frontiers in Political Science

TYPE Original Research
PUBLISHED 29 July 2026
DOI 10.3389/fpos.2026.1745587

Check for updates

OPEN ACCESS

EDITED BY
Joseph Aistrup,
Auburn University, United States

REVIEWED BY
Vaibhav Khatavkar,
DES Pune University, India
Manfred Stede,
Universität Potsdam, Germany

*CORRESPONDENCE
Mauricio Martins
mauricio.martins@fpce.up.pt
Lukas Mayrhofer
lukas.mayrhofer@univie.ac.at

RECEIVED 13 November 2025
REVISED 08 May 2026
ACCEPTED 06 July 2026
PUBLISHED 29 July 2026

CITATION
Mayrhofer L, Foramitti M, Fassnacht S, Köhler JK, Todorova B, Lamm C and Martins M (2026) GPT outperforms BERT and LIWC for stance and anger detection in German news articles and user comments.
Front. Polit. Sci. 8:1745587.
doi: 10.3389/fpos.2026.1745587

COPYRIGHT
© 2026 Mayrhofer, Foramitti, Fassnacht, Köhler, Todorova, Lamm and Martins. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

GPT outperforms BERT and LIWC for stance and anger detection in German news articles and user comments

Lukas Mayrhofer[1]*, Markus Foramitti[2], Simon Fassnacht[3], Jana K. Köhler[4], Boryana Todorova[1], Claus Lamm[1] and Mauricio Martins[1,5,6]*

[1]Social, Cognitive and Affective Neuroscience Unit, Faculty of Psychology, University of Vienna, Vienna, Austria, [2]Department of Clinical and Health Psychology, Faculty of Psychology, University of Vienna, Vienna, Austria, [3]Max Planck Institute for Biological Cybernetics, Tübingen, Germany, [4]Environmental Psychology, Department of Cognition, Emotion, and Methods in Psychology, Faculty of Psychology, University of Vienna, Vienna, Austria, [5]Universidade Lusófona, CICANT, Lisbon, Portugal, [6]Center for Psychology, Faculty of Psychology and Education Sciences, University of Porto, Porto, Portugal

We live in an era of abundant qualitative language data that allows insights into societal trends and public attitudes, such as social media posts and news articles. Natural language processing (NLP) has become essential for analyzing this data, with recent advances in Large Language Models (LLMs) driving rapid progress. These models show strong potential for psychological research, though their validity in detecting attitudes varies across domains. This study evaluates GPT-3.5 and GPT-4 Turbo for identifying stances and expressions of anger in texts about climate activism, comparing them with traditional methods. The dataset included news articles and user comments from eight German outlets. A subset of 320 articles and 330 comments was manually annotated by three human raters for (1) stance toward Fridays for Future and (2) anger expression. GPT models outperformed BERT (Bidirectional Encoder Representations from Transformers) in stance detection and LIWC (Linguistic Inquiry and Word Count) in anger detection. Moreover, GPT achieved higher agreement with the human majority vote than individual raters agreed with each other. GPT-4 Turbo performed slightly better than GPT-3.5, though differences were minor and domain-specific. Overall, GPT models proved reliable and efficient for annotating nuanced German texts where conventional NLP tools often struggle.

KEYWORDS
BERT, climate activism, LIWC, LLM, NLP, stance detection

Introduction

Language analysis has long been used to research human psychology (Boyd and Schwartz, 2021). The analysis of language data generated in real-world contexts is particularly valuable, as it offers insights into the natural dynamics of individuals and society, free from the constraints or biases inherent in experimental settings. However, the sheer volume of data makes manual analysis impractical, necessitating the use of computational methods (Bittermann and Fischer, 2024). The increasing availability of qualitative language data has driven major innovations in computational language modelling, in turn sparking significant interest in related

research (Chen et al., 2025). For example, language analysis has been successfully used to differentiate between personality types and psychological disorders, to investigate health and voting behaviors, affective states, relationship dynamics, and cultural differences in values and attitudes at societal levels (Mihalcea et al., 2024).

Another valuable application of language analysis is the investigation of opinions and attitudes expressed by individuals. This can be effectively achieved through a computational method called stance detection. A stance is defined as an expression of a person's attitudes, feelings, judgments, or commitments towards a message (Biber and Finegan, 1988). Stance detection is closely related to sentiment analysis, which aims to measure the explicit tonality of a text (positive vs. negative). However, the advantage of stance detection is that it is fine-tuned to recognize subtle expressions of opinions towards a topic and usually classifies texts as either *in favor, neutral,* or *against* (Aldayel and Magdy, 2021). For example, the statement "we have wasted enough time already and need to act on climate change now!" might be classified as having negative sentiments. However, its stance towards climate action is very favorable. Stance detection has been employed to study opinions on social media across a wide range of topics (e.g., Grčar et al., 2017; Zhou and Elejalde, 2024; Kim et al., 2025).

Most existing algorithms for stance detection are based on supervised machine learning and require an annotated training dataset (Aldayel and Magdy, 2021). They are typically built upon pre-trained language models that learn from vast amounts of textual data (Acheampong et al., 2021). Among the most popular models of this type is the Bidirectional Encoder Representations from Transformers (BERT; Devlin et al., 2019). BERT models have been developed for various languages and successfully applied in diverse contexts, including sentiment analysis, fraud detection, and biomedical research (Gardazi et al., 2025). However, supervised models like BERT require fine-tuning to better process context-specific language using a dedicated dataset (Alturayeif et al., 2023; Chuang, 2023). In stance detection, most available datasets are highly specific to their training context and do not generalize well to other datasets or topics, which raises questions about their applicability in broader contexts (Schiller et al., 2021; Ng and Carley, 2022). On the other hand, unsupervised machine learning methods structure data in clusters based on similarity. These methods require no pre-annotated training data, making their development significantly more efficient and potentially reducing the biases of supervised methods; however, they still lack extensive research and easy-to-implement solutions for researchers (Aldayel and Magdy, 2021).

Another popular method for analyzing texts, particularly their emotions, is dictionary analysis, which counts words associated with specific categories (Markowitz, 2024; Rathje et al., 2024). One popular method for dictionary-based methods is the Linguistic Inquiry and Word Count (LIWC; Boyd et al., 2022). However, such methods do not consider broader sentence contexts, leading to misinterpretations and sometimes poor performance (Tausczik and Pennebaker, 2010; Rathje et al., 2024).

Large language models (LLMs) are another easy-to-use solution for analyzing texts, which have gained significant attention recently and are showing promising results in the psychological sciences (Rathje et al., 2024; Bunt et al., 2025). While conventional BERT models may still perform better in specific contexts when sufficiently fine-tuned (Liyanage et al., 2024), LLMs offer a significantly faster and more cost-effective solution for text analysis, as no training set needs to be developed. For stance detection in particular, there is some evidence that LLMs yield valid results for social media content (Gambini et al., 2024). However, further validation for texts outside social media posts and in languages other than English is still pending.

