Additional file 9

The adaptive community-response (ACR) method for collecting misinformation on social media

Julian Kauk[1], Helene Kreysa[1], André Scherag[2, 3], and Stefan R. Schweinberger[1, 3, 4]

[1] *Department of General Psychology and Cognitive Neuroscience, Friedrich Schiller University Jena, Jena, Germany*
[2] *Institute of Medical Statistics, Computer and Data Sciences, Jena University Hospital, Jena, Germany*
[3] *Michael Stifel Center Jena, Jena, Germany*
[4] *German Center for Mental Health (DZPG), Site Jena-Magdeburg-Halle, Germany*

February 1, 2024

Validating the ACR method through manual annotation of tweets

To gain a deeper understanding of the ACR method's ability to collect story-related tweets, we conducted a validation analysis. In this analysis, we manually annotated a sample of tweets, thereby comparing the ACR method's automated classification to an external human-based standard. We used the stories with excellent automated classification levels (AUC $\geq$ .9), expecting to observe similarly unambiguous classification by human raters. The performed procedure, results and conclusions are provided in the following sections.

Procedure

In total, there were 37 stories with excellent automated classification levels (AUC $\geq$ .9). To obtain robust estimates of task performance for each of these stories, we randomly sampled 30 tweets per story, which corresponds to a total number of annotated tweets $N = 37 \cdot 30 = 1110$. Two independent raters annotated all tweets. We adhered to the handbook on text annotation provided by Stollenwerk et al [7] to ensure quality of the annotations. The raters were accordingly (i) instructed to work independently, (ii) kept naïve to the purposes and aims of the study, and (iii) provided with explicit instruction how to annotate the tweets.

We determined whether a tweet was related to a story or not (Task I), considering a tweet to be related to a story if it clearly concerned the respective story. If a tweet was

considered to be related to a story, we performed natural language inference (NLI; see, e.g., [5]) to determine whether it was supportive, neutral, or contradictive (Task II). A tweet was considered supportive if it (i) was related to the story and (ii) it repeats the claim of the story without presenting any doubts or contradicting evidence. A tweet was considered neutral if it was related to the story, but without revealing the user's support of the story, e.g., due to irony. A tweet was considered contradictive if it was related to the story and contained clear contradicting sentences or words, e.g., a reference to a fact-check.

To ensure that a tweet in fact relates to a story, we considered a tweet only as related (Task I) if both raters confirmed its relation to the story. If the raters disagreed on their annotations in terms of Task II (NLI: supportive/neutral/contradictive), the corresponding tweet was annotated by a third independent rater, and a majority decision was subsequently used.

We computed Cohen's κ ($\kappa \in [-1, 1]$) to assess inter-rater reliability [see, e.g., 4], whereas values ≤ 0 mean chance agreement between the annotators, while $\kappa = 1$ means perfect agreement. We also report confusion matrices, indicating (dis)agreement between the raters on the level of each category.

Results

Inter-rater reliability

We found that both raters agreed to a substantial degree (for magnitude guidelines, see [2]) on their ratings regarding the relatedness of the tweets, as indicated by a Cohen's κ of .637. Table 1 shows the corresponding confusion matrix, confirming the good agreement between both raters. In total, the raters annotated the same category in 1047 (94.32%) tweets, indicating strong agreement of both raters.

		Rater B	
		Unrelated	Related
Rater A	Unrelated	63 (5.68)	52 (4.68)
	Related	11 (0.99)	984 (88.65)

Table 1: Confusion matrix for the Task I (Relatedness). The numbers in parentheses represent normalized ($N = 1110$) values, in percent.

We observed that Cohen's κ was slightly reduced for the NLI task, as indicated by a Cohen's κ of .537, corresponding to a moderate agreement (cf. [2]) between the annotators. This is also expressed in a lower agreement rate of 79.65%, corresponding to 783 tweets where both raters assigned the same category. Table 2 shows the corresponding confusion matrix. The highest disagreement was observed for the 'Neutral' category, meaning that, for instance, Rater A assigned the 'Neutral' category, while Rater B selected 'Supportive'. Such 'Neutral-Supportive' and 'Neutral-Contradictive' disagreements may be considered less severe, as the raters did not fundamentally disagree in their annotated categories. The annotators even occasionally selected opposing categories (for instance, Rater A choose

'Contradictive', while Rater B selected 'Supportive'); however, this pattern occurred in less than 8% of the tweets.

