model for their NLP-related projects, such as OPT, which can be run on site and perform well on sentiment analysis.

Finally, our study focused on using surveys to understand people's feelings, specifically regarding COVID-19, which was a very important topic at the time. Thus, our conclusions apply specifically to discussions about COVID-19 and may not be true for other subjects. In addition, it is important to highlight that the Stanford data set has an implicit polarity bias: it specifically asks for positive effects ("Although this is a challenging time, can you tell us about any positive effects or 'silver linings' you have experienced during this crisis?") and difficulties ("Have you experienced any difficulties due to the coronavirus crisis?"). The NIH data set poses a single, less-biased question. Therefore, it is crucial to be careful when generalizing our findings beyond the scope of COVID-19 during the studied time frame.

Acknowledgments

The authors would like to thank Kelsey Yeh and Nathaniel Menon, who worked as research assistants in the Social Media Lab at Stanford University, for their assistance in annotating the Stanford test data set. The authors would also like to express gratitude to the many respondents who shared their experiences in both studies. The research reported in this publication was supported in part by the Intramural Research Programs of the National Institute of Mental Health (ZIAMH002922: Principal Investigator JC and ZIC-MH002968: Principal Investigator FP), National Center for Complementary and Integrative Health (ZIA-AT000035: Principal Investigator LA), and National Institute of Arthritis and Musculoskeletal and Skin Diseases (K24AR075060: Principal Investigator EL). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Data Availability

The test data sets generated and analyzed during this study are deidentified and freely available in the FigShare repository [89]. The source code for fine-tuning the OPT models and using ChatGPT in the experiments conducted in this study is publicly accessible on GitHub [110].

Authors' Contributions

JALV and FP contributed to conceiving the study idea and design. LA and JC led the collection of the National Institutes of Health (NIH) data set, whereas EL led the collection of the Stanford data set. The annotation of the training NIH and Stanford data sets was conducted by RW and AYL, respectively. The annotation of the NIH test data set was conducted by LA, RW, and EPG, whereas AYL and 2 research assistants annotated the Stanford test data set. JALV set up the applications and performed the evaluation. JALV and FP wrote the initial draft and revised the subsequent versions. All authors read, revised, and approved the final manuscript.

Conflicts of Interest

None declared.

References

1. Feldman R. Techniques and applications for sentiment analysis. Commun ACM. Apr 2013;56(4):82-89. [FREE Full text] [doi: 10.1145/2436256.2436274]
2. Medhat W, Hassan A, Korashy H. Sentiment analysis algorithms and applications: a survey. Ain Shams Eng J. Dec 2014;5(4):1093-1113. [doi: 10.1016/j.asej.2014.04.011]
3. Birjali M, Kasri M, Beni-Hssane A. A comprehensive survey on sentiment analysis: approaches, challenges and trends. Knowl Based Syst. Aug 2021;226:107134. [doi: 10.1016/j.knosys.2021.107134]
4. Tsao SF, Chen H, Tisseverasinghe T, Yang Y, Li L, Butt ZA. What social media told us in the time of COVID-19: a scoping review. Lancet Digit Health. Mar 2021;3(3):e175-e194. [doi: 10.1016/s2589-7500(20)30315-0]
5. Alamoodi AH, Zaidan BB, Zaidan AA, Albahri OS, Mohammed KI, Malik RQ, et al. Sentiment analysis and its applications in fighting COVID-19 and infectious diseases: a systematic review. Expert Syst Appl. Apr 01, 2021;167:114155. [FREE Full text] [doi: 10.1016/j.eswa.2020.114155] [Medline: 33139966]
6. He L, He C, Reynolds TL, Bai Q, Huang Y, Li C, et al. Why do people oppose mask wearing? a comprehensive analysis of U.S. tweets during the COVID-19 pandemic. J Am Med Inform Assoc. Jul 14, 2021;28(7):1564-1573. [FREE Full text] [doi: 10.1093/jamia/ocab047] [Medline: 33690794]
7. Wang X, Hegde S, Son C, Keller B, Smith A, Sasangohar F. Investigating mental health of US college students during the COVID-19 pandemic: cross-sectional survey study. J Med Internet Res. Sep 17, 2020;22(9):e22817. [FREE Full text] [doi: 10.2196/22817] [Medline: 32897868]