As part of a larger pre-registered study on perceptions and portrayals of climate activism in German news outlets (Mayrhofer et al., 2026), we developed and validated methods for classifying stances and emotions towards climate activists in extensive text data. The current study presents the details of this procedure and contrasts several text annotation methods. In contrast to much previous work in stance detection, which has focused primarily on English-language social media posts, our study evaluates annotation methods on non-English, long-form political texts drawn from real-world news coverage and user discussions.

We collected news articles on climate protests and related user comments from eight major German news outlets. Three human raters classified a random subset of texts based on their stances towards the activist group Fridays for Future (FFF) and the level of anger conveyed in the texts. Anger was selected for investigation as it represents one crucial emotion driving people towards protest activities (Contreras et al., 2024; Stanley et al., 2021, 2023). We then compared the performance of GPT-3.5, GPT-4 Turbo, LIWC, and two GermanBERT models fine-tuned on validated stance-detection datasets to human annotations.

Thus, the contribution of this study is not to introduce a novel NLP architecture, but to provide a transparent validation of accessible annotation methods for German long-form political texts. More specifically, we evaluate whether readily deployable LLM-based approaches can provide reliable stance and emotion annotations in contexts where researchers often lack the resources to construct large task-specific training corpora or fine-tune supervised models.

Methods

Original dataset collection

For the current study, we randomly selected a set of articles and comments from a larger dataset (Mayrhofer et al., 2026), collected as follows.

News outlets were selected based on the following criteria: (1) nationwide reach, (2) generalist reporting (i.e., not focused on a single topic), (3) a minimum readership of 100,000, and (4) the presence of comment sections. Outlets were categorized as left- or right-leaning using political classifications from eurotopics.net and ground.news. We searched for articles related to FFF and LG via the search engine DuckDuckGo, using the keywords "letzte generation", "letzten generation", "aufstand der letzten generation", "fridays for future", and "fridays-for-future". URLs were collected from August 2018, marking the start of FFF with Greta Thunberg's school strikes, through December 2023, the time of data collection. For LG, we included only URLs posted after January 22, 2022, when they first blocked a street in Germany to protest. To avoid distortions from unrelated controversies, we excluded articles published after October 19, 2023, when Greta Thunberg began publicly discussing the Palestine conflict.

We then built custom scrapers for each news outlet to extract article texts, metadata (e.g., date, title, subtitle), and user comments from each URL. Scraping was conducted using Python (v3.10.11) and

Selenium (v4.15.2) between December 28, 2023, and January 6, 2024. In total, we retrieved 5,708 articles (FFF: 2,307; LG: 3,401) and 378,631 comments (FFF: 153,968; LG: 224,663). After exclusions, we analyzed 2,376 articles (FFF: 859; LG: 1,517) and 192,973 comments (FFF: 66,265; LG: 126,708; see Table 1). Data exclusion details are provided in Supplementary Table S1. Article distribution was reasonably balanced across outlets (most accounted for 9%–12% of the overall article count), except for Bild.de (7.8%), Welt.de (21%), and Zeit.de (17%). Comment data was highly skewed, with 86% of all comments and 93.5% of FFF-related comments coming from Welt.de and Zeit.de. Therefore, only these two outlets were included in the commentary analysis.

We validated our scraping by comparing article counts with those of Dablander et al. (2025); (see Supplementary Table S2 for details). After exclusions, they retrieved more articles (3,210 vs. 1,793). This might have been caused by (1) different methods used for filtering (Dablander et al. used GPT to filter articles, as we applied a conservative rule-based filter requiring explicit mentions of FFF or LG), and (2) different search engines used to collect relevant news articles (Dablander et al. used Google or media websites, while we used DuckDuckGo). While our approach was less exhaustive, we see no indication of systematic bias between the collection methods.

Language processing and model comparison

We compared OpenAI's GPT-3.5 and GPT-4 Turbo, and two BERT models fine-tuned on existing German general-topic training sets, to evaluate stances towards FFF. We limited ourselves to stances towards FFF as this study was part of a larger project investigating diachronic trends in support towards FFF and did not include stances on any other groups or topics. Anger was annotated using GPT-3.5, GPT-4 Turbo, and the Linguistic Inquiry and Word Count version 22 (LIWC-22) (Boyd et al., 2022).

GPT-3.5, GPT-4 Turbo annotated stances on a 5-point scale with *1 = strongly against*, *2 = slightly against*, *3 = neutral*, *4 = slightly in favor*, *5 = strongly in favor* (later recoded to −2 to 2), and the anger level of a text using a 3-point scale with *1 = no anger*, *2 = little anger*, *3 = high anger* (later recoded to 0 to 2). For user comments, we included an *unrelated* label in the stance annotations. This label was assigned when the text did not directly or indirectly refer to FFF or its members, or when it was unclear who the comment was targeted at. In news articles, we filtered texts beforehand by excluding those that did not include one of the terms *"fff," "fridays," "greta thunberg," "letzte generation," "letzten generation,"* or the frequently used nicknames *"chaoten" or "kleber"*.

The GPT model's temperature was set to 0, resulting in near-deterministic decoding. Other sampling parameters were left at their default value, meaning that no additional randomness was introduced (top_p = 1), only a single output was generated per prompt ($n = 1$), and no penalties for token repetition or reuse were applied (frequency_penalty = 0, presence_penalty = 0). Following common practice, we adjusted the temperature while keeping top_p at its default, as both parameters serve similar roles in controlling sampling variability. Output constraints were primarily imposed via the user prompt, not via API-side structured-output settings. Specifically, the prompt instructed the model to return exactly one label for support and one label for anger in the format: Support = [.]; Anger = [.]. The exact prompts are provided in Supplementary materials B, C.

For our BERT models, we used GermanBERT (Chan et al., 2020) and fine-tuned it on two multi-target training sets, X-Stance and CHeeSE, producing two separate BERT-based classifiers. X-Stance contains 48,612 German comments written by Swiss election candidates in response to 150 questions, labeled as either *in favor* or *against* (Vamvas and Sennrich, 2020). CHeeSE contains 1,970 news articles paired with 91 multi-topic questions, resulting in a collection of 3,693 pairs manually annotated into four categories: *in favor*, *against*, *discussing*, and *unrelated* (Mascarell et al., 2021).

Both training sets offer distinct advantages and limitations for our purpose. While X-Stance is based on a large corpus of training texts, the average token length (i.e., the number of subunits in texts) is significantly shorter than that of most news articles. However, it is well-suited for user comments. Additionally, the model lacks a neutral annotation category, which might be especially relevant when analyzing news articles that often aim to provide unbiased accounts of events. In contrast, CHeeSE was trained on real news articles but relies on a smaller and highly imbalanced training set. Nearly 40% of the texts are labeled as *unrelated*, meaning they do not contribute stance information, while only 8% are labeled as *against*.

