Hostility on Twitter in the Aftermath of Terror Attacks

Christian S. Czymara (Tel Aviv University, Tel Aviv-Yafo, Israel & Goethe University Frankfurt, Frankfurt am Main, Germany),

Anastasia Gorodzeisky (Tel Aviv University, Tel Aviv-Yafo, Israel)

Published in *Journal of Computational Social Science*

https://doi.org/10.1007/s42001-024-00272-9

Supplementary Material

Validating ethnic hostility predictions

We used [1] to ask ChatGPT 3.5 to state the probability that a given Tweet in the sample is ethnically hostile (zero-shot classification). Similar to training human coders, we used the following instruction:

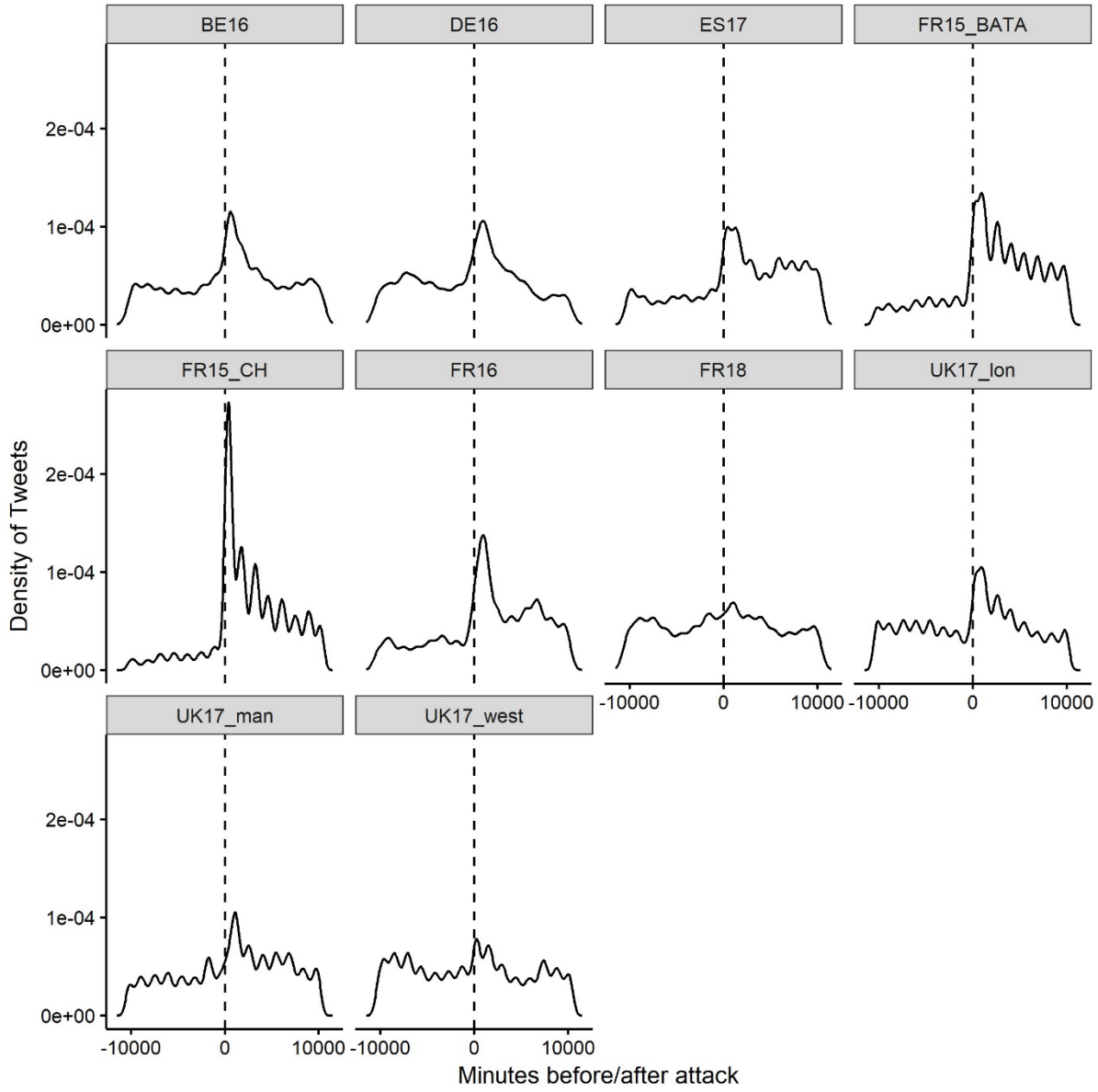
> *"Analyze the following Tweet and assign a probability (0-100) of it containing ethnic hostility, based on the language and tone. Ethnic hostility is any language that is derogatory, insulting, or offensive toward individuals or groups based on their ethnicity, race, or cultural background.*
>
> *Your classification should reflect how likely it is that this tweet contains ethnic hostility from 0 (definitely not hostile) to 100 (definitely hostile). Provide an integer response between 0 and 100 without further text.*
>
> *This tweet will be in [language]. Please classify its probability to be ethnically insulting:*
>
> *[Tweet]."*