8. Daly M, Sutin AR, Robinson E. Longitudinal changes in mental health and the COVID-19 pandemic: evidence from the UK Household Longitudinal Study. Psychol Med. Oct 2022;52(13):2549-2558. [FREE Full text] [doi: 10.1017/S0033291720004432] [Medline: 33183370]
9. Schuster C, Weitzman L, Sass Mikkelsen K, Meyer-Sahling J, Bersch K, Fukuyama F, et al. Responding to COVID-19 through surveys of public servants. Public Adm Rev. 2020;80(5):792-796. [FREE Full text] [doi: 10.1111/puar.13246] [Medline: 32836447]
10. Lossio-Ventura JA, Lee AY, Hancock JT, Linos N, Linos E. Identifying silver linings during the pandemic through natural language processing. Front Psychol. Sep 3, 2021;12:712111. [FREE Full text] [doi: 10.3389/fpsyg.2021.712111] [Medline: 34539512]
11. TextBlob: simplified text processing. TextBlob. URL: https://textblob.readthedocs.io/en/ [accessed 2022-03-15]
12. Hutto C, Gilbert E. VADER: a parsimonious rule-based model for sentiment analysis of social media text. Proc Int AAAI Conf Web Soc Media. May 16, 2014;8(1):216-225. [doi: 10.1609/icwsm.v8i1.14550]
13. Qi P, Zhang Y, Zhang Y, Bolton J, Manning CD. Stanza: a python natural language processing toolkit for many human languages. Presented at: 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations; July 5-10, 2020, 2020; Online. URL: https://aclanthology.org/2020.acl-demos.14 [doi: 10.18653/v1/2020.acl-demos.14]
14. He L, Zheng K. How do general-purpose sentiment analyzers perform when applied to health-related online social media data? Stud Health Technol Inform. Aug 21, 2019;264:1208-1212. [FREE Full text] [doi: 10.3233/SHTI190418] [Medline: 31438117]
15. He L, Yin T, Hu Z, Chen Y, Hanauer DA, Zheng K. Developing a standardized protocol for computational sentiment analysis research using health-related social media data. J Am Med Inform Assoc. Jun 12, 2021;28(6):1125-1134. [FREE Full text] [doi: 10.1093/jamia/ocaa298] [Medline: 33355353]
16. He L, Yin T, Zheng K. They may not work! an evaluation of eleven sentiment analysis tools on seven social media datasets. J Biomed Inform. Aug 2022;132:104142. [FREE Full text] [doi: 10.1016/j.jbi.2022.104142] [Medline: 35835437]
17. Liu S, Liu J. Public attitudes toward COVID-19 vaccines on English-language Twitter: a sentiment analysis. Vaccine. Sep 15, 2021;39(39):5499-5505. [FREE Full text] [doi: 10.1016/j.vaccine.2021.08.058] [Medline: 34452774]
18. ChatGPT homepage. ChatGPT. URL: https://chat.openai.com/ [accessed 2023-04-17]
19. Zagibalov T, Carroll J. Automatic seed word selection for unsupervised sentiment classification of Chinese text. In: Proceedings of the 22nd International Conference on Computational Linguistics - Volume 1. Presented at: COLING '08; August 18-22, 2008, 2008; Manchester, UK. [doi: 10.3115/1599081.1599216]
20. Ofoghi B, Mann M, Verspoor K. Towards early discovery of salient health threats: a social media emotion classification technique. Presented at: Biocomputing 2016: Proceedings of the Pacific Symposium; January 4-8, 2016, 2016; Kohala Coast, HI. [doi: 10.1142/9789814749411_0046]
21. Davis MA, Zheng K, Liu Y, Levy H. Public response to Obamacare on Twitter. J Med Internet Res. May 26, 2017;19(5):e167. [FREE Full text] [doi: 10.2196/jmir.6946] [Medline: 28550002]
22. Pavlopoulos J, Androutsopoulos I. Multi-granular aspect aggregation in aspect-based sentiment analysis. Presented at: 14th Conference of the European Chapter of the Association for Computational Linguistics; April 26-30, 2014, 2014; Gothenburg, Sweden. [doi: 10.3115/v1/e14-1009]
23. Livas C, Delli K, Pandis N. "My invisalign experience": content, metrics and comment sentiment analysis of the most popular patient testimonials on YouTube. Prog Orthod. Jan 22, 2018;19(1):3. [FREE Full text] [doi: 10.1186/s40510-017-0201-1] [Medline: 29354889]
24. Pennebaker JW, Francis ME, Booth RJ. Linguistic Inquiry and Word Count: LIWC. Mahwah, NJ. Lawrence Erlbaum Associates; 2001.
25. Bradley MM, Lang PJ. Affective norms for English words (ANEW): instruction manual and affective ratings. The Center for Research in Psychophysiology, University of Florida. 1999. URL: https://pdodds.w3.uvm.edu/teaching/courses/2009-08UVM-300/docs/others/everything/bradley1999a.pdf [accessed 2024-01-09]
26. Baccianella S, Esuli A, Sebastiani F. SentiWordNet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In: Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10). Presented at: LREC'10; May 17-23, 2010, 2010; Valletta, Malta. URL: http://www.lrec-conf.org/proceedings/lrec2010/pdf/769
27. Lu Y, Castellanos M, Dayal U, Zhai C. Automatic construction of a context-aware sentiment lexicon: an optimization approach. In: Proceedings of the 20th international conference on World Wide Web. Presented at: WWW '11; March 28-April 1, 2011, 2011; Hyderabad, India. URL: https://doi.org/10.1145/1963405.1963456 [doi: 10.1145/1963405.1963456]
28. Huang S, Niu Z, Shi C. Automatic construction of domain-specific sentiment lexicon based on constrained label propagation. Knowl Based Syst. Jan 2014;56:191-200. [doi: 10.1016/j.knosys.2013.11.009]
29. Miller GA. WordNet: a lexical database for English. Commun ACM. Nov 1995;38(11):39-41. [doi: 10.1145/219717.219748]
30. Pedersen T. Rule-based and lightly supervised methods to predict emotions in suicide notes. Biomed Inform Insight. Jan 30, 2012;5s1 [doi: 10.4137/bii.s8953]
31. Wang W, Chen L, Tan M, Wang S, Sheth AP. Discovering fine-grained sentiment in suicide notes. Biomed Inform Insight. Jan 30, 2012;5s1 [doi: 10.4137/bii.s8963]

