medRxiv preprint doi: https://doi.org/10.1101/2024.02.06.24302383; this version posted February 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

Use of large language models as a scalable approach to understanding public health discourse

Short title:

Large language models for understanding public health discourse

Laura Espinosa[1*], *Marcel Salathé*[1]

[1] Digital Epidemiology Lab, School of Life Sciences, School of Computer and Communication Sciences, EPFL, Switzerland

* Corresponding author. Email: laura.espinosamontalban@epfl.ch

Abstract

Online public health discourse is becoming more and more important in shaping public health dynamics. Large Language Models (LLMs) offer a scalable solution for analysing the vast amounts of unstructured text found on online platforms. Here, we explore the effectiveness of Large Language Models (LLMs), including GPT models and open-source alternatives, for extracting public stances towards vaccination from social media posts. Using an expert-annotated dataset of social media posts related to vaccination, we applied various LLMs and a rule-based sentiment analysis tool to classify the stance towards vaccination. We assessed the accuracy of these methods through comparisons with expert annotations and annotations obtained through crowdsourcing. Our results

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2024.02.06.24302383; this version posted February 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

demonstrate that few-shot prompting of best-in-class LLMs are the best performing methods, and that all alternatives have significant risks of substantial misclassification. The study highlights the potential of LLMs as a scalable tool for public health professionals to quickly gauge public opinion on health policies and interventions, offering an efficient alternative to traditional data analysis methods. With the continuous advancement in LLM development, the integration of these models into public health surveillance systems could substantially improve our ability to monitor and respond to changing public health attitudes.

Authors summary

We examined how Large Language Models (LLMs), including GPT models and open-source versions, can analyse online discussions about vaccination from social media. Using a dataset with expert-checked posts, we tested various LLMs and a sentiment analysis tool to identify public stance towards vaccination. Our findings suggest that using LLMs, and prompting them with labelled examples, is the most effective approach. The results show that LLMs are a valuable resource for public health experts to quickly understand the dynamics of public attitudes towards health policies and interventions, providing a faster and efficient option compared to traditional methods. As LLMs continue to improve, incorporating these models into digital public health monitoring could greatly improve how we observe and react to dynamics in public health discussions.

Introduction

Public health is “the art and science of preventing disease, prolonging life and promoting health through the organized efforts of society” [1]. The overall aim of public health is to sustainably promote the health and wellbeing of the entire population, strengthening public health services and reducing inequalities [2], including surveillance, prevention and control of infectious diseases, promoting healthy lifestyles, researching disease, and preventing injury. Designated populations can vary, depending on context, from small neighbourhoods to global regions [3].

One of the greatest achievements of public health in the 20th century is vaccination against infectious diseases, which saves lives, improves individual and population health,

medRxiv preprint doi: https://doi.org/10.1101/2024.02.06.24302383; this version posted February 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

and reduces health care costs by billions of dollars annually. Despite the success of vaccines, uptake of vaccination is still inadequate, often delayed, and unstable [4]. Understanding the factors that affect vaccine acceptance or refusal is key to increasing vaccination uptake. In October 2014, World Health Organization (WHO) Strategic Advisory Group of Experts on Immunization (SAGE) published a report of the SAGE Working Group on Vaccine Hesitancy outlying several factors that can influence vaccine acceptance or refusal grouped in four categories: populations, subgroups, communities, and individuals [5].

Social listening is "an active process of attending to, observing, interpreting, and responding to a variety of stimuli through mediated, electronic, and social channels" [6]. It helps public health experts to better engage with the population [7]. Social media has proven to be useful in public health, for example, for understanding the public perception of an epidemic [8,9]. In the context of COVID-19 pandemic, a scoping review analysed 81 studies on social media from 2020; from these and using a modified version of the framework Social Media and Public Health Epidemic and Response (SPHERE) [10], 48 studies were categorised in the public attitudes theme [11]. A systematic review performed in 2019 on existing opinion mining approaches using data from social media platforms found that 94% of the 461 included studies use data only from one social media platform and the vast majority of the studies used data from X (formerly known as Twitter), followed by Sina Weibo and Facebook [12]. Furthermore, in another systematic review of techniques for stance detection published in 2023 [13], social media was the data source of 72% of the 96 included articles in the study.

Several studies have used natural language processing algorithms to extract the general sentiment of a text [14], and machine learning approaches with labelled data in order to infer public health opinion from social media data. In addition, there are open-source tools available to support such processes, such as Crowdbreaks [15] and epitweetr [16] which use machine learning methods and labelled data to monitor X for extraction of stance towards public health interventions and early detection of public health threats. However, large annotated datasets are required to have enough representation of the specific tasks for ensuring proper learning and to have enough domain-specific examples when applying machine or deep learning techniques to extract or infer the stance from a text, especially in unstructured data such as social media [17].

medRxiv preprint doi: https://doi.org/10.1101/2024.02.06.24302383; this version posted February 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

Furthermore, having non-experts annotators will require a higher number of annotators to equal the results from the expert annotators, and increase costs [18]. These aspects may be underestimated or not properly considered when conceptualising such studies.

In recent years, the emergence of Large Language Models (LLMs) have revolutionised the field of natural language (NLP) processing. The automated identification of a text author's position towards a specific topic (stance detection) is key in public health and other fields. LLMs, thanks to their vast training data and enormous model size with often billions of parameters, have demonstrated remarkable proficiency in this domain. A study from Brown et al., 2020, illustrated the effectiveness of models like GPT-3 in accurately determining stance from text, outperforming NLP models in both precision and recall [19]. Furthermore, the adaptability of LLMs to diverse datasets and their capacity to understand nuanced expressions have made them invaluable tools for researchers and practitioners alike. For instance, Devlin et al., 2018, highlighted the versatility of the Bidirectional Encoder Representations from Transformers in handling context-dependent stance detection tasks, setting a precedent for the more advanced LLMs [20].