We performed this task twice for each Tweet. We then checked the plausibility of the results, removed expressions such as "out of 100" or "/100", and only kept numeric characters up to 3 digits. The two judgments of ChatGPT correlated almost perfectly with each other (*r=0.96, ICC=0.96, F1 with 90% threshold=0.92*) and both correlated highly with the prediction of the Jigsaw's API (*r=0.8, ICC=0.77, F1 with 90% threshold=0.89 and 0.9*).

Reference

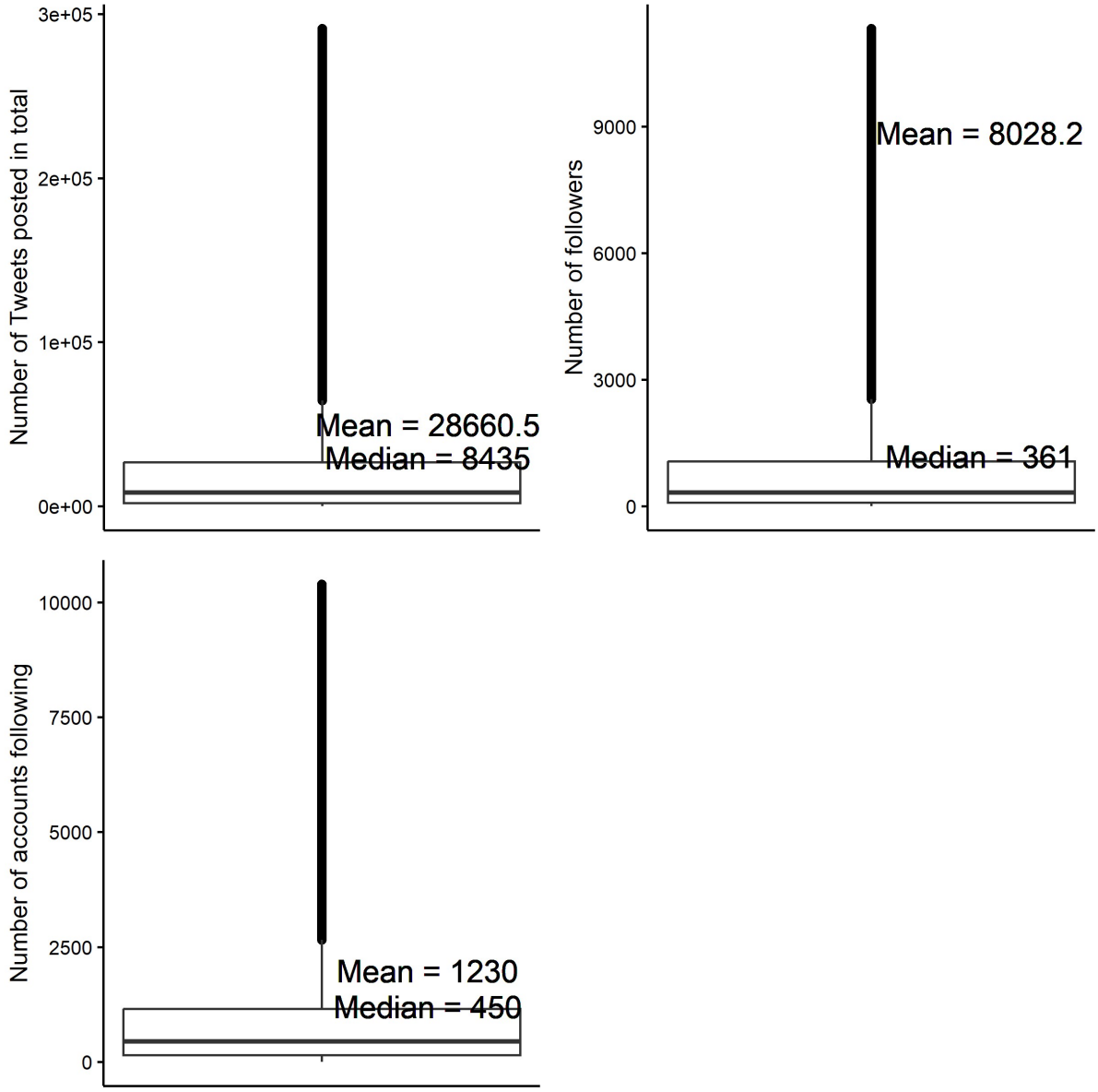
1. Rodriguez, J. C. (2023). *chatgpt: Interface to 'ChatGPT' from R* [Manual]. https://CRAN.R-project.org/package=chatgpt