32. Desai T, Shariff A, Shariff A, Kats M, Fang X, Christiano C, et al. Tweeting the meeting: an in-depth analysis of Twitter activity at Kidney Week 2011. PLoS One. Jul 5, 2012;7(7):e40253. [FREE Full text] [doi: 10.1371/journal.pone.0040253] [Medline: 22792254]
33. Cole-Lewis H, Varghese A, Sanders A, Schwarz M, Pugatch J, Augustson E. Assessing electronic cigarette-related tweets for sentiment and content using supervised machine learning. J Med Internet Res. Aug 25, 2015;17(8):e208. [FREE Full text] [doi: 10.2196/jmir.4392] [Medline: 26307512]
34. Daniulaityte R, Chen L, Lamy FR, Carlson RG, Thirunarayan K, Sheth A. "When 'bad' is 'good'": identifying personal communication and sentiment in drug-related tweets. JMIR Public Health Surveill. Oct 24, 2016;2(2):e162. [FREE Full text] [doi: 10.2196/publichealth.6327] [Medline: 27777215]
35. Appel O, Chiclana F, Carter J, Fujita H. A hybrid approach to the sentiment analysis problem at the sentence level. Knowl Based Syst. Sep 2016;108:110-124. [doi: 10.1016/j.knosys.2016.05.040]
36. Ahmad M, Aftab S, Ali I. Sentiment analysis of tweets using SVM. Int J Comput Appl. Nov 15, 2017;177(5):25-29. [doi: 10.5120/ijca2017915758]
37. Gupta I, Joshi N. Enhanced Twitter sentiment analysis using hybrid approach and by accounting local contextual semantic. J Intell Syst. 2019;29(1):1611-1625. [doi: 10.1515/jisys-2019-0106]
38. Zhang L, Wang S, Liu B. Deep learning for sentiment analysis: a survey. WIREs Data Min Knowl. 2018;8(4):e1253. [doi: 10.1002/widm.1253]
39. Yadav A, Vishwakarma DK. Sentiment analysis using deep learning architectures: a review. Artif Intell Rev. Dec 02, 2019;53(6):4335-4385. [doi: 10.1007/s10462-019-09794-5]
40. Zhou X, Wan X, Xiao J. Attention-based LSTM network for cross-lingual sentiment classification. Presented at: 2016 Conference on Empirical Methods in Natural Language Processing; November 1-5, 2016, 2016; Austin, Texas. [doi: 10.18653/v1/d16-1024]
41. Xu J, Chen D, Qiu X, Huang X. Cached long short-term memory neural networks for document-level sentiment classification. Presented at: 2016 Conference on Empirical Methods in Natural Language Processing; November 1-5, 2016, 2016; Austin, Texas. URL: https://aclanthology.org/D16-1172 [doi: 10.18653/v1/d16-1172]
42. Ma Y, Peng H, Cambria E. Targeted aspect-based sentiment analysis via embedding commonsense knowledge into an attentive LSTM. Presented at: Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18); February 2-7, 2018, 2018; New Orleans, LA. [doi: 10.1609/aaai.v32i1.12048]
43. Rehman AU, Malik AK, Raza B, Ali W. A hybrid CNN-LSTM model for improving accuracy of movie reviews sentiment analysis. Multimed Tools Appl. Jun 11, 2019;78(18):26597-26613. [doi: 10.1007/s11042-019-07788-7]
44. Ouyang X, Zhou P, Li CH, Liu L. Sentiment analysis using convolutional neural network. Presented at: IEEE International Conference on Computer and Information Technology; Ubiquitous Computing and Communications; Dependable, Autonomic and Secure Computing; Pervasive Intelligence and Computing; October 26-28, 2015, 2015; Liverpool, UK. [doi: 10.1109/cit/iucc/dasc/picom.2015.349]
45. Tang D, Qin B, Liu T. Document modeling with gated recurrent neural network for sentiment classification. Presented at: 2015 Conference on Empirical Methods in Natural Language Processing; September 17-21, 2015, 2015; Lisbon, Portugal. URL: https://aclanthology.org/D15-1167 [doi: 10.18653/v1/d15-1167]
46. Chen T, Xu R, He Y, Wang X. Improving sentiment analysis via sentence type classification using BiLSTM-CRF and CNN. Expert Syst Appl. Apr 2017;72:221-230. [doi: 10.1016/j.eswa.2016.10.065]
47. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AL, et al. Attention is all you need. arXiv. Preprint posted online on June 12, 2017. [FREE Full text]
48. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. Presented at: 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers); June 2-7, 2019, 2019; Minneapolis, MN. URL: https://aclanthology.org/N19-1423 [doi: 10.18653/v1/n18-2]
49. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, et al. RoBERTa: a robustly optimized BERT pretraining approach. arXiv. Preprint posted online on July 26, 2019. [doi: 10.1090/mbk/121/79]
50. Peters ME, Neumann M, Iyyer M, Gardner M, Clark C, Lee K, et al. Deep contextualized word representations. Presented at: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers); June 1-6, 2018, 2018; New Orleans, LA. URL: https://aclanthology.org/N18-1202 [doi: 10.18653/v1/n18-1202]
51. Radford A, Narasimhan K, Salimans T, Sutskever I. Improving language understanding by generative pre-training. OpenAI. URL: https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf [accessed 2024-01-08]
52. Chowdhery A, Narang S, Devlin J, Bosma M, Mishra G, Roberts A, et al. PaLM: scaling language modeling with pathways. arXiv. Preprint posted online on April 5, 2022.
53. Sun C, Huang L, Qiu X. Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentence. arXiv. Preprint posted online on March 22, 2019. [FREE Full text] [doi: 10.18653/v1/n18-2043]

