Additional file 1

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

Model of the low prevalence problem on Twitter

For better understanding the low prevalence problem on Twitter, we consider a model of a false story with approximately 100k tweets during a period of one day with the claim 'Two explosions in the White House and Barack Obama is injured'[1]. This story has been debunked within minutes, but it nevertheless spread through Twitter and even induced a negative stack market response [1]. Please note that the numerical values were arbitrarily chosen for illustrative purposes.

However, we also assume that there will be 200 million irrelevant tweets at the same time. Thus, it holds that

$$\begin{aligned} |T| &= 10^5 \text{ and} \\ |Z| &= 2 \cdot 10^8. \end{aligned} \tag{1}$$

To evaluate the performance of the query, we consider $\text{Recall}(q_i)$, $\text{Precision}(q_i)$ and $\text{Precise}(q_i)$ (for their respective definitions, see the main text), as well as $\text{Specificity}(q_i)$ and $\text{Fall} - \text{out}(q_i)$, as given by

[1] see, e.g., https://eu.usatoday.com/story/theoval/2013/04/23/obama-carney-associated-press-hack-white-house/2106757/

$$
\begin{aligned}
\mathrm{Fall-out}(q_i) &= \frac{|\mathrm{FP}(q_i)|}{|Z|} \text{ and} \\
\mathrm{Specificity}(q_1) &= \frac{|\mathrm{TN}(q_i)|}{|Z|} \\
&= 1 - \mathrm{Fall-out}(q_i).
\end{aligned} \tag{2}
$$

$\mathrm{Fall-out}(q_i)$ is reflecting the proportion of irrelevant tweets that would be classified as story-supportive, while $\mathrm{Specificity}(q_i)$ stands for the fraction of irrelevant tweets that would be correctly classified as irrelevant.

However, when considering an exemplary query q_1 to detect tweets in T, let

$$
\begin{aligned}
\mathrm{Recall}(q_1) &= .8, \\
\mathrm{Fall-out}(q_1) &= 10^{-3} \text{ and} \\
\mathrm{Specificity}(q_1) &= 1 - \mathrm{Fall-out}(q_1) \\
&= .999.
\end{aligned} \tag{3}
$$

Prima facie, these measures can be interpreted as 'convincing': While 80% of the story-supporting tweets would be detected (recall), only one out of thousand irrelevant tweets would be misclassified as positive (fall-out) and, vise versa, 999 would be correctly classified as negative (specificity). However, because of the low prevalence of T, the number of false positive tweets yielded by this query would be

$$
\begin{aligned}
|\mathrm{FP}(q_1)| &= \mathrm{Fall-out}(q_1) \cdot |Z| \\
&= 10^{-3} \cdot 2 \cdot 10^8 \\
&= 2 \cdot 10^5,
\end{aligned} \tag{4}
$$

meaning that this query would fetch 200k false positive tweets. The number of true positive tweets, reflecting the number of tweets in T detected by the query, is given by

$$
\begin{aligned}
|\mathrm{TP}(q_1)| &= \mathrm{Recall}(q_1) \cdot |T| \\
&= .8 \cdot 10^5 \\
&= 8 \cdot 10^4,
\end{aligned} \tag{5}
$$

meaning that 80k story-supporting tweets would be detected by the query. In total, this query would therefore yield $N(q_1) = 280\text{k}$ tweets. The precision of q_1 is therefore given by

$$
\begin{aligned}
\mathrm{Precision}(q_1) &= \frac{|\mathrm{TP}(q_1)|}{N(q_1)} \\
&= \frac{8 \cdot 10^4}{2.8 \cdot 10^5} \\
&= .286,
\end{aligned} \tag{6}
$$

meaning that almost only one out of four of the matched tweets would truly support the story. The high number of false positive tweets yielded by this hypothetical query leads to an unfavorable signal-to-noise ratio, rendering (statistical) conclusions drawn from these data impossible. Thus, defining a lower bound of precision (as defined in the main document by 0.9) seems reasonable.

Consequently, we refrain from using q_1 for tweet retrieval, as $\text{Precision}(q_1) < 0.9$ (see also Table 1). Therefore, identifying a more specific query seems mandatory, but may lead to a loss of recall (specificity-recall tradeoff). Such a decrease in recall may be considered to be 'acceptable', as a story typically involves thousands of tweets, meaning that even a relatively small subset of them should be representative according to the law of large numbers.

Table 1: **Different queries aiming to collect tweets supporting the White House explosions story.**

	q_1: OBAMA	q_2: OBAMA WHITE HOUSE	q_3: OBAMA EXPLOSIONS	q_4: OBAMA INJURED EXPLOSIONS
$\|TP\|$	80k	70k	45k	35k
$\|FP\|$	200k	100k	2k	1k
N	280k	170k	47k	36k
Recall	.8	.7	.45	.35
$\text{Fall}-\text{out}$	10^{-3}	$.5 \cdot 10^{-3}$	10^{-5}	$.5 \cdot 10^{-5}$
Precision	.286	.412	.957	.972
Precise	0	0	1	1

Note: We set the number of supporting tweets arbitrarily to 100k and all the results shown were chosen arbitrarily for illustrative purposes.

However, increasing specificity in the context of tweet retrieval can be achieved by adding one or more conjuncts, i.e., keywords, to the query. Query q_2 was made more specific by adding the keywords 'white house'. This leads to a relatively small recall decrease ($\Delta = 0.1$), while fall-out halves. The query nevertheless remains relatively imprecise: Not even every second of the matched tweets is related to the story. The apparent problem with q_1 and q_2 is that they are not specific enough regarding the story of interest: These keywords may also occur in tweets related to other (irrelevant) stories, e.g., 'Obama announces tax rise during White House press briefing'.

This problem can be addressed by adding story-specific keywords, e.g., 'explosions'. Query q_3 shows significant recall loss relative to q_1 ($\Delta = 0.35$), but fall-out decreases disproportionally on the order of 10^2. Consequently, precision increased substantially to a value greater than 0.9, leaving q_3 as a favorable choice for tweet collection.

There may also be other queries exceeding the precision threshold. Query q_4 (which adds the keyword 'injured') is more precise compared to q_3, but it is also accompanied by a loss of recall ($\Delta = 0.1$) compared to q_3. However, according to our precision criterion, both queries q_3 and q_4 would be classified as appropriate for tweet collection.

References

[1] Vosoughi S, Roy D, Aral S (2018) The spread of true and false news online. Science 359(6380), DOI 10.1126/science.aap9559