To measure the relative amount of anger in the texts, we employed a bag-of-words approach using LIWC, a widely used tool for word frequency analysis (Eichstaedt et al., 2021; Kučera and Mehl, 2022; Berger and Packard, 2022). To analyze German news articles and comments, we employed the German word list (Meier et al., 2019).

TABLE 1 Number of news articles and user comments of the original dataset (Mayrhofer et al., 2026) after data exclusion for both FFF and LG.

Group	Left wing				Right wing				Total
	Spiegel	SZ	TAZ	Zeit	Bild	FAZ	Focus	Welt	
News articles									
FFF	97	105	89	135	051	127	074	181	859
LG	146	204	132	276	135	160	157	307	1,517
Total	243	309	221	411	186	287	231	488	2,376
User comments									
FFF	—	1	2,255	24,290	—	810	1,441	41,975	70,772
LG	—	141	5,729	58,843	—	6,786	14,985	67,865	154,349
Total	—	142	7,984	83,133	—	7,596	16,426	109,840	225,121

Human annotation and model validation in the random subsample

To assess the relative accuracy of the models' performance, we asked three German-speaking members of the research team to annotate the stances of 160 news articles on FFF (20 articles per outlet; 990 tokens on average per text, calculated using tiktoken v0.12.0 and cl100k_base encoder) and 150 related user comments (30 per outlet; 128 tokens on average per text). They also annotated the level of anger conveyed in 160 news articles on LG (20 articles per outlet; 813 tokens per text on average) and 180 related user comments (30 per outlet; 91 tokens per text on average). Annotation instructions are provided in Supplementary material A.

Before annotating the final test set, all three raters completed a pilot round on a small set of independent texts and then met to discuss disagreements. Annotators then labeled the full test set without outlet names to minimize bias. Texts were annotated and rated on the same five-point scale used for the GPT prompts, then aggregated and mapped to fit the X-Stance and CHeeSE models' label schemes, as shown in Figure 1. Final stances and anger levels were determined by a majority vote among the three human raters' annotations. Texts for which no majority on a single label was reached were excluded from the test set and not used for model evaluation.

We used various metrics to assess interrater agreement. Kendall's W (W) assesses the agreement of ordinal ratings from two or more raters and was tie-corrected to reduce bias (Gisev et al., 2013). Intraclass correlation (ICC) is used to assess consistency (rather than absolute agreement) between raters (Koo and Li, 2016). $F1_{binary}$ represents the harmonic mean between precision (true positive detections) and recall (the proportion of positive cases detected), and can be used for ratings with two classes, such as X-Stance in our case. For the multi-class case (as in our GPT and CHeeSE models), $F1_{micro}$ and $F1_{macro}$ can be used. $F1_{micro}$ pools all ratings and calculates the average, creating a bias towards more frequent classes. $F1_{macro}$ calculates the performance for each class and averages the metrics, giving all classes the same weight (Jurafsky and Martin, 2009). In stance annotations for user comments, we also reported Fleiss' Kappa (K) for the *unrelated* category, as this label does not follow an ordinal scaling like the other levels (Gisev et al., 2013). To compare the ordinal human majority vote with metric LIWC scores, we computed Kendall's Tau (T), which is appropriate for assessing rank correlations between ordinal and continuous measures. We applied the same statistics to GPT outputs to enable a direct comparison of GPT and LIWC performance (Khamis, 2008). Confidence intervals for Kendall's W, F1 scores, Fleiss' kappa, and Kendall's tau were estimated by bootstrap and are reported as bias-corrected and accelerated (BCa) 95% confidence intervals, based on 2,000 bootstrap replicates, in accordance with common recommendations for non-parametric confidence interval estimation (Grothe & Schmid, 2011; Chernick, 2012; Zapf et al., 2016).

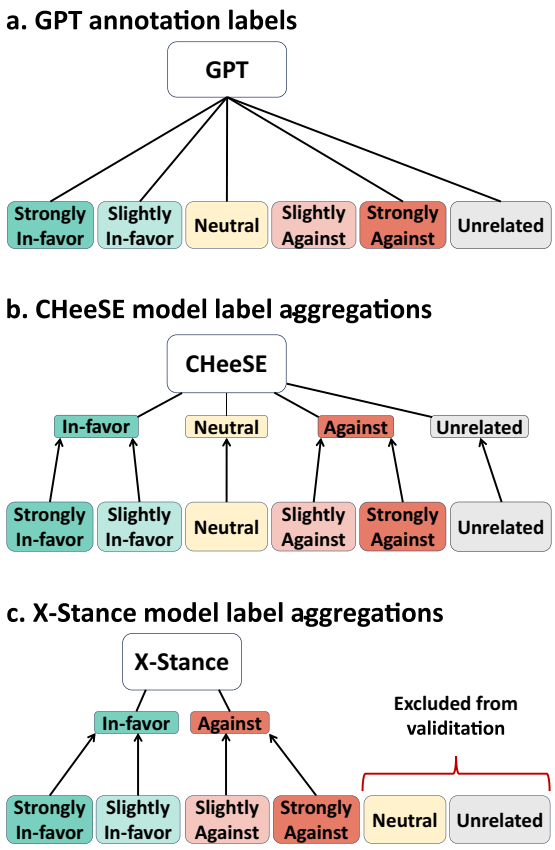


FIGURE 1
Label aggregation of human annotations for each language model. Human raters annotated texts on five levels (*Strongly against* to *Strongly in-favor*) and the additional label *unrelated* for user comments. **(a)** GPT annotations used the same five levels as human annotations. **(b)** CHeeSE included four categories, so the two favorable and two opposing human labels were aggregated. **(c)** X-Stance only provides binary labels to indicate if a text is written *in-favor* or *against*, so human labels were again aggregated as in Cheese. Texts labeled as *Neutral* or *Unrelated* were excluded from the model evaluation.

Results

Humans agreed with each other on stance and anger annotations

There was significant agreement between human raters both for stances ($W = 0.78$, $ICC = 0.72$, $K = 0.66$, all $p < 0.001$) and anger annotations ($W = 0.62$, $ICC = 0.4$, all $p < 0.001$). Annotations from language models were contrasted with the majority vote of the human ratings. No majority was reached in 40 out of 310 texts for stance labels (23 news articles and 17 user comments) and in 42 out of 340 texts for anger labels (20 news articles and 22 user comments). These cases were excluded from model evaluation. Label distributions were often imbalanced, especially for stance annotations in user comments, where the majority of comments were labeled as *unrelated* to FFF, and the remaining ones were mostly labeled as *strongly against* (Table 2).

GPT-4 Turbo outperformed employed BERT models for stance annotations

For the stance detection task, we compared four language models in total (two versions of GPT and two BERT models) to the human majority vote. First, we found that GPT-4 Turbo performed best on most metrics, both for news articles ($W = 0.81$, $ICC = 0.68$, both $p < 0.001$) and user comments ($W = 0.91$, $ICC = 0.86$, $K = 0.64$, all $p < 0.001$; Table 3). The agreement between GPT-4 Turbo and the majority human vote was usually higher than the agreement between humans. The raw distribution of stance annotations is shown in Supplementary Table S3.