Figure A1: Distribution of Tweets by case



Note: The figure displays the distribution of Tweets over time, with minutes before and after an incident on the x-axis and the Kernel density estimate of Tweets on the y-axis.

Figure A2: Distributions of sample characteristics



Note: The first panel refers to the Tweets an account has posted and kept on its timeline. The second panel shows the number of followers (i.e. incoming connections); the third panel shows the number of followings (i.e., outgoing connections). The total number of Tweets and the number of accounts following exclude the top 1 percent of the data and the number of followers the top 5 percent to enhance readability.

Table A1: Regression models

	M1: Mean comparison (full model)	**M2: Time trends (full model)**	**Varying trends (full model)**
(Intercept)	11.722 *** (11.429 – 12.015)	12.009 *** (11.694 – 12.325)	11.665 *** (10.888 – 12.442)
attack befaft01	5.414 *** (5.321 – 5.507)	9.877 *** (9.721 – 10.033)	8.057 *** (7.284 – 8.829)
case [FR15_BATA]	-0.542 *** (-0.774 – -0.310)	-1.020 *** (-1.247 – -0.794)	-2.881 *** (-3.528 – -2.234)
case [BE16]	-0.656 * (-1.163 – -0.149)	-0.917 *** (-1.423 – -0.412)	0.326 (-0.450 – 1.103)
case [FR16]	1.839 *** (1.377 – 2.301)	1.392 *** (0.927 – 1.857)	-0.851 (-1.829 – 0.128)
case [DE16]	43.443 *** (42.527 – 44.359)	42.817 *** (41.922 – 43.711)	40.718 *** (39.211 – 42.226)
case [UK17_west]	6.471 *** (6.143 – 6.799)	6.363 *** (6.023 – 6.702)	8.230 *** (7.414 – 9.047)
case [UK17_man]	7.898 *** (7.598 – 8.199)	7.665 *** (7.352 – 7.978)	9.648 *** (8.819 – 10.477)
case [UK17_lon]	9.925 *** (9.633 – 10.216)	9.394 *** (9.092 – 9.696)	7.117 *** (6.315 – 7.919)
case [ES17]	-0.790 *** (-1.142 – -0.438)	-0.851 *** (-1.208 – -0.495)	-3.830 *** (-4.668 – -2.993)
case [FR18]	-2.503 *** (-2.836 – -2.170)	-2.641 *** (-2.985 – -2.297)	-1.482 *** (-2.340 – -0.623)
attack diff minutes		0.000 (-0.000 – 0.000)	0.000 (-0.000 – 0.000)
attack diff minutes × attack befaft01		-0.001 *** (-0.001 – -0.001)	-0.001 *** (-0.001 – -0.000)
attack diff minutes × case [FR15_BATA]			-0.000 ** (-0.000 – -0.000)
attack diff minutes × case [BE16]			0.000 ** (0.000 – 0.000)

attack diff minutes × case [FR16]	-0.000 * (-0.000 – -0.000)
attack diff minutes × case [DE16]	-0.000 (-0.000 – 0.000)
attack diff minutes × case [UK17_west]	0.000 (-0.000 – 0.000)
attack diff minutes × case [UK17_man]	0.000 *** (0.000 – 0.000)
attack diff minutes × case [UK17_lon]	-0.000 *** (-0.000 – -0.000)
attack diff minutes × case [ES17]	-0.000 *** (-0.000 – -0.000)
attack diff minutes × case [FR18]	0.000 (-0.000 – 0.000)
attack befaft01 × case [FR15_BATA]	4.019 *** (3.346 – 4.691)
attack befaft01 × case [BE16]	-0.800 * (-1.576 – -0.025)
attack befaft01 × case [FR16]	4.080 *** (3.097 – 5.063)
attack befaft01 × case [DE16]	6.471 *** (4.940 – 8.003)
attack befaft01 × case [UK17_west]	-1.021 * (-1.824 – -0.218)
attack befaft01 × case [UK17_man]	0.312 (-0.506 – 1.131)

