Additional file 8

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*

August 29, 2023

Evaluation of the ACR method for true stories

Please note that the following evaluation adheres to the analysis performed in the main document.

Descriptive statistics

Our dataset consisted of 2767 true stories that were either fact-checked by Snopes (1464 stories), PolitiFact (1290 stories), or both (13 stories). A majority of these stories ($N = 1795$) had a verdict score of 5, meaning that they were considered to be true, while the remaining stories had a verdict score of 4 (mostly true stories; $N = 972$). Compared to the false stories, we therefore had substantially fewer true stories in our dataset, possibly due to a tendency of fact-checking sites to check false stories.

Replies

Links to the respective fact-checking sites were found in $6.04 \cdot 10^4$ replies, with an average of 21.86 replies per story. We again found that the distribution of the number of replies was highly right-skewed, which was also expressed by the median Md = 2, as well as by the fact that 858 stories (31.01%) did not occur in any reply. This is also expressed in Fig. 1a, which reflects the log_{10}-transformed distribution of the number of replies per story.

Despite this transformation, the distribution remains substantially right-skewed. We again observed that (significant) initial fact-checking efforts on Twitter coincided with 2016 US presidential election, as shown in Fig. 1b.

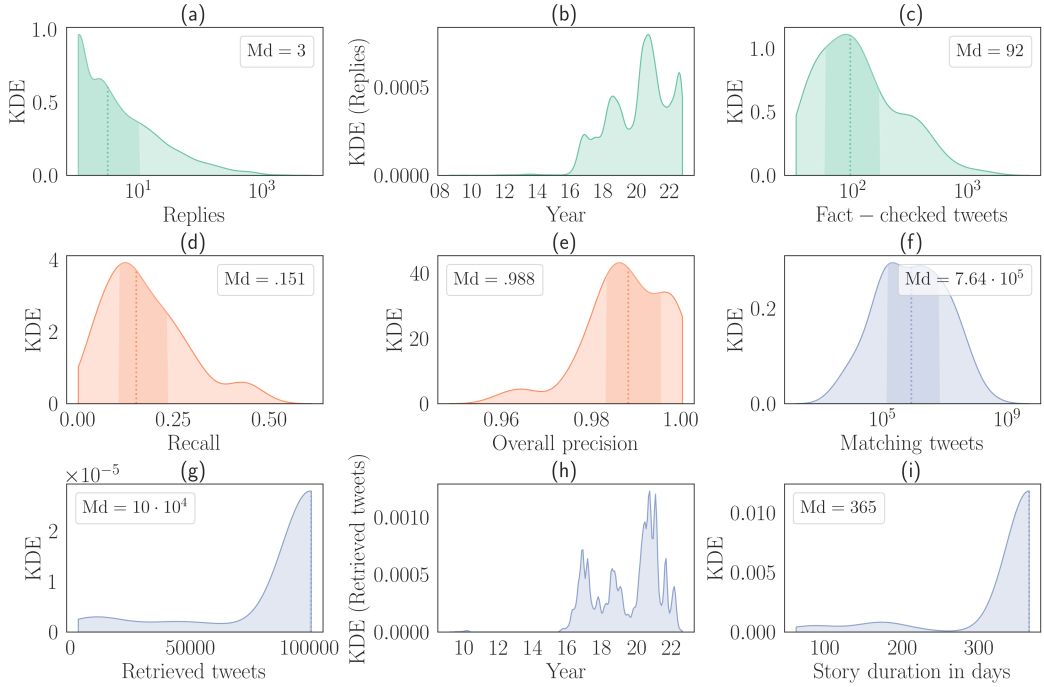

Figure 1: Relevant distributions of the dataset. Please note the following points: First, we performed kernel density estimations (KDEs; bandwidth selection according to Scott's rule) to approximate the underlying probability density functions. Second, the dashed line and the colored area reflect the median and interquartile range (IQR), respectively. Third, the full dataset, i.e., before story exclusion, was used for panels a and b. (a) KDE of the number of replies per story. Please note that the x-axis was $\log_{10}(x+1)$-transformed. (b) Time series (KDE-approximated) of all replies ($N = 6.04 \cdot 10^4$) (c) KDE of the number of fact-checked tweets per story. Please note that the x-axis was $\log_{10}(x+1)$-transformed. (d, e) KDEs of Recall (d) and Overall Precision (e). Please note that these measures are estimates. Density evaluation was restricted to $[0, 1]$. (f) KDE of the number of matching tweets. Please note that this data was accessed via the Tweet count endpoints and that the number of retrieved tweets for a given story might be lower because of down-sampling. (g) Time series (KDE-approximated) of all retrieved tweets ($N = 8.32 \cdot 10^6$). (h) KDE of the estimated story duration. Please note that density evaluation was also restricted because we limited the observation period to one year.