54. Bataa E, Wu J. An investigation of transfer learning-based sentiment analysis in Japanese. Presented at: 57th Annual Meeting of the Association for Computational Linguistics; July 28-August 2, 2019, 2019; Florence, Italy. URL: https://aclanthology.org/P19-1458 [doi: 10.18653/v1/p19-1458]
55. Yin D, Meng T, Chang KW. SentiBERT: a transferable transformer-based architecture for compositional sentiment semantics. Presented at: 58th Annual Meeting of the Association for Computational Linguistics; July 5-10, 2020, 2020; Online. URL: https://aclanthology.org/2020.acl-main.341 [doi: 10.18653/v1/2020.acl-main.341]
56. Büyüköz B, Hürriyetoğlu A, Özgür A. Analyzing ELMo and DistilBERT on socio-political news classification. Presented at: Workshop on Automated Extraction of Socio-political Events from News 2020; May 12, 2020, 2020; Marseille, France. URL: https://aclanthology.org/2020.aespen-1.4
57. Zhang L, Fan H, Peng C, Rao G, Cong Q. Sentiment analysis methods for HPV vaccines related tweets based on transfer learning. Healthcare (Basel). Aug 28, 2020;8(3):307. [FREE Full text] [doi: 10.3390/healthcare8030307] [Medline: 32872330]
58. Dai J, Yan H, Sun T, Liu P, Qiu X. Does syntax matter? A strong baseline for aspect-based sentiment analysis with RoBERTa. Presented at: Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; June 6-11, 2021, 2021; Online. URL: https://aclanthology.org/2021.naacl-main.146 [doi: 10.18653/v1/2021.naacl-main.146]
59. You L, Han F, Peng J, Jin H, Claramunt C. ASK-RoBERTa: a pretraining model for aspect-based sentiment classification via sentiment knowledge mining. Knowl Based Syst. Oct 11, 2022;253:109511. [doi: 10.1016/j.knosys.2022.109511]
60. Brown TB, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, et al. Language models are few-shot learners. arXiv. Preprint posted online on May 28, 2020. [FREE Full text]
61. Kojima T, Gu SS, Reid M, Matsuo Y, Iwasawa Y. Large language models are zero-shot reasoners. arXiv. Preprint posted online on May 24, 2022. [FREE Full text]
62. Rae JW, Borgeaud S, Cai T, Millican K, Hoffmann J, Song F, et al. Scaling language models: methods, analysis and insights from training gopher. arXiv. Preprint posted online on December 8, 2021. [FREE Full text]
63. Rodríguez-Ibánez M, Casánez-Ventura A, Castejón-Mateos F, Cuenca-Jiménez PM. A review on sentiment analysis from social media platforms. Expert Syst Appl. Aug 2023;223:119862. [doi: 10.1016/j.eswa.2023.119862]
64. Zunic A, Corcoran P, Spasic I. Sentiment analysis in health and well-being: systematic review. JMIR Med Inform. Jan 28, 2020;8(1):e16023. [FREE Full text] [doi: 10.2196/16023] [Medline: 32012057]
65. Bian J, Zhao Y, Salloum RG, Guo Y, Wang M, Prosperi M, et al. Using social media data to understand the impact of promotional information on laypeople's discussions: a case study of lynch syndrome. J Med Internet Res. Dec 13, 2017;19(12):e414. [FREE Full text] [doi: 10.2196/jmir.9266] [Medline: 29237586]
66. Lim S, Tucker CS, Kumara S. An unsupervised machine learning model for discovering latent infectious diseases using social media data. J Biomed Inform. Feb 2017;66:82-94. [FREE Full text] [doi: 10.1016/j.jbi.2016.12.007] [Medline: 28034788]
67. Gohil S, Vuik S, Darzi A. Sentiment analysis of health care tweets: review of the methods used. JMIR Public Health Surveill. Apr 23, 2018;4(2):e43. [FREE Full text] [doi: 10.2196/publichealth.5789] [Medline: 29685871]
68. Shofiya C, Abidi S. Sentiment analysis on COVID-19-related social distancing in Canada using Twitter data. Int J Environ Res Public Health. Jun 03, 2021;18(11):5993. [FREE Full text] [doi: 10.3390/ijerph18115993] [Medline: 34204907]
69. Es-Sabery F, Es-Sabery K, Qadir J, Sainz-De-Abajo B, Hair A, Garcia-Zapirain B, et al. A MapReduce opinion mining for COVID-19-related tweets classification using enhanced ID3 decision tree classifier. IEEE Access. 2021;9:58706-58739. [doi: 10.1109/access.2021.3073215]
70. Basiri ME, Nemati S, Abdar M, Asadi S, Acharrya UR. A novel fusion-based deep learning model for sentiment analysis of COVID-19 tweets. Knowl Based Syst. Sep 27, 2021;228:107242. [FREE Full text] [doi: 10.1016/j.knosys.2021.107242] [Medline: 36570870]
71. Yan C, Law M, Nguyen S, Cheung J, Kong J. Comparing public sentiment toward COVID-19 vaccines across Canadian cities: analysis of comments on Reddit. J Med Internet Res. Sep 24, 2021;23(9):e32685. [FREE Full text] [doi: 10.2196/32685] [Medline: 34519654]
72. Chandra R, Krishna A. COVID-19 sentiment analysis via deep learning during the rise of novel cases. PLoS One. Aug 19, 2021;16(8):e0255615. [FREE Full text] [doi: 10.1371/journal.pone.0255615] [Medline: 34411112]
73. Zang S, Zhang X, Xing Y, Chen J, Lin L, Hou Z. Applications of social media and digital technologies in COVID-19 vaccination: scoping review. J Med Internet Res. Feb 10, 2023;25:e40057. [FREE Full text] [doi: 10.2196/40057] [Medline: 36649235]
74. Loria S. textblob Documentation. Release 0.15, 2. TextBlob. 2018. URL: https://textblob.readthedocs.io/en/dev/changelog.html [accessed 2024-01-09]
75. Yu S, Eisenman D, Han Z. Temporal dynamics of public emotions during the COVID-19 pandemic at the epicenter of the outbreak: sentiment analysis of Weibo posts from Wuhan. J Med Internet Res. Mar 18, 2021;23(3):e27078. [FREE Full text] [doi: 10.2196/27078] [Medline: 33661755]