More recently, LLMs have entered the public debate with the launch of ChatGPT, a chatbot originally based on the GPT-3.5 model, and more recently the GPT-4 model. These LLMs by OpenAI have triggered unprecedented public discourse around artificial intelligence across all domains. Shortly after the launch of ChatGPT, the first open LLMs were released, notably the Llama models by Meta (e.g. Llama 2 [CITE]). During the year of 2023, new incumbents such as Mistral AI released a number of highly performant open LLMs as a free alternative to commercially available models such as the GPT models. Early results have shown the incredible power in text understanding across domains that these models provided out of the box, pre-trained. This opened up the opportunity for the use of such models for large-scale social media analysis in many domains, including public health, without the resource-intensive steps of data annotation and model training for every individual use case.

The objective of this study was to assess how these new methods would perform for extracting the public's stance towards vaccination using social media posts. The theme of vaccination was chosen for its relevance to public health, as well for its importance in the public discourse and the broad range of views therein. Further, our focus on stance,

medRxiv preprint doi: https://doi.org/10.1101/2024.02.06.24302383; this version posted February 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

rather than just sentiment, is based on the notion that stance is a better measure of vaccination intention. However, stance can be more difficult to assess, because stance can be very different from sentiment (for example, the sentence "I am angry that I haven't been vaccinated yet" expresses a negative sentiment but a positive stance towards vaccination). Finally, our choice of methods was based on practical utility and methods broadly used (or usable in the future) in various domains: annotation by experts, crowds, lexicon- and rule-based sentiment analysis tools, and large language models (LLMs).

Materials and methods

Data collection

Crowdbreaks was a platform that collected and filtered Twitter data (which was renamed X). Crowdbreaks used crowdsourcing annotation and trained machine learning classifiers to determine a tweet's stance towards a specific topic (e.g., vaccination) [15].

We sampled 1,000 random English-language tweets on vaccination from tweets that Crowdbreaks collected between 2nd of December, 2019 and 11th of March, 2022. We used the following keywords: vaccine, vaccination, vaxxer, vaxxed, vaccinated, vaccinating, vaccine, overvaccinate, undervaccinate, and unvaccinated.

We used six methods to classify the user's stance towards vaccination from those 1,000 tweets: experts, crowdsourcing, and the four large language models GPT (Generative Pre-training Transformer, versions 3.5 and 4) [21,22], Mistral 7B Dense Transformer (Mistral) [23] and Mixtral 8x7B (Mixtral) [24]. In addition, we used a lexicon and rule-based feeling analysis instrument named Valence Aware Dictionary and sEntiment Reasoner (VADER) [25] to classify the general sentiment from the same 1,000 tweets as an automated, non-LLM baseline.

Classifying the vaccine stance using experts

Participants

We asked a convenience sample of four experts from the EPFL Digital Epidemiology Lab to identify the general sentiment and stance towards vaccination of Tweets: two public health experts (co-authors of this publication), one expert annotator, and one expert in

medRxiv preprint doi: https://doi.org/10.1101/2024.02.06.24302383; this version posted February 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

administration. These experts were selected based on their knowledge in the field (public health), on their knowledge in the technique (annotation) and as a comparison to the crowdsourcing (expert in administration) with whom we could discuss in case of disagreements.

Data collection

Each of the four experts individually classified the tweets according to the general sentiment and the stance of the user towards vaccination based solely on the text of the same list of tweets.

Each expert classified the tweets on a tab of a Google sheet (one tab for each expert) containing three variables: tweet text, general sentiment of the tweet, and stance towards vaccination of the tweet. The file was saved in a restricted folder in which only the selected experts could access it.

Each expert classified independently the tweets as either negative, neutral, or positive regarding the general sentiment and stance towards vaccination.

Data analysis

We used R 4.1.3 and RStudio version 2023.06.1 for analysing the individual classifications of all tweets [26]. A new dataset was created with the following variables: tweet text, classification of the general sentiment and stance toward vaccination by each expert, agreement (binary variable with 1, if more than three experts labelled equally the tweet for each category, or 0 if it was otherwise), percentage of each class (negative, neutral, and positive) per class (general sentiment and stance towards vaccination), and agreed class (in case of disagreement, we selected the class “neutral”).

Classifying the vaccine stance using crowdsourcing

Participants

We used crowd annotators from the platform Amazon Mechanical Turk (Mturk) [27] to classify the same list of tweets according to the stance towards vaccination. Only crowd annotators with at least 5,000 previous approved tasks and at least 98% of approval rate in the platform had access to the task. In addition, black-listed crowd annotators on the platform were excluded.

medRxiv preprint doi: https://doi.org/10.1101/2024.02.06.24302383; this version posted February 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

Data collection

We published the job in Amazon Mechanical Turk (Mturk) with several information for the crowd annotators. We included the description of the task, expected time to complete the task, a question to classify the tweet based on the "attitude of the author regarding vaccines" (i.e. stance), two examples for each of the classes (positive, neutral and negative) and the indication to classify a tweet as "neutral" if the tweet was news, factual, objective or generally ambiguous.

Each crowd annotator could decide how many and which tweets to classify. Each crowd annotator could classify a maximum of 1,000 tweets, had to wait at least 2 seconds before providing the class to avoid automatic selection by a machine, and was rewarded with 0.04 USD per tweet. At the end, each tweet was classified by at least three crowd annotators.

Data analysis

For each tweet, we had the following variables: identifier of the specific task in Amazon Mturk, tweet identifier, date of the tweet, crowd annotator identifier, tweet text, and total duration and date of the classification per tweet.

Classifying the vaccine stance using LLMs

Software

We used Jupyter Notebook 6.5.2 and Python 3.9.12 to classify the same tweets according to the stance towards vaccination using GPT versions 3.5 (GPT-3.5) and 4 (GPT-4), and Ollama 0.1.17 and Python 3.11.6 to classify the same tweets according to the stance towards vaccination using Mistral and Mixtral. The specific scripts are available in the repository [26]. These LLMs were selected based on its popularity of use (GPT-3.5), expected good performance (GPT-4), and usability with low computing power needed and no cost (Mistral) and expected good performance (Mixtral).