TABLE 2 Distribution of human majority labels for stance and anger annotations.

Label	News articles		User comments	
	n	%	*n*	%
Stance annotations				
Strongly in-favour	56	35	11	8
Slightly in-favour	38	24	2	2
Neutral	22	14	3	2
Slightly against	7	4	4	3
Strongly against	14	9	34	26
Unrelated	—	—	79	59
No majority achieved	23	14	17	8
Anger annotations				
No anger	89	56	99	55
Little anger	36	23	44	25
High anger	15	9	15	8
No majority achieved	20	12	22	12

TABLE 3 Inter-rater agreements between all language models and the human majority vote for stances.

Language model	K_{Fleiss}	$W_{Kendall}$	$ICC_{consistency}$	$F1_{macro}$	$F1_{micro}$
News articles					
GPT-3.5	—	0.80*** [0.72, 0.85]	**0.71*** [0.61, 0.78]	**0.48** [0.39, 0.59]	**0.50** [0.42, 0.59]
GPT-4 Turbo	—	**0.81*** [0.74, 0.86]	0.68*** [0.58, 0.76]	0.35 [0.27, 0.46]	0.44 [0.36, 0.52]
CheeSE (BERT)	—	0.55*** [0.47, 0.62]***	0.06*** [−0.11, 0.23]	0.29 [0.23, 0.37]	0.36 [0.28, 0.44]
X-Stance (BERT)	—	0.51*** [0.42, 0.61]***	0.03*** [−0.16, 0.21]	0.27[a] [0.14, 0.41]	
Human annotators	—	0.69*** [0.62, 0.75]	0.57*** [0.49, 0.65]	—	—
User comments					
GPT-3.5	0.26*** [0.09, 0.41]	0.82*** [0.68, 0.92]	0.66*** [0.46, 0.79]	0.26 [0.20, 0.34]	0.54 [0.32, 0.65]
GPT-4 Turbo	**0.64*** [0.49, 0.76]	**0.91*** [0.81, 0.96]	**0.86*** [0.75, 0.92]	**0.39** [0.31, 0.51]	**0.73** [0.20, 0.82]
CheeSE (BERT)	0.12*** [−0.05, 0.31]	0.39**** [0.13, 0.66]	−0.02**** [−0.68, 0.40]	0.28 [0.21, 0.33]	0.62 [0.63, 0.70]
X-Stance (BERT)	—	0.59*** [0.43, 0.69]	0.17*** [−0.11, 0.42]	0.56[a] [0.39, 0.71]	
Human annotators	0.45*** [0.34, 0.56]	0.84*** [0.68, 0.92]	0.88*** [0.80, 0.93]	—	—

Inter-rater reliabilities for stances toward FFF, comparing language-model labels to the human majority vote and human annotators with each other. Reported statistics: Fleiss' kappa (only for the "Unrelated" category), Kendall's W (tie-corrected), $F1_{micro}$ and $F1_{macro}$ for GPT and CHeeSE models, and F1 (binary) for X-Stance models. $F1_{macro}$ better reflects performance across individual class levels when data are highly imbalanced, as in our user comment annotations, whereas $F1_{micro}$ reflects overall accuracy. Numbers in brackets represent bias-corrected and accelerated (BCA) 95% confidence intervals obtained via bootstrapping with 2000 repetitions. Bold values indicate the highest score across models for each metric.
*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$, F1 scores do not provide statistical tests.
[a]F1 scores for X-stance represent F1 binary scores.

Second, GPT-3.5 performed similarly well as GPT-4 Turbo for news articles ($W = 0.81$, $ICC = 0.71$, all $p < 0.001$), but slightly worse in user comments, although the difference was still significant ($W = 0.82$, $ICC = 0.66$, $K = 0.26$, all $p < 0.001$). GPT-3.5 particularly struggled with identifying user comments "unrelated" to FFF (Figure 2). After excluding texts labeled as *unrelated,* there was no