76. Melton CA, Olusanya OA, Ammar N, Shaban-Nejad A. Public sentiment analysis and topic modeling regarding COVID-19 vaccines on the Reddit social media platform: a call to action for strengthening vaccine confidence. J Infect Public Health. Oct 2021;14(10):1505-1512. [FREE Full text] [doi: 10.1016/j.jiph.2021.08.010] [Medline: 34426095]
77. Huangfu L, Mo Y, Zhang P, Zeng DD, He S. COVID-19 vaccine tweets after vaccine rollout: sentiment-based topic modeling. J Med Internet Res. Feb 08, 2022;24(2):e31726. [FREE Full text] [doi: 10.2196/31726] [Medline: 34783665]
78. Boukobza A, Burgun A, Roudier B, Tsopra R. Deep neural networks for simultaneously capturing public topics and sentiments during a pandemic: application on a COVID-19 tweet data set. JMIR Med Inform. May 25, 2022;10(5):e34306. [FREE Full text] [doi: 10.2196/34306] [Medline: 35533390]
79. Ali SH, Foreman J, Capasso A, Jones AM, Tozan Y, DiClemente RJ. Social media as a recruitment platform for a nationwide online survey of COVID-19 knowledge, beliefs, and practices in the United States: methodology and feasibility analysis. BMC Med Res Methodol. May 13, 2020;20(1):116. [FREE Full text] [doi: 10.1186/s12874-020-01011-0] [Medline: 32404050]
80. Geldsetzer P. Use of rapid online surveys to assess people's perceptions during infectious disease outbreaks: a cross-sectional survey on COVID-19. J Med Internet Res. Apr 2, 2020;22(4):e18790. [FREE Full text] [doi: 10.2196/18790] [Medline: 32240094]
81. Hlatshwako TG, Shah SJ, Kosana P, Adebayo E, Hendriks J, Larsson EC, et al. Online health survey research during COVID-19. Lancet Digit Health. Feb 2021;3(2):e76-e77. [doi: 10.1016/s2589-7500(21)00002-9]
82. Shuja J, Alanazi E, Alasmary W, Alashaikh A. COVID-19 open source data sets: a comprehensive survey. Appl Intell (Dordr). 2021;51(3):1296-1325. [FREE Full text] [doi: 10.1007/s10489-020-01862-6] [Medline: 34764552]
83. Ben-Ezra M, Hamama-Raz Y, Goodwin R, Leshem E, Levin Y. Association between mental health trajectories and somatic symptoms following a second lockdown in Israel: a longitudinal study. BMJ Open. Sep 02, 2021;11(9):e050480. [FREE Full text] [doi: 10.1136/bmjopen-2021-050480] [Medline: 34475179]
84. de Koning R, Egiz A, Kotecha J, Ciuculete AC, Ooi SZ, Bankole ND, et al. Survey fatigue during the COVID-19 pandemic: an analysis of neurosurgery survey response rates. Front Surg. Aug 12, 2021;8:690680. [FREE Full text] [doi: 10.3389/fsurg.2021.690680] [Medline: 34458314]
85. Wang Z, Xie Q, Ding Z, Feng Y, Xia R. Is ChatGPT a good sentiment analyzer? a preliminary study. arXiv. Preprint posted online on April 10, 2023. [FREE Full text]
86. Chung JY, Gibbons A, Atlas L, Ballard E, Ernst M, Japee S, et al. COVID-19 and mental health: predicted mental health status is associated with clinical symptoms and pandemic-related psychological and behavioral responses. medRxiv. Preprint posted online on October 14, 2021 (forthcoming) [FREE Full text] [doi: 10.1101/2021.10.12.21264902] [Medline: 34671781]