Story selection

In 50% of the cases, the ACR method identified a valid query. The exclusion process finally collected tweets of 96 stories (out of 2767), corresponding to a dropout rate of 96.53%. This low success rate of 3.47% again indicates that the ACR method is relatively strict in terms of data selection, and maybe even more strict for true stories. Please note that the upcoming sections only present results where the ACR method was *not* terminated.

Fact-checked tweets

In total, we retrieved $1.49 \cdot 10^4$ fact-checked tweets. The fact-checked tweets showed a similar pattern as the replies in terms of the distribution shape (see Fig. 1c): After $\log_{10}$-transformation, the distributions remained right-skewed, indicating that fact-checking of true stories is relatively rare on Twitter and maybe even more rare than for false stories.

Query selection and performance metrics

The average number of subqueries per story is 2.2 (max. 6), again indicating that our early stopping approach successfully restricted the number of queries. Concerning the number of available baseline periods, we found that most stories had three baselines, followed by four baselines (average: 3.23 baselines). Notably, baseline b_4 (post-story) was most likely missing due to time constraints, and we also observed that a b_4-dropout was relatively more prevalent than false stories.

We found that Recall followed a right-skewed distribution (see Fig. 1c) with an average and median of .175 and .151, respectively. This pattern of low average recall and positive skewness can again be explained by the relatively strict precision threshold(s), leading to recall loss according to the precision-recall tradeoff. However, we found that recall for true stories was reduced relative to that for false stories.

Precision (see Fig. 1d), on the other hand, again behaved as expected: Overall Precision was kept above 0.95, with an average and a median of .988 and .988, respectively.

Retrieved tweets

We retrieved $8.32 \cdot 10^6$ tweets belonging to 96 true stories. On average, we collected $8.67 \cdot 10^4$ tweets per story (median: $10 \cdot 10^4$, see Fig. 1g). We performed tweet sampling for 79 (82.29%) stories because the respective number of matching tweets exceeded the threshold of 10^5. The distribution of the number of matching tweets (see Fig. 1f) was again highly right-skewed.

With respect to the temporal features of our dataset, we again found that most stories emerged beginning from 2016 (see Fig. 1h) with a peak in 2020/21. In general, the time series again resembles the time series of the replies, as shown in Fig. 1b. The estimated duration of the stories (Fig. 1i) indicates that most were again estimated to last for one year or longer.

Main indicators of the ACR method

LMEM of mean text similarity across baselines and story period

Mean text similarity was significantly reduced for all baselines compared with the story period, as indicated by negative baseline coefficients (Fig. 2a), and an expected pattern of mean text similarity across baselines: While the pre-story baselines b_1, b_2, and b_3 showed the lowest text similarity, the text similarity of post-story baseline b_4 was again slightly less reduced. We again observed substantial effect heterogeneity across stories, with a minority of stories (48.96%) not showing the expected pattern of a similarity peak during the story period, potentially reflecting failures of the ACR method. Welch's t-tests confirmed for 49 (51.04%) stories that the similarity during the story period was significantly higher than that for all other baselines. However, the failure rate of the ACR method seems to be slightly enhanced for true stories compared with false stories.

Table 1: Fixed effects of the random effects model for mean text similarity.