		Rater B		
		Contradictive	Neutral	Supportive
Rater A	Contradictive	149 (15.16)	8 (0.81)	70 (7.12)
	Neutral	4 (0.41)	16 (1.63)	101 (10.27)
	Supportive	7 (0.71)	10 (1.02)	618 (62.87)

Table 2: Confusion matrix for the Task II (NLI). Numbers in parentheses represent normalized ($N = 984$) values, in percent.

Task I: Relatedness

We found that 984 of the 1110 annotated tweets were related to their respective story, corresponding to an overall performance of 88.65% (95% CI[1] [86.63%, 90.46%]). We observed that the performance varied from story to story (see Figure 1), ranging from 50.0% to 100.0% success rate (median: 93.33%). Most of the stories ($22 \mathrel{\widehat{=}} 59.46\%$) showed performances $\geq .9$, and 10 stories (27.03%) even showed perfect performance, meaning that all annotated tweets in the sample were considered to be related.

Task II: Natural language inference (NLI)

Supporting tweets. We found that 750 out of the 984 related tweets were supportive for the respective stories, corresponding to a proportion of 76.3% (95% CI [73.51%, 78.92%]). We observed considerable heterogeneity between the stories (see Figure 2A), as indicated by proportions ranging from 14.81% to 100.0%. However, most of the stories showed high proportions of supportive tweets, as indicated by a (i) median proportion of 86.21% and (ii) low number of stories having less than 50% supportive tweets ($N = 6$ [16.22%]).

Neutral tweets. The 'Neutral' category was used relatively infrequently and only 70 (7.12%; 95% CI [5.59%, 8.91%]) tweets were annotated accordingly. On the level of the different stories (see Figure 2B), we observed that a majority of the stories had less than 10% neutral tweets ($N = 27$ [72.97%]), which was also expressed in (i) the low median proportion of neutral tweets (3.45%) and (ii) the high number of stories not having any neutral tweet ($N = 16$ [43.24%]).

Contradicting tweets. We found that 163 (16.58%; 95% CI [14.31%, 19.06%]) tweets were annotated to be contradictive. We observed considerable heterogeneity between the stories (see Figure 2C). While a majority of the stories had less than 10% contradicting tweets ($N = 24$ [64.86%]), there were 5 (13.51%) stories having at least 50% contradicting tweets.

[1] We used Clopper–Pearson interval [see, e.g., 8] to estimate binomial proportion confidence intervals.