attack befaft01 × case [UK17_lon]	6.191 *** (5.384 – 6.998)
attack befaft01 × case [ES17]	7.244 *** (6.377 – 8.112)
attack befaft01 × case [FR18]	-1.143 * (-2.050 – -0.237)
(attack diff minutes × attack befaft01) × case [FR15_BATA]	-0.000 ** (-0.000 – -0.000)
(attack diff minutes × attack befaft01) × case [BE16]	-0.000 *** (-0.000 – -0.000)
(attack diff minutes × attack befaft01) × case [FR16]	-0.000 (-0.000 – 0.000)
(attack diff minutes × attack befaft01) × case [DE16]	-0.001 *** (-0.001 – -0.000)
(attack diff minutes × attack befaft01) × case [UK17_west]	-0.000 *** (-0.001 – -0.000)
(attack diff minutes × attack befaft01) × case [UK17_man]	-0.001 *** (-0.001 – -0.001)
(attack diff minutes × attack befaft01) ×	-0.000 *** (-0.001 – -0.000)

case [UK17_lon]			
(attack diff minutes × attack befaft01) × case [ES17]			-0.001 *** (-0.001 – -0.000)
(attack diff minutes × attack befaft01) × case [FR18]			-0.000 ** (-0.000 – -0.000)
Observations	4596300	4596300	4596300
AIC	39872345.708	39791994.051	39773090.265

** p<0.05 ** p<0.01 *** p<0.001*

Note: The first model compares the average probability that a Tweet is hostile before and after an attack (full version of model M1 in Table 1). The second model adds trends in the before and after period and estimates the effect immediately after an attack (full version of model M2 in Table 1). The third allows these trends and the attack effect to vary across cases (underlying Figure 2). The reference category for the case variable is the Charlie Hebdo attacks.

Table A2: Fixed effects interaction model

	FE mean difference varying by case	**FE varying trends by case**
attack befaft01 [1]	2.189 ***	3.017 ***
	(1.971 – 2.407)	(2.630 – 3.403)
case [FR15_BATA]	-0.580 ***	-0.883 ***
	(-0.846 – -0.313)	(-1.360 – -0.406)
case [BE16]	0.301 *	-0.378
	(0.031 – 0.571)	(-0.868 – 0.111)
case [FR16]	0.215	-0.542 *
	(-0.069 – 0.499)	(-1.057 – -0.026)
case [DE16]	41.391 ***	40.912 ***
	(40.815 – 41.967)	(40.138 – 41.687)
case [UK17_west]	6.089 ***	6.027 ***
	(5.781 – 6.396)	(5.594 – 6.459)
case [UK17_man]	6.525 ***	7.242 ***
	(6.216 – 6.833)	(6.809 – 7.675)
case [UK17_lon]	7.559 ***	5.711 ***
	(7.251 – 7.866)	(5.276 – 6.145)
case [ES17]	2.594 ***	1.683 ***
	(2.107 – 3.082)	(1.061 – 2.305)
case [FR18]	-2.937 ***	-3.309 ***
	(-3.226 – -2.649)	(-3.799 – -2.820)
attack befaft01 [1] × case [FR15_BATA]	0.722 ***	1.461 ***
	(0.432 – 1.013)	(0.936 – 1.986)
attack befaft01 [1] × case [BE16]	-0.226	0.677 *
	(-0.534 – 0.081)	(0.123 – 1.231)
attack befaft01 [1] × case [FR16]	1.185 ***	2.456 ***
	(0.869 – 1.502)	(1.878 – 3.034)
attack befaft01 [1] × case [DE16]	1.347 ***	2.722 ***
	(0.935 – 1.759)	(1.974 – 3.471)
attack befaft01 [1] × case [UK17_west]	-0.761 ***	-0.205
	(-0.987 – -0.535)	(-0.608 – 0.198)
attack befaft01 [1] × case [UK17_man]	-0.081	-0.567 **
	(-0.308 – 0.145)	(-0.970 – -0.163)