					95% confidence interval	
Time period	Coefficient	SE	z	$p > \|z\|$	0.025	0.975
Intercept s	.43	.015	28.32	≈ 0	.401	.46
Baseline b_1	$-.088$	.016	-5.59	$1.13 \cdot 10^{-8}$	$-.118$	$-.057$
Baseline b_2	$-.098$	.014	-7.14	≈ 0	$-.125$	$-.071$
Baseline b_3	$-.081$	.012	-6.73	≈ 0	$-.105$	$-.058$
Baseline b_4	$-.079$	.014	-5.78	$3.68 \cdot 10^{-9}$	$-.105$	$-.052$

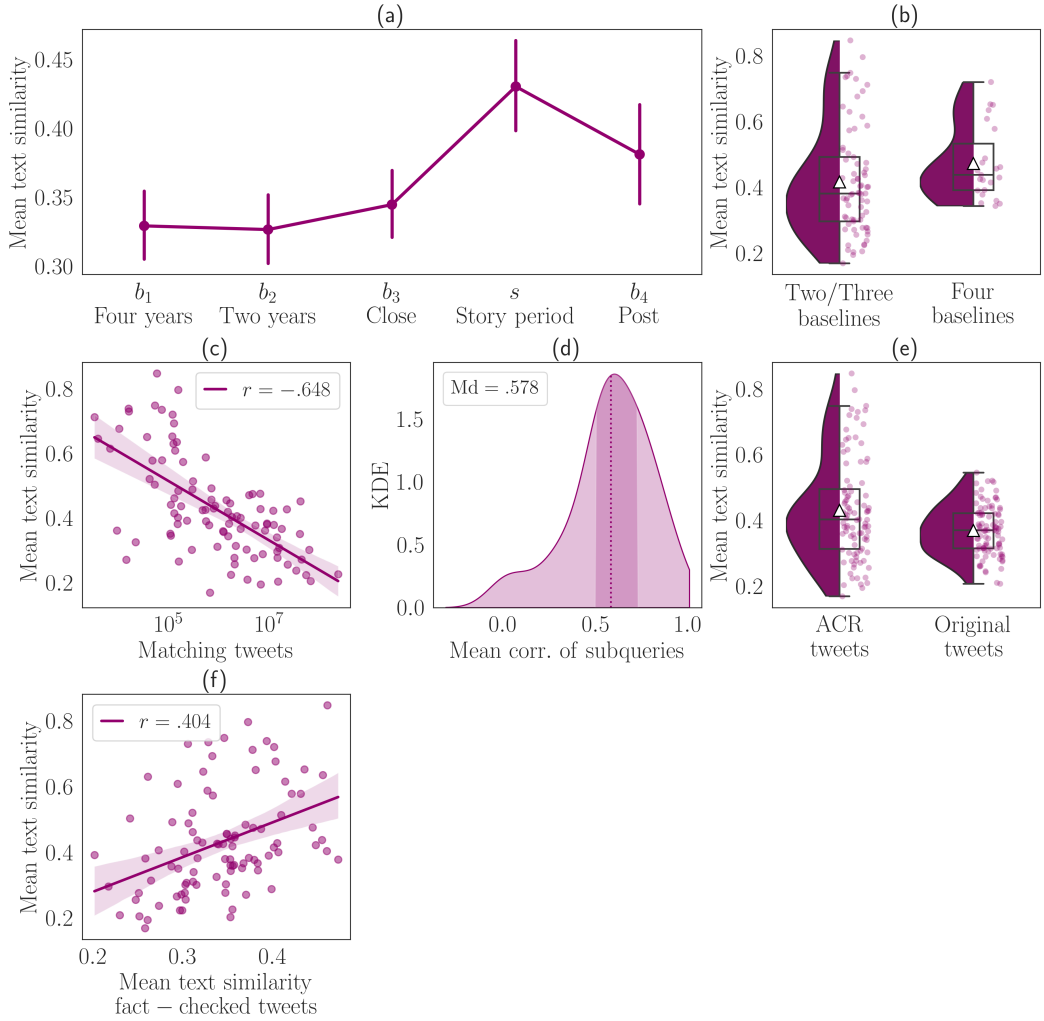

Figure 2: Main indicators of the ACR method. Please note that the regression lines reflect ordinary least squares regression, with 95% confidence interval (estimation via bootstrapping). Please also note that in panels (b) and (g), colored areas and white triangles reflect KDEs and means, respectively. (a) Mean text similarity across baselines and story period. Please note that error bars reflect 95% confidence intervals. The measure showed the expected pattern of a peak during the story period. (b) Mean text similarity was significantly higher for stories with four baselines, indicating that ACR methods is more reliable when more baselines are available. (c) A strong association between the number of matching tweets and mean text similarity is a signature of occasional tweet overload. (d) KDE of the mean correlation between subqueries. Most subqueries are highly correlated. (f) A text similarity comparison of the ACR-retrieved and original tweets indicates that ACR-retrieved may outperform the original tweets. (g) An association between the text similarity of the ACR-retrieved and fact-checked tweets indicates that the reliability of the ACR depends upon a good training set.