Prompt engineering

We used different prompts to classify the tweets based on instructions and examples provided [S1 Prompts used for large language models]. Prompts 1 and 2 were zero-shot prompts, i.e. without labelled examples, and thus the simplest instructions provided to these LLM asking for the sentiment regarding vaccination. The remaining prompts 3 to

medRxiv preprint doi: https://doi.org/10.1101/2024.02.06.24302383; this version posted February 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

8 used few-shot classification, i.e., few labelled examples for each class were provided for the model to extract the features for each class and use these learned features to predict the class of new tweets [19,28,29]. Prompt 3 contained the same examples provided to crowd annotators (i.e., two examples per class). Prompts 4 to 7 were built based on the previous prompts using a random sample of 100 tweets. Prompts were tested in this sample and the wrongly classified tweets were used as examples for the following prompt. For example, if a tweet considered by experts to have a negative class was classified otherwise, this tweet was included in the next prompt as an example for the negative class. Furthermore, examples of tweets misclassified as neutral were added as exceptions for this class. All prompts except from prompt 1 requested the LLM model to provide an explanation on the classification.

The same parameters were used for all prompts within each LLM. For GPT-3.5, we used the following parameters: text-davinci-003 as model, 0.8 as temperature, and 0 as frequency and presence penalty. For GPT-4, we used the following parameters: gpt-4 as model, "user" as role, 0.8 as temperature, and 0 as frequency and presence penalty. For Mistral and Mixtral, we used 0.8 as temperature.

Classifying the general sentiment using natural language processing algorithms

Software

We used Jupyter Notebook 6.5.2 and Python 3.9.12 to classify the same tweets according to the general sentiment expressed in the tweet using VADER.

Data collection

VADER assigned to each tweet a probability for having a negative, neutral and positive sentiment in the text and an overall score between -1 (most negative) and +1 (most positive).

Data analysis

We used R 4.1.3 and RStudio version 2023.06.1 for classifying the tweets into negative, positive or neutral according to the overall score provided by VADER. Tweets with scores between -0.5 and 0.5 were classified as neutral, more than 0.5 as positive and less than -0.5 as negative. analysing the individual classifications of all tweets

medRxiv preprint doi: https://doi.org/10.1101/2024.02.06.24302383; this version posted February 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

Comparison of the classification methods

Statistical analysis

We used R 4.1.3 and RStudio version 2023.06.1 to compare the classes for the same list of tweets provided by experts, crowdsourcing, LLMs and VADER. Experts were considered the gold standard for the comparison.

We did two comparisons based on the level of agreement by the experts: partial and full agreement in stance towards vaccination. Tweets with at least three out of four experts assigning the same class were considered to have a partial agreement.

We calculated a confusion matrix with three classes (negative, neutral, and positive) for each pair of comparisons: experts and crowdsourcing, experts and GPT, experts and Mistral, experts and Mixtral, and experts and VADER. Moreover, a confusion matrix with three classes was also calculated for the comparison of experts and selecting the majority class (i.e., neutral) for all tweets which was considered the baseline. We calculated the overall accuracy with a 95% confidence interval and a one-side test to see if the accuracy was better than the fourth confusion matrix considering that the system would always select that majority class in the classification process.

Furthermore, for each class we calculated the sensitivity, specificity, and F1 score. For this, we used the experts classes as the gold standard and the crowdsourcing and LLM as the comparison method separately. Moreover, the baseline classifying all tweets as neutral was also applied here.

Ethics

This work has been approved by the EPFL Human Research Ethics Committee (HREC No. 012.2018 / 05.03.2018).

Crowdbreaks anonymises all tweets' text changing the mention or reference to usernames to "@username" and the URLs to "url", and complies with the X terms of service.

We have implemented mitigation measures to ensure an ethical use of the Amazon Mturk platform and avoid any abuse of the system. We provided an overestimation of the time required for the task, provide a communication channel via email in case crowd annotators wanted to contact us, had a prompt communication in case of being

medRxiv preprint doi: https://doi.org/10.1101/2024.02.06.24302383; this version posted February 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

contacted, accepting all jobs in maximum one day, providing comprehensive and clear description of the research context and instructions and job description, providing a monetary compensation substantially higher than the US federal government minimum wage in which the platform is based. In addition, quality criteria for selecting the crowd annotators was established to minimise the risk of having inappropriate users or users misusing the platform.

Results

Extracting the vaccine stance using experts

Seventy-six percent of the 1,000 tweets had a partial agreement between experts (i.e., three out of four experts agreed on the class per tweet). From these 756 tweets, 74.1%, 19.4%, and 6.5% were classified as neutral, positive, and negative stance towards vaccination, respectively.

Thirty-seven percent of those 756 tweets had a full agreement between experts (i.e., all experts agreed on the class per tweet). From these, 77.3%, 17.6%, and 5.0% were classified as neutral, positive, and negative stance towards vaccination, respectively.

Only the vaccine stance from the tweets with at least partial agreement among EPFL experts was also extracted using crowdsourcing and LLMs.

Extracting the vaccine stance using crowdsourcing

A total of 467 crowd annotators classified the tweets with the majority of tweets (64.5%) being classified by three different crowd annotators, followed by 14.7% and 7.4% of the tweets being classified by 10 and four crowd annotators, respectively.

The majority of the 756 tweets had an agreement: 85.6% had a partial agreement (i.e., 75% of the crowd annotators agreed on the stance per tweet) and, from these, 91.2% had full agreement.