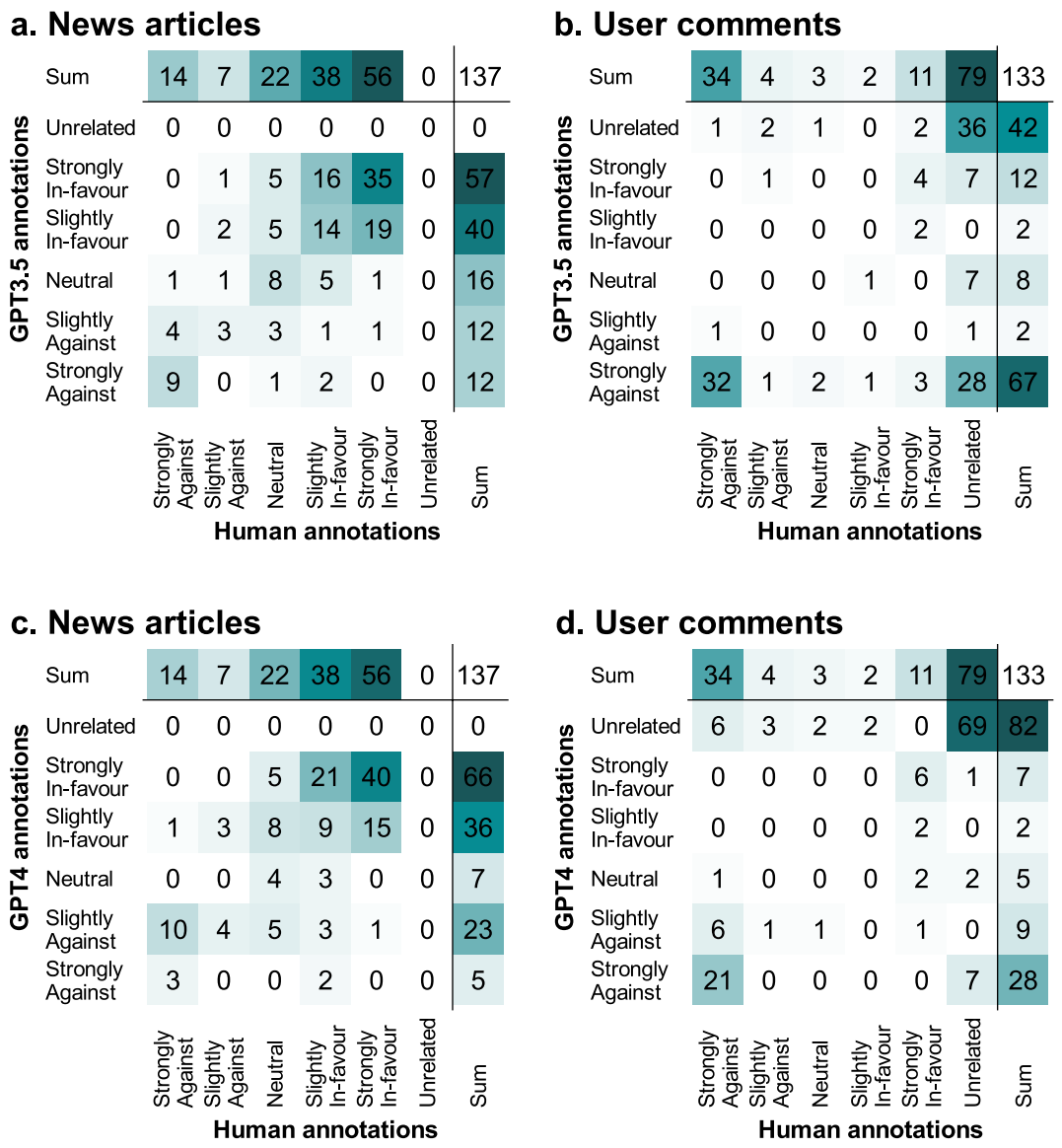


FIGURE 2
Confusion plots for stance annotations for GPT-3.5 [**(a)** news articles; **(b)** user comments] and GPT-4 Turbo [**(c)** news articles; **(d)** user comments]. Texts for which no majority for a single label was reached were removed from the dataset.

significant difference between GPT-3.5 annotations and the human majority vote in news articles ($W = 9{,}499$, $p = 0.43$) or user comments ($W = 2{,}193$, $p = 0.18$).

Third, none of the BERT models achieved significant agreement with the human majority vote neither for news articles (CHeeSE: $W = 0.54$, $p = 0.26$, $ICC = 0.06$, $p = 0.23$; X-Stance: $W = 0.51$, $p = 0.40$, $ICC = 0.03$, $p = 0.39$) nor user comments (CHeeSE: $W = 0.39$, $p = 0.66$, $ICC = -0.02$, $p = 0.74$, $K = 0.12$, $p = 0.17$; X-Stance: $W = 0.59$, $p = 0.19$, $ICC = 0.17$, $p = 0.11$).

GPT-4 Turbo outperformed LIWC for anger annotations

For anger, we compared three language models in total (two versions of GPT and LIWC) to the human majority vote. First, we found that GPT-4 Turbo achieved the best performance with significant agreement with the human majority vote, both for news articles ($T = 0.52$, $W = 0.77$, $ICC = 0.57$, all $p < 0.001$) and user comments ($T = 0.49$, $W = 0.76$, $ICC = 0.52$, all $p < 0.001$). Again, the agreement between GPT and the majority vote was higher than the agreement between humans (Table 4). However, GPT − 4 Turbo showed a significant bias towards higher anger ratings compared to the human majority vote (Figure 3; Supplementary Table S4), both in news articles ($W = 12{,}814$, $p < 0.001$) and user comments ($W = 15{,}414$, $p < 0.001$). A qualitative description of common disagreements between GPT and human ratings is provided in Supplementary material E.

Second, GPT-3.5 performance was equivalent to GPT-4 Turbo, although agreement was slightly lower both for news articles ($T = 0.46$, $W = 0.75$, $ICC = 0.49$, all $p < 0.001$) and user comments ($T = 0.43$, $W = 0.73$, $ICC = 0.44$, all $p < 0.001$). The agreement with the majority vote was still higher than the agreement between humans and GPT-3.5, which showed a similar bias towards higher anger ratings as GPT-4 Turbo (Figure 3), both in news articles ($W = 12{,}910$, $p < 0.001$) and user comments ($W = 15{,}925$, $p < 0.001$).

Third, the correlation between the majority vote and LIWC was statistically significant for both news articles ($T = 0.28$, $p < 0.001$) and

TABLE 4 Inter-rater agreements between all language models and the human majority vote for anger.

Language model	$T_{Kendall}$	$W_{Kendall}$	$ICC_{consistency}$	$F1_{macro}$	$F1_{micro}$
News articles					
GPT-3.5	0.46*** [0.34, 0.57]	0.75*** [0.69, 0.81]	0.49*** [0.35, 0.60]	0.44 [0.37, 0.54]	0.49 [0.41, 0.58]
GPT-4 Turbo	**0.52***** [0.40, 0.61]	**0.77***** [0.71, 0.83]	**0.57***** [0.45, 0.67]	**0.52** [0.43, 0.62]	**0.52** [0.44, 0.60]
LIWC	0.28*** [0.14, 0.39]	—	—	—	—
Human annotators	0.32*** [0.18, 0.44] 0.41*** [0.28, 0.53] 0.50*** [0.38, 0.59]	0.62*** [0.55, 0.69]	0.40*** [0.33, 0.47]	—	—
User comments					
GPT-3.5	0.43*** [0.29, 0.54]	0.73*** [0.65, 0.79]	0.44*** [0.31, 0.56]	0.45 [0.38, 0.54]	0.53 [0.44, 0.60]
GPT-4 Turbo	**0.49***** [0.37, 0.58]	**0.76***** [0.70, 0.81]	**0.52***** [0.40, 0.63]	**0.50** [0.42, 0.59]	**0.56** [0.48, 0.64]
LIWC	0.34*** [0.20, 0.47]	—	—	—	—
Human annotators	0.35*** [0.20, 0.48] 0.44*** [0.32, 0.53] 0.45*** [0.33, 0.55]	0.63*** [0.56, 0.69]	0.42*** [0.33, 0.51]	—	—

Inter-rater reliabilities for stances toward FFF, comparing language-model labels to the human majority vote and within human annotators. Reported statistics: Kendall's W (tie-corrected), $F1_{micro}$ and $F1_{macro}$ for GPT and CHeeSE models, and F1 (binary) for X-Stance models. $F1_{macro}$ better reflects performance across individual class levels when data are highly imbalanced, as in our user comment annotations, whereas $F1_{micro}$ reflects overall accuracy. Kendall's Tau (to compare metric LIWC scores to ordinal human ratings). For human annotators, Tau scores between all rater pairs are presented. Numbers in brackets represent bias-corrected and accelerated (BCA) 95% confidence intervals obtained via bootstrapping with 2000 repetitions. alues indicate the highest score across models for each metric.
*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$, F1 scores do not provide statistical tests.

user comments ($T = 0.34$, $p < 0.001$). However, the correlation was lower than for both versions of GPT and lower than the correlation between humans.

Discussion

Stance detection algorithms represent valuable tools for measuring people's attitudes towards specific topics in natural settings and large datasets. However, current methods are often found to lack generalizability outside the contexts in which they were trained (Schiller et al., 2021; Ng and Carley, 2022). LMMs have the potential to overcome this limitation; however, they still lack validation for such purposes, particularly outside social media posts and in languages other than English. In this study, we addressed this methodological gap by comparing the performance of two versions of GPT in a stance detection task to more conventional algorithms based on the language model BERT when annotating German news articles and user comments on climate activism. Additionally, we compared GPT's performance at identifying angry language in texts with that of the state-of-the-art dictionary method, LIWC. Performance was evaluated by comparing annotations to the majority vote of three human annotators who rated a subset of texts on their stance towards climate activists and the use of angry language.