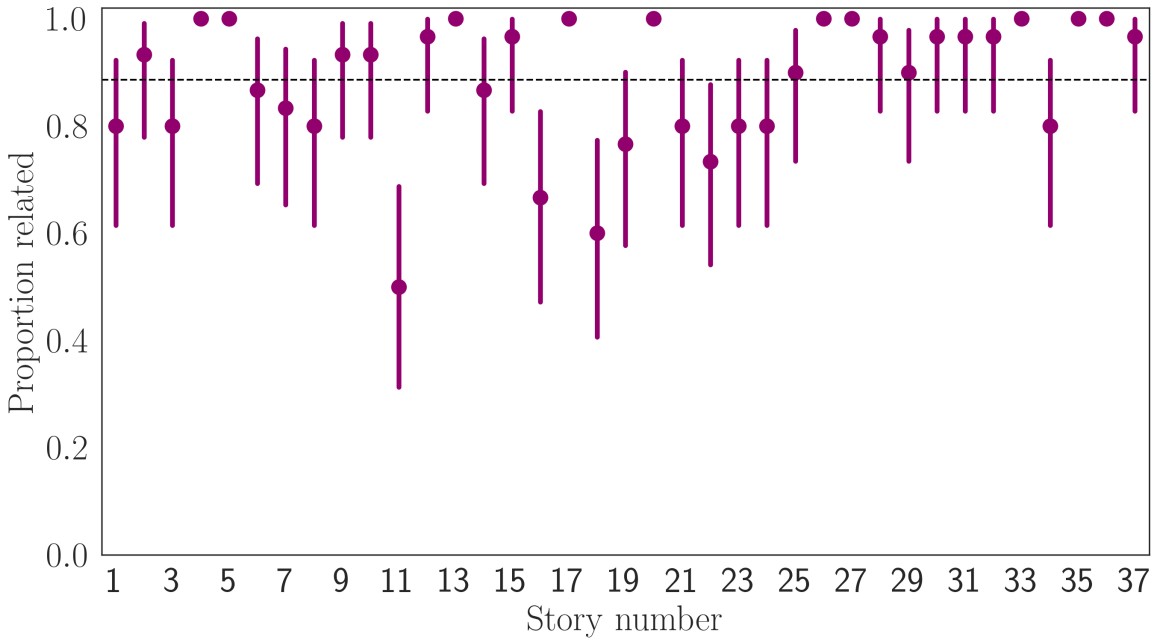

Figure 1: Proportion of related tweets for the considered stories. Error bars reflect 95% binomial proportion confidence intervals. The dashed black line corresponds to the overall success rate of 88.65%.

We observed a relatively low median proportion of contradicting tweets (6.67%) and a considerable number of stories not having any contradicting tweet ($N = 10$ [27.03%]).

Conclusion

In the present validation analysis, we examined a sample of the tweets collected via the ACR method and checked whether the tweets, in fact, correspond to the respective stories by manually annotating them. Consistent with our expectations, we observed that the ACR method collects related tweets with high precision: We found that 88.65% of the tweets were related to the respective stories, closely approximating the intended precision of .9.

We also examined whether the tweets were also supportive of the respective stories, as tweets may be related to a specific story, but may be in contradiction or neutral to the claim of the story. Among the tweets being related to the stories, we found that a majority of the tweets were also supportive (76.3%), indicating that the ACR method in fact collects tweets supporting the respective stories.

We also observed a significant amount of tweets being neutral (7.12%) or contradictive (16.58%) to the stories, and these proportions varied considerably from story to story (cf. Figure 2). The observed heterogeneity is consistent with other ACR validation analyses, indicating that the ACR method performs differently for each story. In this context, it is important to note that even human annotators occasionally fail to provide consistent judg-