The stance towards vaccination of tweets with partial agreement was distributed as follows: 89.5% as neutral and 10.5% as positive, with no tweet classified as negative. The stance towards vaccination of tweets with full agreement were 98.1% classified as neutral and 1.9% classified as positive, with no tweet classified as negative (Fig 1).

medRxiv preprint doi: https://doi.org/10.1101/2024.02.06.24302383; this version posted February 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

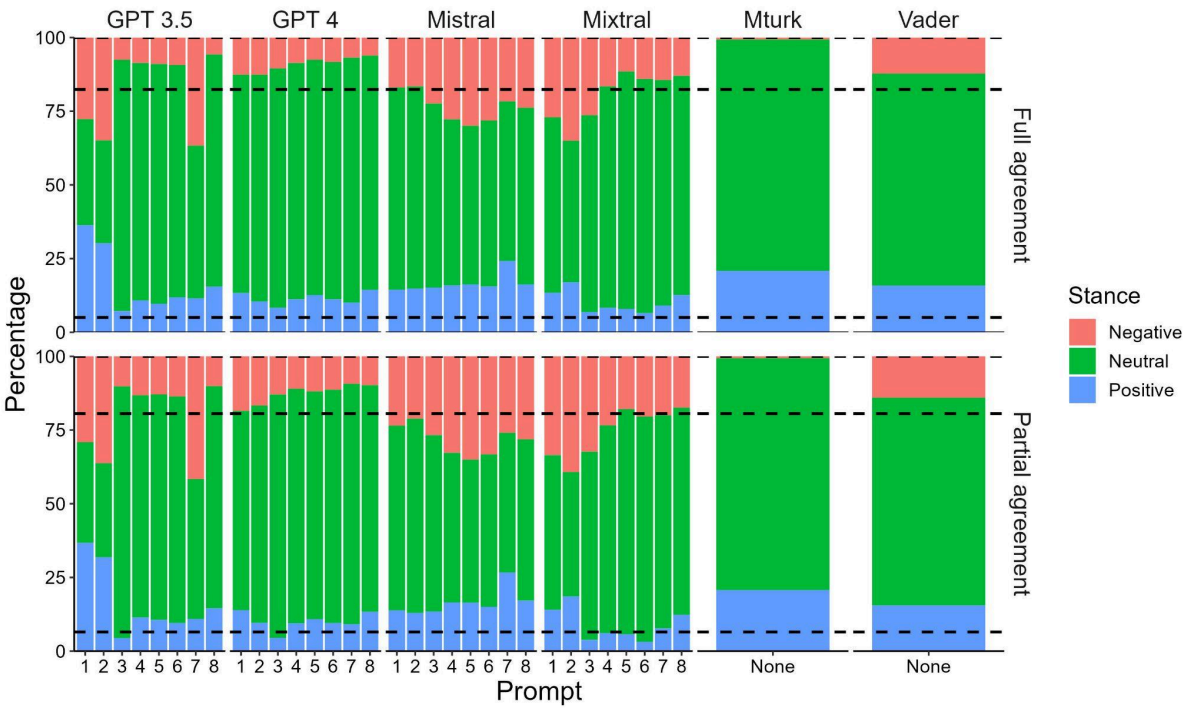

Figure 1. Distribution of the tweets per stance, prompt and large language model (percentage of total number of tweets). The dashed lines represent the stance distribution of positive, neutral and negative tweets according to the experts' full and partial agreement.

Extracting the vaccine stance using LLMs

The majority class was neutral for all prompts, LLMs and tweets with partial or full agreement by experts, except for prompts 1 and 2 of GPT-3.5 for which the majority class was positive and negative, respectively (Fig 1).

Extracting the general sentiment using VADER

The majority class was neutral (Fig 1) with an overall score ranging from -0.97 to 0.99, median of 0 and interquartile range of -0.30 and 0.01.

Comparison of the extraction methods

Overall performance for tweets with partial and full agreement

The overall accuracy for all methods was worse than just selecting the majority class for all tweets, except for prompts 2-8 of GPT-4 in tweets with partial agreement among

medRxiv preprint doi: https://doi.org/10.1101/2024.02.06.24302383; this version posted February 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

experts and for all prompts of GPT-4, prompt 3, 4 and 6 of GPT-3.5 and prompts 5, 7 and 8 of Mixtral in tweets with full agreement among experts (Fig 2).

In the tweets with partial agreement among experts, the accuracy was 0.73 for crowdsourcing and 0.66 for VADER, and ranged between 0.80 and 0.72 for GPT-4, between 0.74 and 0.43 for GPT-3.5, between 0.68 and 0.54 for Mistral and between 0.76 and 0.52 for Mixtral. In the tweets with full agreement among experts, the accuracy had similar ranges with a higher upper value for GPT version 4 (0.91), GPT-3.5 (0.83) and Mixtral (0.84), and higher values for Mistral with a range between 0.79 and 0.62 and crowdsourcing with a value of 0.81.

All prompts of GPT-4 except for prompt 1, half of the Mixtral prompts (4, 5, 7 and 8) and prompt 3 of GPT-3.5 performed better than crowd annotators when comparing the results with the experts' classification in tweets with partial agreement among experts. In addition to those, prompt 1 of GPT-4 and prompts 4, 5 and 6 of GPT-3.5 performed better than crowd annotators in tweets with full agreement among experts.

Some of the prompts of GPT-3.5, Mistral and Mixtral performed worse than extracting the general sentiment of the tweet using VADER.

Overall, Mistral's performance was lower than that of crowd annotators. Interestingly, prompts 1 and 2 had the best accuracy with Mistral in tweets with partial and full agreement among experts, differently to the other models.

medRxiv preprint doi: https://doi.org/10.1101/2024.02.06.24302383; this version posted February 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