First, we found that the agreement between both GPT-3.5 and GPT-4 Turbo and the human majority vote was relatively high for both stances and anger. Furthermore, the agreement between the two GPT versions and the majority vote was usually higher than the agreement among humans, indicating that GPT is a valid rater for such annotation tasks. These results align with recent results and discussions on the usefulness of GPT in behavioral research (Rathje et al., 2024; Feuerriegel et al., 2025). There was, however, a significant bias towards higher anger ratings in both GPT-3.5 and GPT-4 Turbo, indicating that GPT rated texts as angrier than our human annotators did. We could not find such a bias in stance ratings. While previous research also suggests that GPT-4 might overattribute anger (Rathje et al., 2024), our results may also stem from a domain-specific bias, in which anger is often associated with climate change news and comments. Further research is thus needed to evaluate GPT's capabilities in this domain and how fine-tuning LLMs may improve their performance. Importantly, inter-rater agreement for anger was only moderate, indicating that discrete emotions were often difficult to discern reliably in our texts. This uncertainty can affect model evaluation, but by excluding texts without a majority vote, we restricted comparisons to cases with higher annotator consensus, thereby focusing evaluation on texts with relatively greater certainty.

Second, while GPT-4 Turbo performed slightly better than GPT-3.5, the performance difference was relatively small, except for user comments, where GPT was prompted to decide whether texts were related to climate activists. While GPT-4 Turbo accurately identified these cases when compared to human ratings, GPT-3.5 did not. Still, this suggests that, depending on the task, older versions of GPT may

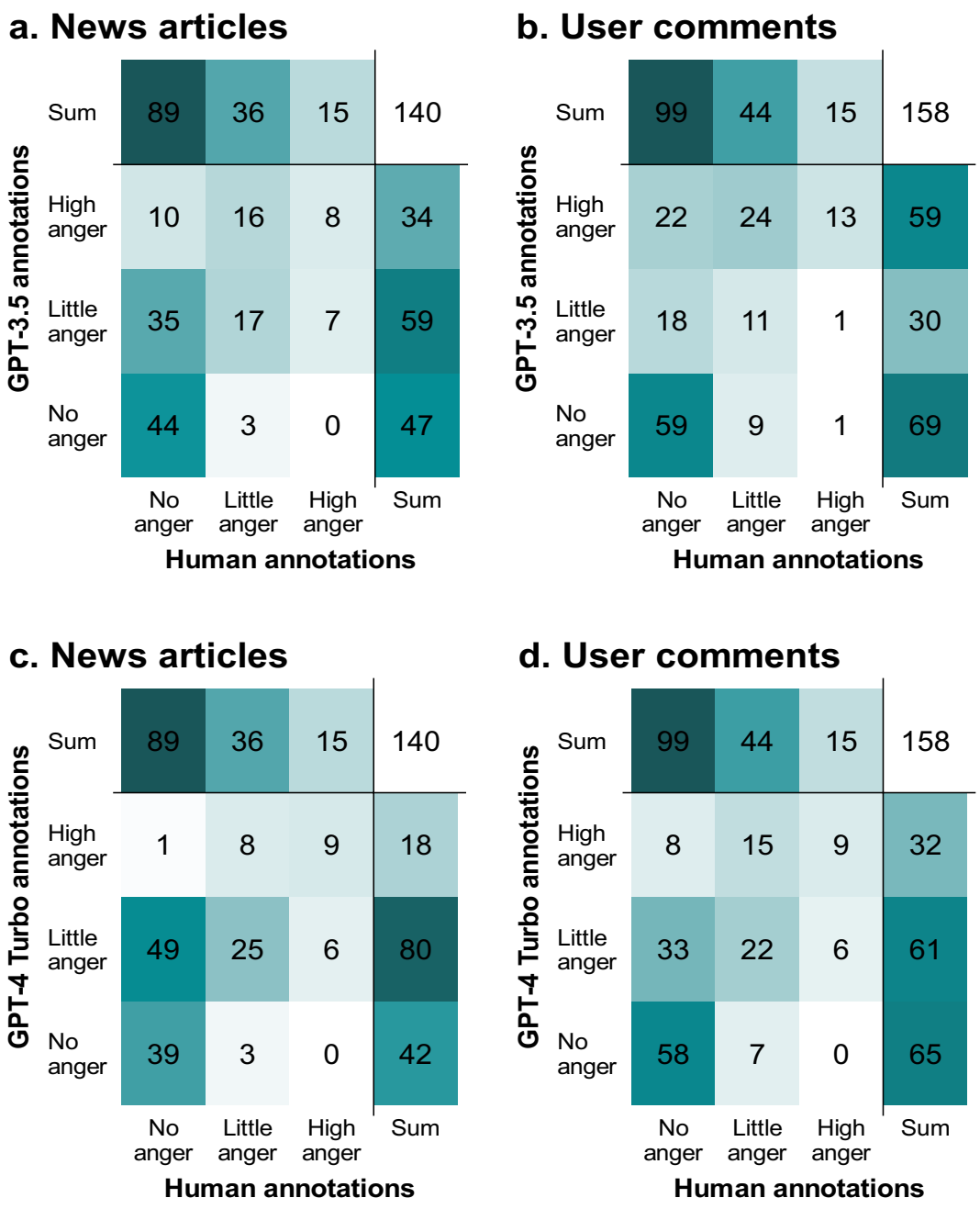


FIGURE 3
Confusion plots for anger annotations for GPT-3.5 [**(a)** news articles; **(b)** user comments] and GPT-4 Turbo [**(c)** news articles; **(d)** user comments]. Texts for which no majority for a single label was reached were removed from the dataset.

perform well enough for annotation while being significantly more cost-efficient.

Third, neither X-Stance nor CheeSE showed significant agreement with the human majority vote. Importantly, texts from both the X-Stance and CHeeSE dataset differ substantially from the texts we examined (e.g., in token length or topics), which likely harmed model performance. Developing tailored training sets would likely boost performance. However, this requires considerable time and resources and is therefore impractical for many researchers. Consequently, our analysis reflects the "out-of-the-box" performance of stance-detection models trained on broad, general-purpose datasets and, consistent with other work, highlights their limited generalizability to texts that diverge from their training data (Schiller et al., 2021; Ng and Carley, 2022), limiting their usability for researchers.

The evaluation of the BERT models' performance is also restricted by the labeling procedure our human annotators underwent. We used more labels when annotating the data than CHeeSE and X-Stance provide, allowing for additional differentiation in attitudes. To minimize potential biases, we selected our levels to facilitate easy aggregation. Furthermore, our labels were highly imbalanced, especially for user comments, where center labels for stance were underrepresented. While this might reflect strongly polarized commenter stances, it also reflects an imbalance in model evaluation across underrepresented levels. It is therefore difficult to assess how well our models

distinguished between centrist stance categories. This issue is evident in substantially lower $F1_{macro}$ vs. $F1_{micro}$ scores for stance on user comments, indicating weaker performance on less common classes.