87. Weger R, Lossio-Ventura JA, Rose-McCandlish M, Shaw JS, Sinclair S, Pereira F, et al. Trends in language use during the COVID-19 pandemic and relationship between language use and mental health: text analysis based on free responses from a longitudinal study. JMIR Ment Health. Mar 01, 2023;10:e40899. [FREE Full text] [doi: 10.2196/40899] [Medline: 36525362]
88. Nelson LM, Simard JF, Oluyomi A, Nava V, Rosas LG, Bondy M, et al. US public concerns about the COVID-19 pandemic from results of a survey given via social media. JAMA Intern Med. Jul 01, 2020;180(7):1020-1022. [FREE Full text] [doi: 10.1001/jamainternmed.2020.1369] [Medline: 32259192]
89. Lossio-Ventura JA, Weger R, Lee A, Guinee E, Chung J, Atlas L, et al. Sentiment analysis test dataset created from two COVID-19 surveys: National Institutes of Health (NIH) and Stanford University. FigShare. Nov 23, 2023. URL: http://tinyurl.com/2s39yutm [accessed 2024-01-09]
90. Nakov P, Ritter A, Rosenthal S, Sebastiani F, Stoyanov V. SemEval-2016 task 4: sentiment analysis in Twitter. Presented at: 10th International Workshop on Semantic Evaluation (SemEval-2016); June 16-17, 2016, 2016; San Diego, California. URL: https://aclanthology.org/S16-1001/ [doi: 10.18653/v1/s16-1001]
91. Rosenthal S, Farra N, Nakov P. SemEval-2017 task 4: sentiment analysis in Twitter. Presented at: 11th International Workshop on Semantic Evaluation (SemEval-2017); August 3-4, 2017, 2017; Vancouver, Canada. URL: https://aclanthology.org/S17-2088 [doi: 10.18653/v1/s17-2088]
92. Pennebaker JW, Boyd RL, Jordan K, Blackburn K. The development and psychometric properties of LIWC2015. University of Texas at Austin. 2015. URL: https://repositories.lib.utexas.edu/server/api/core/bitstreams/b0d26dcf-2391-4701-88d0-3cf50ebee697/content [accessed 2024-01-09]
93. Introducing LIWC-22. Linguistic Inquiry and Word Count. URL: https://www.liwc.app/ [accessed 2022-03-15]
94. Thelwall M, Buckley K, Paltoglou G, Cai D, Kappas A. Sentiment strength detection in short informal text. J Am Soc Inf Sci Technol. Dec 15, 2010;61(12):2544-2558. [doi: 10.1002/asi.21416]
95. Thelwall M, Buckley K, Paltoglou G. Sentiment strength detection for the social web. J Am Soc Inf Sci Technol. Oct 13, 2011;63(1):163-173. [doi: 10.1002/asi.21662]
96. SentiStrength- sentiment strength detection in short texts. SentiStrength. URL: http://sentistrength.wlv.ac.uk/ [accessed 2022-03-15]
97. VADER-sentiment-analysis. GitHub. URL: https://github.com/cjhutto/vaderSentiment [accessed 2022-03-15]
98. Stanza – a python NLP package for many human languages. Stanza. URL: https://stanfordnlp.github.io/stanza/ [accessed 2022-03-15]