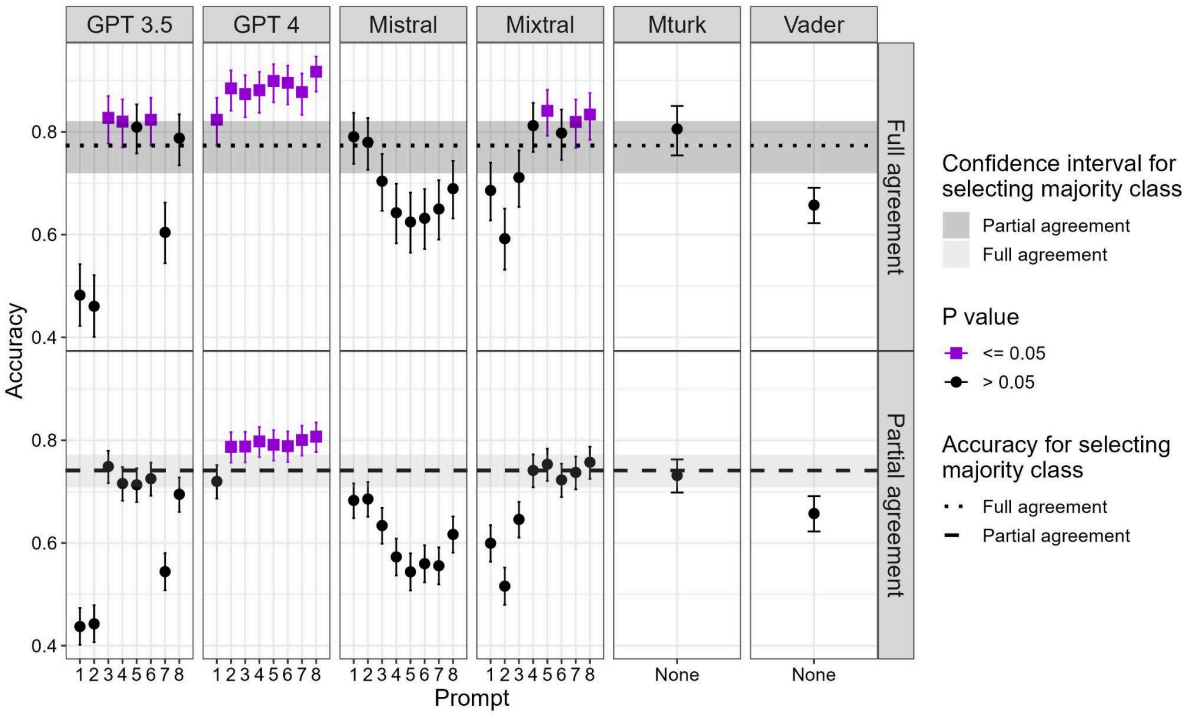

Figure 2. Accuracy and 95% confidence interval per prompt, model and experts' agreement.

Performance per class for partial and full agreement

The performance metrics per class and method showed high variability according to the experts' agreement on stance towards vaccination with an overall highest performance for neutral tweets based on the F1 score and sensitivity (Fig 3).

medRxiv preprint doi: https://doi.org/10.1101/2024.02.06.24302383; this version posted February 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

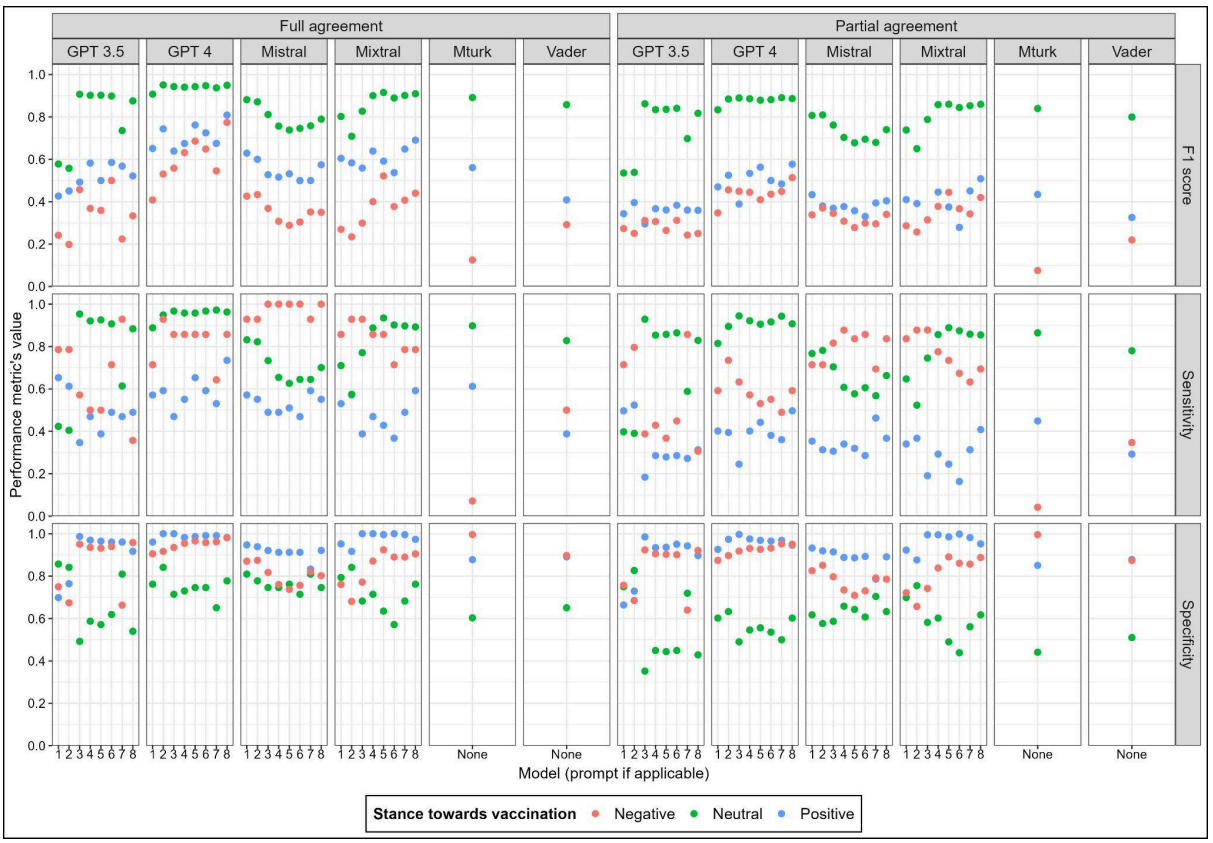

Figure 3. F1 score, sensitivity and specificity per model, prompt (if applicable), stance and experts' agreement.