Finally, although LIWC scores correlated significantly with the human majority vote, GPT demonstrated substantially higher correlations and significant agreement scores. LIWC is only one example of a dictionary tool, and alternative lexicons might yield better results (e.g., Kim and Klinger, 2018; Fehle et al., 2021). However, our findings are consistent with a growing body of literature indicating that, while dictionary-based approaches offer greater interpretability, they often fall short in terms of accuracy (Rathje et al., 2024; Klähn et al., 2025; Feuerriegel et al., 2025). Future research is needed to determine how to effectively combine these approaches to strike a balance between accuracy and interpretability.

Conclusion

This study highlights the potential of LLMs to analyze attitudes and expressions of anger in news articles and user commentaries. Our findings show that (1) GPT annotated stances toward climate activists and expressions of anger closely aligned with human judgments, (2) conventional BERT-based stance-detection models struggled to transfer to texts that differed substantially from their training data, and (3) GPT outperformed LIWC in detecting anger.

Beyond comparing model performance, our findings provide a practical validation of LLM-based annotation for non-English, long-form political texts drawn from real-world media contexts. While much previous work has focused on English social media data or highly task-specific training settings, our results suggest that readily deployable LLMs can provide reliable, cost-efficient annotation solutions in settings where researchers may lack the resources to build large training corpora or fine-tune supervised models. More broadly, these findings contribute to the growing methodological literature on computational language analysis and help clarify the conditions under which LLMs may be preferable to conventional dictionary-based or supervised approaches for psychological and social-science research.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found at: https://osf.io/yr7gk/.

Author contributions

LM: Conceptualization, Formal analysis, Funding acquisition, Methodology, Software, Visualization, Writing – original draft. MF: Software, Writing – review & editing. SF: Conceptualization, Methodology, Writing – review & editing. JK: Conceptualization, Writing – review & editing. BT: Conceptualization, Visualization, Writing – review & editing. CL: Funding acquisition, Supervision, Writing – review & editing. MM: Conceptualization, Methodology, Software, Supervision, Writing – review & editing.

Funding

The author(s) declared that financial support was received for this work and/or its publication. This work was supported by a student grant from the University of Vienna. BT was partly funded by an Austrian Science Fund (FWF) "DK Cognition and Communication 2": W1262-B29 [10.55776/W1262]. Open access funding provided by University of Vienna.

Conflict of interest

The author(s) declared that this work was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declared that Generative AI was used in the creation of this manuscript. During the preparation of this work, the authors used OpenAI's GPT-4o to improve the readability of the initial submission manuscript. After usage, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpos.2026.1745587/full#supplementary-material

References

Acheampong, F. A., Nunoo-Mensah, H., and Chen, W. (2021). Transformer models for text-based emotion detection: a review of BERT-based approaches. *Artif. Intell. Rev.* 54, 5789–5829. doi: 10.1007/s10462-021-09958-2

Aldayel, A., and Magdy, W. (2021). Stance detection on social media: state of the art and trends. *Inf. Process. Manag.* 58:102597. doi: 10.1016/j.ipm.2021.102597

Alturayeif, N., Luqman, H., and Ahmed, M. (2023). A systematic review of machine learning techniques for stance detection and its applications. *Neural Comput. Applic.* 35, 5113–5144. doi: 10.1007/s00521-023-08285-7

Berger, J., and Packard, G. (2022). Using natural language processing to understand people and culture. *Am. Psychol.* 77, 525–537. doi: 10.1037/amp0000882

Biber, D., and Finegan, E. (1988). Adverbial stance types in English. *Discourse Process.* 11, 1–34. doi: 10.1080/01638538809544689

Bittermann, A., and Fischer, A. (2024). Natural language processing in psychology. *Z. Psychol.* 232, 143–146. doi: 10.1027/2151-2604/a000568

Boyd, R. L., and Schwartz, H. A. (2021). Natural language analysis and the psychology of verbal behavior: the past, present, and future states of the field. *J. Lang. Soc. Psychol.* 40, 21–41. doi: 10.1177/0261927X20967028

Boyd, R. L., Ashokkumar, A., Seraj, S., and Pennebaker, J. W. (2022). *The Development and Psychometric Properties of LIWC-22*. Austin, TX: University of Texas at Austin.

Bunt, H. L., Goddard, A., Reader, T. W., and Gillespie, A. (2025). Validating the use of large language models for psychological text classification. *Front. Soc. Psychol.* 3:1460277. doi: 10.3389/frsps.2025.1460277

Chan, B., Schweter, S., and Möller, T. (2020). "German's next language model," in *Proceedings of the 28th International Conference on Computational Linguistics*. Eds. D. Scott, N. Bel and C. Zong (International Committee on Computational Linguistics), 6788–6796. doi: 10.18653/v1/2020.coling-main.598

Chen, G., Tan, B., Laham, N., Tracey, T. J. G., Lapinski, S., and Liu, Y. (2025). A bibliometric review of natural language processing applications in psychology from 1991 to 2023. *Basic Appl. Soc. Psychol.* 47, 105–119. doi: 10.1080/01973533.2024.2433720

Chernick, M. R. (2012). Resampling methods. *WIREs Data Mining and Knowledge Discovery*, 2, 255–262. doi: 10.1002/widm.1054

Chuang, Y-S (2023) *Tutorials on Stance Detection using Pre-Trained Language Models: Fine-Tuning BERT and Prompting Large Language Models*. (arXiv:2307.15331). arXiv. doi: 10.48550/arXiv.2307.15331

Contreras, A., Blanchard, M. A., Mouguiama-Daouda, C., and Heeren, A. (2024). When eco-anger (but not eco-anxiety nor eco-sadness) makes you change! A temporal network approach to the emotional experience of climate change. *J. Anxiety Disord.* 102:102822. doi: 10.1016/j.janxdis.2023.102822

Dablander, F., Wimmer, S., and Haslbeck, J. (2025). Media coverage of climate activist groups in Germany. *Climatic Change* 178:144. doi: 10.1007/s10584-025-03959-8

Devlin, J, Chang, M-W, Lee, K, and Toutanova, K (2019) BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: J Burstein, C Doran and T Solorio *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*

Eichstaedt, J. C., Kern, M. L., Yaden, D. B., Schwartz, H. A., Giorgi, S., Park, G., et al. (2021). Closed- and open-vocabulary approaches to text analysis: a review, quantitative comparison, and recommendations. *Psychol. Methods* 26, 398–427. doi: 10.1037/met0000349

Fehle, J, Schmidt, T, and Wolff, C (2021) Lexicon-based sentiment analysis in German: systematic evaluation of resources and preprocessing techniques. In: K Evang, L Kallmeyer and R Osswald *Proceedings of the 17th Conference on Natural Language Processing (KONVENS 2021)*