Again, text similarity during story periods was significantly higher for stories having four baselines (average: .473) compared to stories with two or three baselines (average: .417), as indicated by Fig. 2b and confirmed by Welch's t-test ($t(54.47) = 1.85, p = .035, \text{one} - \text{tailed}$).

ROC analysis of text similarity between baselines and story period

The ROC analysis again confirmed the substantial performance heterogeneity of the ACR method: While a mean AUC of .632 (SD = .172) indicates fair but not excellent classification performance, we observed good ($\geq .7$), very good ($\geq .8$) and even excellent ($\geq .9$) performance for 29 (31.87%), 18 (19.78%) and 9 (9.89%) stories, respectively. We observed ACR failures for 43 (47.25%) stories, as indicated by AUCs being equal to or smaller than .6, corresponding to poor or even uninformative classification. Slightly reduced AUCs for true stories relative to false stories again indicate that the ACR may perform slightly better for false stories.

We also observed a relatively strong association between AUC and mean text similarity during story periods ($r(89) = .583, p = 1.3 \cdot 10^{-9}, \text{two} - \text{tailed}$), indicating that better discrimination between tweets of the story and baselines periods is accompanied by an absolute increase in mean text similarity.

Tweet overload phenomenon

We again observed large numbers of matching tweets (46 (47.92%) stories exceeded 10^6), which could be related to at least two different factors. A substantial negative correlation between the number of matching tweets and text similarity ($r(94) = -.648, p \approx 0, \text{two} - \text{tailed}$; see Fig. 2c) again indicates that overload with false positives leads to reduced text similarity. This finding was again confirmed by a robust association between the number of matching tweets and AUC ($r(89) = -.381, p = 1.92 \cdot 10^{-4}, \text{two} - \text{tailed}$), showing that tweet overload also led to reduced classification performance. The correlation coefficients seem to be even more pronounced for true stories relative to false stories, indicating that overload with false positives was slightly more likely for true stories, supporting the impression that the ACR method may perform slightly worse for true stories.

Time series correlation between subqueries

Again, for most stories (62.5%), multiple subqueries were identified. On average, the time series of the subqueries were robustly correlated ($\overline{r} = .575, \text{SD} = .224$; see Fig. 2d), again indicating that these time series reflect the same underlying process. A control analysis accounting for tweet intersections between subqueries confirmed the robustness of the average correlation ($\overline{r} = .529, \text{SD} = .217$).

Unlike for false stories, we did *not* observe a correlation between mean text similarity and (mean) correlation of subqueries ($r(58) = .031, p = .407, \text{one} - \text{tailed}$), which might have occurred due to a lack of power, as we had substantially fewer true stories in our dataset.

Comparing text similarity between ACR tweets and original tweets

Again, we found that the ACR tweets showed higher mean text similarity (average: .43) relative to the original tweets (average: .37), as confirmed by a paired t-test ($t(95) = -4.26, p = 4.81 \cdot 10^{-5}, \text{two} - \text{tailed}$; see also Fig. 2e). We again also found that the 'quality' of the fact-checked tweets was somehow predictive for the ability of the ACR method to detect story-related tweets, as indicated by a moderate association between the mean text similarity of fact-checked and ACR tweets ($r(94) = .404, p = 4.51 \cdot 10^{-5}, \text{two} - \text{tailed}$; see also Fig. 2f).

Discussion

The evaluation of the ACR method for true stories confirmed that the ACR method is also a valid tool for fetching tweets that belong to true stories. Almost all indicators showed the expected patterns, but we observed a slight performance decrease relative to true stories. This might be attributable to a prevalent loss of the b_4 (post-story) baseline, which evidently increases the ACR method significantly.