For the tweets with partial agreement, the performance of crowd annotators was the lowest for the minority class (negative stance towards vaccination) having the lowest F1 score and sensitivity (0.08); however, crowd annotators had the highest specificity (0.99). Furthermore, for the minority class, all prompts except for prompt 4 of GPT-4 and prompts 4, 5 and 8 of Mixtral were the top 10 best performing methods based on the F1 score; and most of the prompts of Mistral (3, 4, 5, 6, and 8), prompts 1, 2 and 3 of Mixtral, and prompts 2 and 7 of GPT-3.5 had the highest sensitivity. For the majority class (neutral stance towards vaccination), the performance was variable according to the method used with all prompts except for prompt 1 of GPT-4, prompts 5 and 8 of Mixtral, and prompt 3 of GPT-3.5 with the highest F1 score; and prompts 1, 2, and 7 of GPT-3.5, prompts 1 and 2 of Mixtral, prompts 4, 5 and 7 of Mistral with the highest specificity. For the positive class, all prompts except prompt 3 of GPT-4 had the highest F1 scores followed by prompts 4, 7 and 8 of Mixtral; and prompts 3, 4, 5, 6 and 7 of Mixtral and prompt 3 of GPT-4 and GPT-3.5 had the highest specificity.

medRxiv preprint doi: https://doi.org/10.1101/2024.02.06.24302383; this version posted February 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

When comparing the results of the tweets with full agreement with the results of the tweets with the partial agreement, there were some commonalities and some differences. The differences for the negative class were that prompt 1 of GPT-4 and prompts 4 and 8 of Mixtral were not among the top 10 methods with the highest F1 score which were substituted in that ranking by prompt 4 of GPT-4, and prompts 3 and 6 of GPT-3.5. The differences for the neutral class were that all prompts of GPT-4 were in the top 10 of methods with the highest F1 score; and prompts 4 and 5 of Mistral were not in the top methods with the highest specificity, which were substituted by prompt 2 of GPT-4 and prompt 1 of Mistral. The differences for the positive class were that prompt 4 of Mixtral was not among the top 10 methods with the highest F1 score, which was substituted by prompt 3 of GPT-4; and prompt 3 of GPT-3.5 was not among the models with the highest specificity which was substituted by prompt 2 of GPT-4.

Discussion

Understanding the population's perception of public health interventions, such as vaccination, and current public health events is critical for public health experts and epidemiologists. This understanding helps them adjust risk communication and community engagement activities. Achieving this is challenging and requires a holistic approach. Our study concentrates on using social media to assess users' stances towards vaccination. However, our method could be easily extended to cover other public health interventions and objectives, as well as other platforms or sources of information. With the continuous advancement in LLM technology and better understanding of their performance in different studies and contexts, their usage and integration into public health surveillance systems could greatly improve our ability to monitor and respond to changing public health attitudes.

Our study provides an analysis of attitudes towards vaccination, derived from 1,000 randomly selected English-language tweets collected between 2019 and 2022. This analysis is based on a consensus among four experts regarding the vaccination stance expressed in each tweet. Moreover, we extracted vaccination stances using six alternative methods: crowd annotators, a rule-based sentiment analysis tool, and four large language models (LLMs), specifically focusing on tweets where at least three out of four experts agreed on the tweet's classification (i.e., partial expert agreement). These

medRxiv preprint doi: https://doi.org/10.1101/2024.02.06.24302383; this version posted February 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

six methods were chosen because each offers unique advantages. Crowd annotators are considered the gold standard for scalable human annotation, but are nonetheless slow and costly. Rule-based sentiment analysis, while not incorporating the latest machine learning technologies, is valued for its simplicity, speed, and the ability to automate and run on local computers.

The four LLMs, which all offer automation and speed, were selected because they represent the leading edge in their field. The two GPT models from OpenAI, known for underpinning ChatGPT, are among the most recognized and utilised. GPT-3.5, the foundation of the free version of ChatGPT, and GPT-4, the most advanced model publicly available via the OpenAI platform and the basis for ChatGPT Plus, are considered state-of-the-art in LLM technology as of this writing. Access to both models is restricted to OpenAI's platform. Mistral and Mixtral, from Mistral AI, are considered state-of-thep-art for open models. The Mistral 7B model, smaller with 7 billion parameters, can operate on a personal computer, such as a laptop. Mixtral, a more complex “mixture of experts” model, may not run on most personal computers but can be installed on local server environments. This versatility implies that Mistral AI's open models provide privacy benefits, as they allow data to stay on-premises.

We included a lexicon and rule-based sentiment analysis tool, specifically designed to extract only the general sentiment of a text, as a baseline for comparison with the more advanced LLMs. Our findings indicated that this method achieved higher accuracy than some LLMs, particularly those utilising zero-shot prompting. This suggests that inadequate prompt engineering can result in poorer outcomes than simply determining the general sentiment with a straightforward tool like VADER. Similar to crowdsourcing, VADER demonstrated high specificity but low sensitivity for minority classes, indicating that while these methods are less effective in identifying minority classes, they benefit from a reduced rate of false positive detections.

We observed a notable class imbalance among the three identified stances or classes (negative, neutral and positive), similar to what has been found in previous studies [30,31]. The majority were neutral, followed by positive, with negative being the least common stance. A high proportion of tweets (76% of the 1,000 annotated tweets) achieved a partial agreement among experts, but full agreement was limited to 37% of

medRxiv preprint doi: https://doi.org/10.1101/2024.02.06.24302383; this version posted February 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

these. This pattern was consistent across all the other methods, with the exception of one GPT-3.5 prompt showing a more balanced class distribution.

In the analysis of the tweets classified using crowdsourcing, 86% of the tweets showed partial agreement among annotators, similar to the experts pattern. However, a significantly higher 91% of these reached full agreement. This might be due to the homogeneity of Amazon Mturk annotators. There was no partial or full agreement in negative tweets, with all neutrally-classified tweets reaching full agreement. This could be attributed to our methodological choice of not including a category for unclear or irrelevant tweets as in other studies [12,13], and instructing annotators to opt for neutrality in cases of uncertainty.

In the analysis of the tweets classified using LLMs, the predominant prediction was neutral, except in GPT-3.5 zero-shot prompts where a more even class distribution was found, which could be a consequence of the simplicity of these prompts. GPT-4 demonstrated a consistent class distribution across prompts, as did Mistral and Mixtral for few-shot prompts, suggesting less prompt-dependency. This is a notable improvement over GPT-3.5, where simpler prompts showed a significantly different class distribution compared to more complex ones.