Feuerriegel, S., Maarouf, A., Bär, D., Geissler, D., Schweisthal, J., Pröllochs, N., et al. (2025). Using natural language processing to analyse text data in behavioural science. *Nat. Rev. Psychol.* 4, 96–111. doi: 10.1038/s44159-024-00392-z

Gambini, M., Senette, C., Fagni, T., and Tesconi, M. (2024). Evaluating large language models for user stance detection on X (twitter). *Mach. Learn.* 113, 7243–7266. doi: 10.1007/s10994-024-06587-y

Gardazi, N. M., Daud, A., Malik, M. K., Bukhari, A., Alsahfi, T., and Alshemaimri, B. (2025). BERT applications in natural language processing: a review. *Artif. Intell. Rev.* 58:166. doi: 10.1007/s10462-025-11162-5

Gisev, N., Bell, J. S., and Chen, T. F. (2013). Interrater agreement and interrater reliability: key concepts, approaches, and applications. *Res. Soc. Adm. Pharm.* 9, 330–338. doi: 10.1016/j.sapharm.2012.04.004

Grčar, M., Cherepnalkoski, D., Mozetič, I., and Kralj Novak, P. (2017). Stance and influence of twitter users regarding the Brexit referendum. *Comput. Soc. Netw.* 4:6. doi: 10.1186/s40649-017-0042-6

Grothe, O., and Schmid, F. (2011). Kendall's $\mathscr{W}$ Reconsidered. *Communications in Statistics - Simulation and Computation* 40, 285–305. doi: 10.1080/03610918.2010.538791

Jurafsky, D., and Martin, J. H. (2009). *Speech and Language Processing: an Introduction to natural Language Processing, Computational Linguistics, and speech Recognition*. Upper Saddle River, NJ: Pearson Education.

Khamis, H. (2008). Measures of association: how to choose? *J. Diagn. Med. Sonogr.* 24, 155–162. doi: 10.1177/8756479308317006

Kim, E., and Klinger, R. (2018). Who feels what and why? Annotation of a literature corpus with semantic roles of emotions. In: Bender, EM, Derczynski, L, and Isabelle, P *Proceedings of the 27th International Conference on Computational Linguistics*

Kim, J., Kim, D., and Park, E. (2025). I know your stance! Analyzing twitter users' political stance on diverse perspectives. *J. Big Data* 12:14. doi: 10.1186/s40537-025-01083-z

Klähn, J., Borst-Graetz, J., and Burghardt, M. (2025). From dictionaries to LLMs – an evaluation of sentiment analysis techniques for German language data. *Comput. Human. Res.* 1:e4. doi: 10.1017/chr.2025.10005

Koo, T. K., and Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J. Chiropr. Med.* 15, 155–163. doi: 10.1016/j.jcm.2016.02.012

Kučera, D., and Mehl, M. R. (2022). Beyond English: considering language and culture in psychological text analysis. *Front. Psychol.* 13:819543. doi: 10.3389/fpsyg.2022.819543

Liyanage, C. R., Gokani, R., and Mago, V. (2024). GPT-4 as an X data annotator: unraveling its performance on a stance classification task. *PLoS One* 19:e0307741. doi: 10.1371/journal.pone.0307741

Markowitz, D. M. (2024). Can generative AI infer thinking style from language? Evaluating the utility of AI as a psychological text analysis tool. *Behav. Res.* 56, 3548–3559. doi: 10.3758/s13428-024-02344-0

Mascarell, L., Ruzsics, T., Schneebeli, C., Schlattner, P., Campanella, L., Klingler, S., et al (2021). "Stance detection in German news articles," in *Proceedings of the Fourth Workshop on Fact Extraction and VERification (FEVER)*. Dominican Republic. Association for Computational Linguistics. 66–77. doi: 10.18653/v1/2021.fever-1.8

Mayrhofer, L., Foramitti, M., Fassnacht, S., Köhler, J., Todorova, B., Lamm, C., et al. (2026). Radical climate protests shaped portrayals of moderate activists and reader attitudes in German news media. *NPJ Clim. Action*. doi: 10.1038/s44168-026-00361-7

Meier, T., Boyd, R. L., Pennebaker, J. W., Mehl, M. R., Martin, M., Wolf, M., et al (2019). *"LIWC auf Deutsch": The Development, Psychometrics, and Introduction of DE-LIWC2015*

Mihalcea, R., Biester, L., Boyd, R. L., Jin, Z., Perez-Rosas, V., Wilson, S., et al. (2024). How developments in natural language processing help us in understanding human behaviour. *Nat. Hum. Behav.* 8, 1877–1889. doi: 10.1038/s41562-024-01938-0

Ng, L. H. X., and Carley, K. M. (2022). Is my stance the same as your stance? A cross validation study of stance detection datasets. *Inf. Process. Manag.* 59:103070. doi: 10.1016/j.ipm.2022.103070

Rathje, S., Mirea, D.-M., Sucholutsky, I., Marjieh, R., Robertson, C. E., and van Bavel, J. J. (2024). GPT is an effective tool for multilingual psychological text analysis. *Proc. Natl. Acad. Sci.* 121:e2308950121. doi: 10.1073/pnas.2308950121

Schiller, B., Daxenberger, J., and Gurevych, I. (2021). Stance detection benchmark: How robust is your stance detection? *Künstl. Intell.* 35, 329–341. doi: 10.1007/s13218-021-00714-w

Stanley, , Hogg, T. L., Leviston, Z., and Walker, I. (2021). From anger to action: differential impacts of eco-anxiety, eco-depression, and eco-anger on climate action and wellbeing. *The Journal of Climate Change and Health* 1:100003. doi: 10.1016/j.joclim.2021.100003

Stanley, , Leviston, Z., Hogg, T., and Walker, I. (2023). Anger about climate inaction: the content of eco-anger shapes emotional and behavioral engagement with climate change. *OSF*. doi: 10.31234/osf.io/juc67

Tausczik, Y., and Pennebaker, J. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *J. Lang. Soc. Psychol.* 29, 24–54. doi: 10.1177/0261927X09351676

Vamvas, J, and Sennrich, R. (2020) *X-Stance: A Multilingual Multi-Target Dataset for Stance Detection*. (arXiv:2003.08385). arXiv. doi: 10.48550/arXiv.2003.08385

Zapf, A., Castell, S., Morawietz, L., and Karch, A. (2016). Measuring inter-rater reliability for nominal data – which coefficients and confidence intervals are appropriate? *BMC Medical Research Methodology* 16:93. doi: 10.1186/s12874-016-0200-9

Zhou, Z., and Elejalde, E. (2024). Unveiling the silent majority: stance detection and characterization of passive users on social media using collaborative filtering and graph convolutional networks. *EPJ Data Sci.* 13:28. doi: 10.1140/epjds/s13688-024-00469-y