99. Mikolov T, Sutskever I, Chen K, Corrado G, Dean J. Distributed representations of words and phrases and their compositionality. arXiv. Preprint posted online on October 16, 2013. [FREE Full text]
100. Barbieri F, Camacho-Collados J, Espinosa-Anke L, Neves L. TweetEval: unified benchmark and comparative evaluation for tweet classification. arXiv. Preprint posted online on October 23, 2020. [FREE Full text] [doi: 10.18653/v1/2020.findings-emnlp.148]
101. Pérez JM, Rajngewerc M, Giudici JC, Furman DA, Luque F, Alemany LA, et al. pysentimiento: a python toolkit for opinion mining and social NLP tasks. arXiv. Preprint posted online on June 17, 2021. [FREE Full text]
102. pysentimiento. GitHub. URL: https://github.com/pysentimiento/pysentimiento [accessed 2022-03-15]
103. Nguyen DQ, Vu T, Nguyen AT. BERTweet: a pre-trained language model for English tweets. arXiv. Preprint posted online on May 20, 2020. [FREE Full text] [doi: 10.18653/v1/2020.emnlp-demos.2]
104. nlptown / bert-base-multilingual-uncased-sentiment. Hugging Face. URL: https://huggingface.co/nlptown/bert-base-multilingual-uncased-sentiment [accessed 2022-03-15]
105. Zhang S, Roller S, Goyal N, Artetxe M, Chen M, Chen S, et al. OPT: open pre-trained transformer language models. arXiv. Preprint posted online on May 2, 2022. [FREE Full text]
106. Benavoli A, Corani G, Dems?ar J, Zaffalon M. Time for a change: a tutorial for comparing multiple classifiers through Bayesian analysis. arXiv. Preprint posted online on June 14, 2016.
107. Demšar J. Statistical comparisons of classifiers over multiple data sets. J Mach Learn Res. 2006;7:1-30. [FREE Full text]
108. Ziems C, Held W, Shaikh O, Chen J, Zhang Z, Yang D. Can large language models transform computational social science? arXiv. Preprint posted online on April 12, 2023 [FREE Full text] [doi: 10.1162/coli_a_00502]
109. Gilardi F, Alizadeh M, Kubli M. ChatGPT outperforms crowd workers for text-annotation tasks. Proc Natl Acad Sci USA. Jul 25, 2023;120(30):e2305016120. [FREE Full text] [doi: 10.1073/pnas.2305016120] [Medline: 37463210]
110. juanlossio / sentiment_analysis. GitHub. URL: https://github.com/juanlossio/sentiment_analysis [accessed 2024-01-12]