All methods showed lower performance on the tweets with partial experts' agreement, indicating a common challenge in resolving ambiguity or unclarity. In general, this ambiguity is caused by having single words with multiple meanings, i.e., polysemy, which is more noticeable in short texts as social media posts in which there is not much information on the context [32]. Other factors, such as humour, sarcasm, or general misunderstandings, indicate that a single objective assessment can be challenging, depending as much on the reader's interpretation as on the content itself. These considerations are important when employing scalable methods and evaluating the potential risks of misclassification.

Generally, the highest performing prompts of GPT-3.5 and GPT-4 surpassed crowdsourcing in the overall accuracy, as supported by previous studies [33–36]. Furthermore, the overall accuracy of the best performing prompts of Mixtral and specific accuracy for the minority class of the best performing prompt of Mistral also outperformed crowdsourcing. The cost benefit and time effectiveness of LLMs over

medRxiv preprint doi: https://doi.org/10.1101/2024.02.06.24302383; this version posted February 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

crowd annotators, despite usage fees, has been previously highlighted [33]. Furthermore, a study has shown that up to 46% of crowd workers resorted to LLMs for task completion [37].

Each method showed varying performance per class. Some prompts of GPT-4 and Mixtral stood out with the highest overall accuracy and statistically significant better performance than just selecting the majority class for all tweets. They also had the highest F1 scores across all classes, indicating a reduced impact from the class imbalance. GPT-3.5 had the greatest variability in performance by class, particularly in the tweets with full agreement, showing higher dependency on the prompt than other LLMs. Crowdsourcing showed its weakest performance in the negative class, being the method with the lowest performance, suggesting a need for more diverse examples in the dataset. These differences in performance per class must be considered to use complementary models depending on their performance and the objective of the annotation as suggested in a previous study [35].

Mixtral, an open-source "mixture of experts" model, demonstrated a performance comparable to GPT-4, a strong result given its accessibility and lower computational requirements. It showed consistent performance across different prompts, with simpler prompts often yielding better results. The accuracy of Mixtral in tweets with full experts' agreement was statistically better than merely selecting the majority class, a trend also observed for many GPT-4 prompts, and some GPT 3.5 prompts. Given the open nature of Mixtral and its ability to be installed in a local computing environment, this makes Mixtral a very interesting model for use in data-sensitive situations.

We did not analyse the explanations provided by LLMs, nor collected these from crowd annotators, which could offer insights into the discrepancies with expert classification. A study by Huang et al., 2023, suggests that GPT-3.5 provides statistically clearer explanations than human-written ones [34], indicating a potential area for future research. Additionally, while our study was limited to English tweets, a previous research indicates that GPT's performance is consistent in non-English languages when prompts remain in English [35], allowing for multilingual studies.

In summary, GPT-4 and Mixtral showed a consistent high performance across prompts, while GPT3.5 and Mistral showed more variability and prompt dependency, showing a

medRxiv preprint doi: https://doi.org/10.1101/2024.02.06.24302383; this version posted February 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

higher relevance of the prompt engineering for these models. In general, LLMs, particularly in the minority negative class, outperformed crowdsourcing, highlighting their efficacy even with limited class-specific data. Overall, the use of best-in-class models like the open Mixtral model or the closed GPT-4 model with few-shot prompting are the best methods. All alternatives have significant risks of substantial misclassification.

Data sharing

The data (list of tweets identifiers and summarised anonymised datasets) and R and Python code used can be found in the online repository at https://github.com/digitalepidemiologylab/llm_crowd_experts_annotation.

Acknowledgements

We would like to thank the crowd and experts for their valuable inputs in the annotation of the tweets. This work was supported by Fondation Botnar and the European Union's Horizon H2020 grant VEO (874735).

References

1. Acheson ED. On the state of the public health [The fourth Duncan lecture]. Public Health. 1988;102: 431–437. doi:10.1016/S0033-3506(88)80080-5
2. World Health Organization. Public health services. Available: https://www.euro.who.int/en/health-topics/Health-systems/public-health-services/public-health-services
3. CDC Foundation. What is Public Health? 2021 [cited 20 Sep 2021]. Available: https://www.cdcfoundation.org/what-public-health
4. Armstrong D. Origins of the Problem of Health-related Behaviours: A Genealogical Study. Soc Stud Sci. 2009;39: 909–926. doi:10.1177/0306312709104258
5. World Health Organisation. Report of the SAGE working group on vaccine hesitancy. [cited 3 Nov 2023]. Available: https://cdn.who.int/media/docs/default-source/immunization/sage/2014/october/sage-working-group-revised-report-vaccine-hesitancy.pdf?sfvrsn=240a7c1c_4
6. Stewart MC, Arnold CL. Defining Social Listening: Recognizing an Emerging Dimension of Listening. International Journal of Listening. 2018;32: 85–100. doi:10.1080/10904018.2017.1330656
7. Pomputius A. Can You Hear Me Now? Social Listening as a Strategy for Understanding User Needs. Med Ref Serv Q. 2019;38: 181–186. doi:10.1080/02763869.2019.1588042
8. Dang T, Nguyen NVT, Pham V. HealthTvizer: Exploring Health Awareness in Twitter Data through Coordinated Multiple Views. 2018 IEEE International Conference on

medRxiv preprint doi: https://doi.org/10.1101/2024.02.06.24302383; this version posted February 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