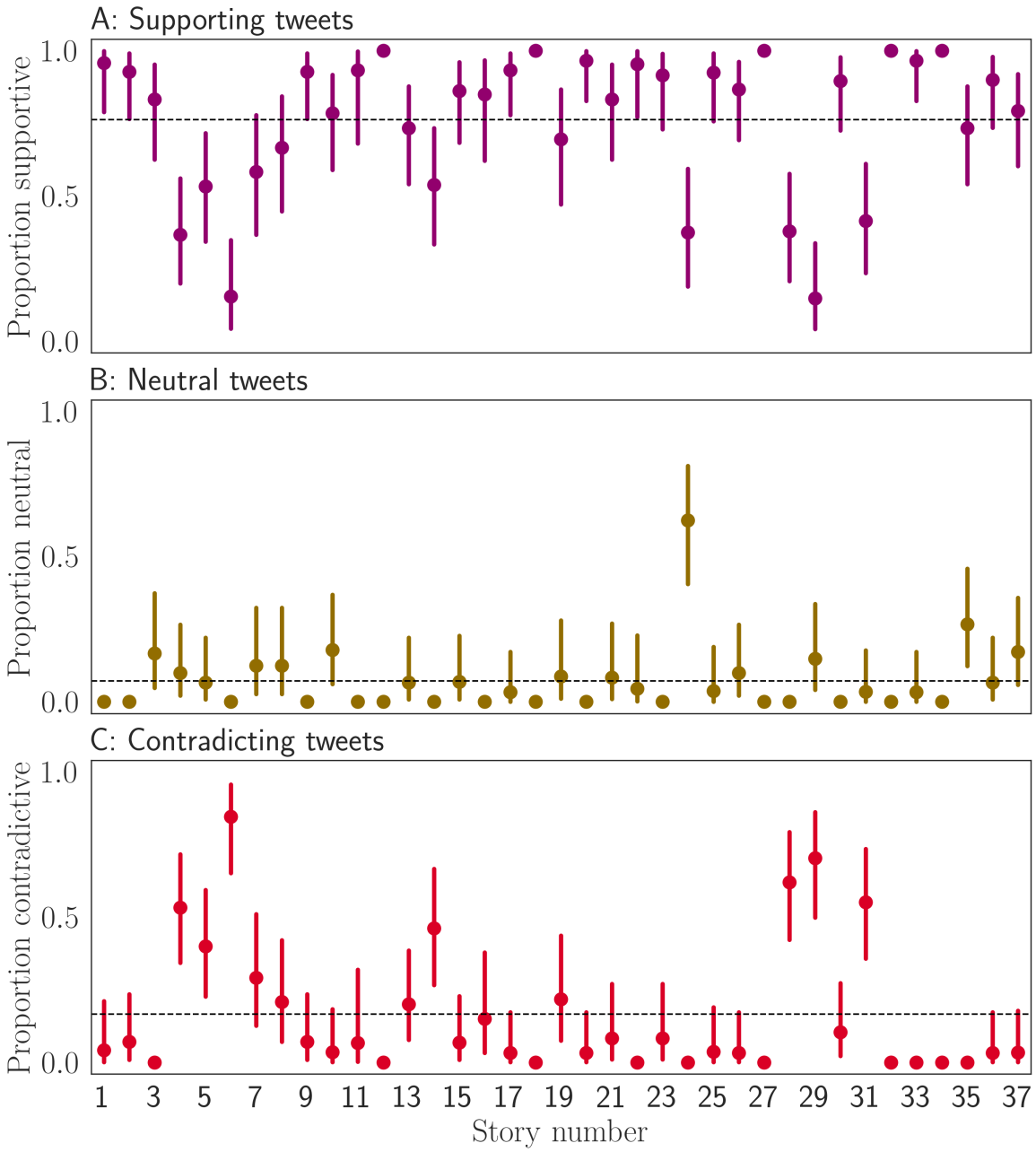

Figure 2: Proportion of supporting (Panel A), neutral (Panel B) and contradicting (Panel C) tweets for the considered stories. Error bars reflect 95% binomial proportion confidence interval. The dashed black line corresponds to the respective overall proportion of supporting, neutral and contradicting tweets.

ments when classifying tweets (see, e.g., [3]). We observed that the inter-rater reliability was acceptable for both tasks (Task I [Relatedness]: Cohen's $\kappa = .637$, Task II [NLI]: $\kappa = .537$), but far-away from a perfect agreement. We suspect that (i) the very fragmented syntactic and grammatical structure (see, e.g., [6, 1]), (ii) the limited length, (iii) irony/sarcasm and (iv) a missing context may complicate manual annotations of the tweets. Overall, these results suggest that, in cases in which optimal performance at the level of specific stories is a priority, it can be beneficial to combine automated ACR-based processing with manual checks of tweets to achieve the best outcome.

References

[1] Kim AE, Hansen HM, Murphy J, Richards AK, Duke J, Allen JA (2013) Methodological considerations in analyzing twitter data. Journal of the National Cancer Institute - Monographs 2013(47), DOI 10.1093/jncimonographs/lgt026

[2] Landis JR, Koch GG (1977) The Measurement of Observer Agreement for Categorical Data. Biometrics 33(1), DOI 10.2307/2529310

[3] Lendvai P, Augenstein I, Bontcheva K, Declerck T (2016) Monolingual social media datasets for detecting contradiction and entailment. In: Proceedings of the 10th International Conference on Language Resources and Evaluation, LREC 2016

[4] McHugh ML (2012) Interrater reliability: The kappa statistic. Biochemia Medica 22(3), DOI 10.11613/bm.2012.031

[5] Nie Y, Williams A, Dinan E, Bansal M, Weston J, Kiela D (2020) Adversarial NLI: A New Benchmark for Natural Language Understanding. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, pp 4885–4901, DOI 10.18653/v1/2020.acl-main.441

[6] Olteanu A, Castillo C, Diaz F, Kıcıman E (2019) Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries. Frontiers in Big Data 2, DOI 10.3389/fdata.2019.00013

[7] Stollenwerk F, Öhman J, Petrelli D, Wallerö E, Olsson F, Bengtsson C, Horndahl A, Gandler GZ (2023) Text Annotation Handbook: A Practical Guide for Machine Learning Projects. arXiv preprint arXiv:231011780

[8] Wallis S (2013) Binomial confidence intervals and contingency tests: Mathematical fundamentals and the evaluation of alternative methods. Journal of Quantitative Linguistics 20(3), DOI 10.1080/09296174.2013.799918