attack befaft01 [1] × case [UK17_lon]	-0.144 (-0.370 – 0.082)	2.511 *** (2.107 – 2.915)
attack befaft01 [1] × case [ES17]	-0.267 (-0.561 – 0.026)	0.828 ** (0.298 – 1.358)
attack befaft01 [1] × case [FR18]	0.230 (-0.087 – 0.548)	1.269 *** (0.693 – 1.844)
attack diff minutes		0.000 (-0.000 – 0.000)
attack befaft01 [1] × attack diff minutes		-0.000 *** (-0.000 – -0.000)
attack diff minutes × case [FR15_BATA]		-0.000 (-0.000 – 0.000)
attack diff minutes × case [BE16]		-0.000 ** (-0.000 – -0.000)
attack diff minutes × case [FR16]		-0.000 ** (-0.000 – -0.000)
attack diff minutes × case [DE16]		-0.000 (-0.000 – 0.000)
attack diff minutes × case [UK17_west]		-0.000 (-0.000 – 0.000)
attack diff minutes × case [UK17_man]		0.000 *** (0.000 – 0.000)
attack diff minutes × case [UK17_lon]		-0.000 *** (-0.000 – -0.000)
attack diff minutes × case [ES17]		-0.000 *** (-0.000 – -0.000)
attack diff minutes × case [FR18]		-0.000 (-0.000 – 0.000)
(attack befaft01 [1] × attack diff minutes) × case [FR15_BATA]		-0.000 (-0.000 – 0.000)
(attack befaft01 [1] × attack diff minutes) × case [BE16]		0.000 (-0.000 – 0.000)
(attack befaft01 [1] × attack diff minutes) × case [FR16]		-0.000 (-0.000 – 0.000)

(attack befaft01 [1] × attack diff minutes) × case [DE16]		-0.000 *** (-0.000 – -0.000)
(attack befaft01 [1] × attack diff minutes) × case [UK17_west]		-0.000 *** (-0.000 – -0.000)
(attack befaft01 [1] × attack diff minutes) × case [UK17_man]		-0.000 *** (-0.000 – -0.000)
(attack befaft01 [1] × attack diff minutes) × case [UK17_lon]		0.000 (-0.000 – 0.000)
(attack befaft01 [1] × attack diff minutes) × case [ES17]		0.000 * (0.000 – 0.000)
(attack befaft01 [1] × attack diff minutes) × case [FR18]		-0.000 * (-0.000 – -0.000)
Observations	4596300	4596300
AIC	35772054.884	35756310.662

** p<0.05 ** p<0.01 *** p<0.001*

Note: The first model estimates a fixed effects model for the average difference in the probability that a Tweet is hostile before and after an attack for each attack separately by interacting the attack dummy with the case variable. The second model allows the trends before and after to vary by case by including a three-way interaction.

Table A3: Logistic regression models with binary outcome

	P(hostility) >= 75 percent	**P(hostility) >= 90 percent**
(Intercept)	0.001 *** (0.000 – 0.001)	0.000 (0.000 – 0.000)
attack diff minutes	1.000 (1.000 – 1.000)	1.000 (1.000 – 1.000)
attack befaft01	2.901 *** (2.741 – 3.070)	2.256 *** (2.026 – 2.512)
attack diff minutes × attack befaft01	1.000 *** (1.000 – 1.000)	1.000 *** (1.000 – 1.000)
Case dummies	✓	✓
Observations	4596300	4596300
AIC	187500.002	34029.315

** p<0.05 ** p<0.01 *** p<0.001*

Note: Models estimate the odds that a Tweet is at least 75% or 90% hostile, respectively, using logistic regression.

Table A4: Model excluding highly active users

	Without highly active users
(Intercept)	12.073 *** (11.867 – 12.279)
attack diff minutes	0.000 (-0.000 – 0.000)
attack befaft01	9.910 *** (9.779 – 10.040)
attack diff minutes × attack befaft01	-0.001 *** (-0.001 – -0.001)
Case dummies	✓
Observations	4280656
AIC	37119822.356

** p<0.05 ** p<0.01 *** p<0.001*

Note: This model replicates M2 of Table 1 excluding Tweets from highly active users. Highly active users here are the top one percent that Tweeted the most overall.

Table A5: Placebo tests

	Only before period	**Randomly timed treatment**
(Intercept)	10.069 *** (9.195 – 10.942)	18.283 *** (18.042 – 18.524)
attack diff minutes	-0.000 *** (-0.000 – -0.000)	0.001 *** (0.001 – 0.001)
placebo	1.422 *** (1.080 – 1.763)	-1.261 *** (-1.420 – -1.101)
attack diff minutes × placebo	0.000 *** (0.000 – 0.000)	-0.001 *** (-0.001 – -0.001)
Case dummies	✓	✓
Observations	1943095	4596300
AIC	16736772.890	39864836.454

** p<0.05 ** p<0.01 *** p<0.001*

Note: The models constitute two placebo tests. The first model of this table only examines the before period by splitting it in half. The second model of this table tests for a randomly timed treatment variable.