Big Data (Big Data). 2018. pp. 3647–3655. doi:10.1109/BigData.2018.8622445
9. Kannan R, Govindasamy MA, Soon L-K, Ramakrishnan K. Social Media Analytics for Dengue Monitoring in Malaysia. 2018 8th IEEE International Conference on Control System, Computing and Engineering (ICCSCE). 2018. pp. 105–108. doi:10.1109/ICCSCE.2018.8685028
10. Schillinger D, Chittamuru D, Ramírez AS. From "Infodemics" to Health Promotion: A Novel Framework for the Role of Social Media in Public Health. Am J Public Health. 2020;110: 1393–1396. doi:10.2105/AJPH.2020.305746
11. Tsao S-F, 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. 2021;3: e175–e194. doi:10.1016/S2589-7500(20)30315-0
12. Cortis K, Davis B. Over a decade of social opinion mining: a systematic review. Artif Intell Rev. 2021;54: 4873–4965. doi:10.1007/s10462-021-10030-2
13. Alturayeif N, Luqman H, Ahmed M. A systematic review of machine learning techniques for stance detection and its applications. Neural Comput Appl. 2023;35: 5113–5144. doi:10.1007/s00521-023-08285-7
14. Ribeiro FN, Araújo M, Gonçalves P, André Gonçalves M, Benevenuto F. SentiBench - a benchmark comparison of state-of-the-practice sentiment analysis methods. EPJ Data Sci. 2016;5: 23. doi:10.1140/epjds/s13688-016-0085-1
15. Müller MM, Salathé M. Crowdbreaks: Tracking Health Trends Using Public Social Media Data and Crowdsourcing. Front Public Health. 2019;7: 81. doi:10.3389/fpubh.2019.00081
16. Espinosa L, Wijermans A, Orchard F, Höhle M, Czernichow T, Coletti P, et al. Epitweetr: Early warning of public health threats using Twitter data. Eurosurveillance. 2022;27: 2200177. doi:10.2807/1560-7917.ES.2022.27.39.2200177
17. Wang R, Zhou D, Jiang M, Si J, Yang Y. A Survey on Opinion Mining: From Stance to Product Aspect. IEEE Access. 2019;7: 41101–41124. doi:10.1109/ACCESS.2019.2906754
18. Snow R, O'Connor B, Jurafsky D, Ng A. Cheap and Fast – But is it Good? Evaluating Non-Expert Annotations for Natural Language Tasks. In: Lapata M, Ng HT, editors. Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing. Honolulu, Hawaii: Association for Computational Linguistics; 2008. pp. 254–263. Available: https://aclanthology.org/D08-1027
19. Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, et al. Language Models are Few-Shot Learners. In: Larochelle H, Ranzato M, Hadsell R, Balcan MF, Lin H, editors. Advances in Neural Information Processing Systems. Curran Associates, Inc.; 2020. pp. 1877–1901. Available: https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf
20. Devlin J, Chang M-W, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv; 2019. Available: http://arxiv.org/abs/1810.04805
21. OpenAI. GPT-3.5. 2022. Available: https://openai.com/
22. OpenAI. GPT-4. 2023. Available: https://openai.com/
23. Jiang AQ, Sablayrolles A, Mensch A, Bamford C, Chaplot DS, Casas D de las, et al. Mistral 7B. 2023 [cited 5 Feb 2024]. doi:10.48550/ARXIV.2310.06825
24. Jiang AQ, Sablayrolles A, Roux A, Mensch A, Savary B, Bamford C, et al. Mixtral of Experts. 2024 [cited 5 Feb 2024]. doi:10.48550/ARXIV.2401.04088
25. Hutto C, Gilbert E. VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. ICWSM. 2014;8: 216–225.

medRxiv preprint doi: https://doi.org/10.1101/2024.02.06.24302383; this version posted February 6, 2024. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

doi:10.1609/icwsm.v8i1.14550

26. Espinosa L, Salathé M. Use of large language models as a scalable approach to understanding public health discourse. In: GitHub repository [Internet]. [cited 5 Feb 2024]. Available: https://github.com/digitalepidemiologylab/llm_crowd_experts_annotation
27. Crowston K. Amazon Mechanical Turk: A Research Tool for Organizations and Information Systems Scholars. In: Bhattacherjee A, Fitzgerald B, editors. Shaping the Future of ICT Research Methods and Approaches. Berlin, Heidelberg: Springer; 2012. pp. 210–221. doi:10.1007/978-3-642-35142-6_14
28. Zhang R, Liu Q. Learning with few samples in deep learning for image classification, a mini-review. Front Comput Neurosci. 2023;16: 1075294. doi:10.3389/fncom.2022.1075294
29. Su H, Kasai J, Wu CH, Shi W, Wang T, Xin J, et al. Selective Annotation Makes Language Models Better Few-Shot Learners. arXiv; 2022. Available: http://arxiv.org/abs/2209.01975
30. Kunneman F, Lambooij M, Wong A, Bosch A van den, Mollema L. Monitoring stance towards vaccination in Twitter messages. arXiv; 2019. Available: http://arxiv.org/abs/1909.00338
31. Martínez RY, Blanco G, Lourenço A. Spanish Corpora of tweets about COVID-19 vaccination for automatic stance detection. Information Processing & Management. 2023;60: 103294. doi:10.1016/j.ipm.2023.103294
32. Deng S, Sinha AP, Zhao H. Resolving Ambiguity in Sentiment Classification: The Role of Dependency Features. ACM Trans Manage Inf Syst. 2017;8: 1–13. doi:10.1145/3046684
33. Gilardi F, Alizadeh M, Kubli M. ChatGPT Outperforms Crowd-Workers for Text-Annotation Tasks. Proc Natl Acad Sci USA. 2023;120: e2305016120. doi:10.1073/pnas.2305016120
34. Huang F, Kwak H, An J. Is ChatGPT better than Human Annotators? Potential and Limitations of ChatGPT in Explaining Implicit Hate Speech. Companion Proceedings of the ACM Web Conference 2023. Austin TX USA: ACM; 2023. pp. 294–297. doi:10.1145/3543873.3587368
35. Kuzman T, Mozetič I, Ljubešić N. ChatGPT: Beginning of an End of Manual Linguistic Data Annotation? Use Case of Automatic Genre Identification. arXiv; 2023. Available: http://arxiv.org/abs/2303.03953
36. Törnberg P. ChatGPT-4 Outperforms Experts and Crowd Workers in Annotating Political Twitter Messages with Zero-Shot Learning. arXiv; 2023. Available: http://arxiv.org/abs/2304.06588
37. Veselovsky V, Ribeiro MH, West R. Artificial Intelligence: Crowd Workers Widely Use Large Language Models for Text Production Tasks. arXiv; 2023. doi:10.48550/arXiv.2306.07899