Abbreviations

BERT: Bidirectional Encoder Representations from Transformers
DNN: deep neural network
ELMo: Embeddings from Language Models
GPT: Generative Pre-trained Transformers
LIWC: Linguistic Inquiry and Word Count
LIWC2015: Linguistic Inquiry and Word Count 2015
LLM: large language model
NIH: National Institutes of Health
NLP: natural language processing
OPT: Open Pre-Trained Transformers
PaLM: Pathways Language Model
PHI: protected health information
RoBERTa: Robustly optimized Bidirectional Encoder Representations from Transformers approach
VADER: Valence Aware Dictionary and Sentiment Reasoner

Edited by J Torous; submitted 21.06.23; peer-reviewed by L He, R Gore; comments to author 02.09.23; revised version received 16.11.23; accepted 17.11.23; published 25.01.24

Please cite as:
Lossio-Ventura JA, Weger R, Lee AY, Guinee EP, Chung J, Atlas L, Linos E, Pereira F
A Comparison of ChatGPT and Fine-Tuned Open Pre-Trained Transformers (OPT) Against Widely Used Sentiment Analysis Tools: Sentiment Analysis of COVID-19 Survey Data
JMIR Ment Health 2024;11:e50150
URL: https://mental.jmir.org/2024/1/e50150
doi: 10.2196/50150
PMID: