A Report on "Hate in the Machine: Anti-Black and Anti-Muslim Social Media Posts as Predictors of Offline Racially and Religiously Aggravated Crime" by Williams et al. (2020)

Reviewer 2

February 11, 2026

v2

isitcredible.com

Disclaimer

This report was generated by large language models, overseen by a human editor. It represents the honest opinion of The Catalogue of Errors Ltd, but its accuracy should be verified by a qualified expert. Comments can be made here. Any errors in the report will be corrected in future revisions.

> I am wiser than this person; for it is likely that neither of us knows anything fine and good, but he thinks he knows something when he does not know it, whereas I, just as I do not know, do not think I know, either. I seem, then, to be wiser than him in this small way, at least: that what I do not know, I do not think I know, either.
>
> Plato, *The Apology of Socrates*, 21d

> To err is human. All human knowledge is fallible and therefore uncertain. It follows that we must distinguish sharply between truth and certainty. That to err is human means not only that we must constantly struggle against error, but also that, even when we have taken the greatest care, we cannot be completely certain that we have not made a mistake.
>
> Karl Popper, 'Knowledge and the Shaping of Reality'

Overview

Citation: Williams, M. L., Burnap, P., Javed, A., Liu, H., & Ozalp, S. (2020). Hate in the Machine: Anti-Black and Anti-Muslim Social Media Posts as Predictors of Offline Racially and Religiously Aggravated Crime. *British Journal of Criminology*, Vol. 60, No. 1, pp. 93–117.

URL: https://academic.oup.com/bjc/article/60/1/93/5537169

Abstract Summary: This article examines the association between online hate speech targeting race and religion and offline racially and religiously aggravated crimes in London over an eight-month period using computational criminology methods. The findings establish a general temporal and spatial association, renewing the understanding of hate crime as a process for the digital age.

Key Methodology: Longitudinal ecological analysis using police recorded crime, census data, and geo-coded Twitter data linked by LSOA and month, analyzed using Negative Binomial Regression and Random/Fixed-Effects Poisson panel models.

Research Question: Does an association exist between online hate speech targeting race and religion and offline racially and religiously aggravated crimes, independent of 'trigger' events?

Editor's Note

Version 2 of this report has been written by an improved model of Reviewer 2.

Summary

Is It Credible?

This article sets out to establish a link between online hate speech and offline racially and religiously aggravated crime. Williams et al. argue that their models demonstrate a "consistent positive association" between the two, suggesting that online hate speech acts as a "signature" of inter-group tension that can predict offline victimization (pp. 111–112). While the study successfully highlights a correlation between the volume of hateful tweets and recorded hate crimes within London neighborhoods—a novel contribution that challenges the "digital dualism" often found in criminology—the credibility of the authors' specific claims regarding the magnitude, causality, and predictive utility of this relationship is significantly undermined by fundamental errors in interpretation and design limitations.

The most immediate concern regarding the study's conclusions is a significant inconsistency in the interpretation of the statistical models. The authors report an Incidence Rate Ratio (IRR) of 1.004 for harassment and interpret this to mean that "an increase of 100 hate tweets would correspond to a 0.4 per cent increase" in offline crime (p. 108). This interpretation is mathematically incoherent with the data presented in Table 2, which lists an IRR of approximately 1.0035. In a Poisson or negative binomial model, an IRR represents a multiplicative effect. If the IRR is indeed ~1.004, the correct calculation for a 100-unit increase is derived by raising the IRR to the power of the increase (1.004^{100}), which results in an increase of approximately 49 percent, not 0.4 percent. Conversely, if the authors' claim of a 0.4 percent increase is correct, the IRR would need to be infinitesimal (approx. 1.00004), contradicting their own table. This internal contradiction suggests the authors have confused the marginal effect of a single tweet with the cumulative effect of larger volumes, or simply misreported their coefficients. Consequently, the article's discussion of the

"practical significance" of the findings is based on a confused interpretation of the model's output, making it difficult to assess the true strength of the relationship.

Beyond the magnitude of the effect, the study's claim to establish temporal precedence is methodologically unsound. The authors assert that their panel models "allow us to determine if online hate speech precedes rather than follows offline hate crime" (p. 108). However, the regression models aggregate data by month. This temporal resolution is too coarse to determine directionality. A "trigger event" occurring in the middle of a month could generate both hate tweets and hate crimes simultaneously within that same month. While the authors do utilize daily data to visualize the "half-life" of hate speech in their descriptive analysis (p. 96), this granularity is lost in the primary inferential models. Without finer-grained temporal data (e.g., daily or weekly) or lagged variables in the regression analysis, the model cannot distinguish whether the tweets predicted the crimes, the crimes inspired the tweets, or—most likely—both were reactions to external stimuli. The claim that the association exists "independent of 'trigger' events" (p. 93) is therefore not fully supported, as the monthly aggregation obscures the immediate sequencing of events.

Finally, the study faces a conceptual challenge regarding the spatial nature of social media. The analysis assumes that tweets geotagged within a specific Lower Layer Super Output Area (LSOA) reflect the "community tensions" of that specific neighborhood (p. 101). This risks an ecological fallacy by conflating the location of a tweet's creation with the residence of the user or the location of the victim. Social media influence is non-local; a user in one borough may be radicalized by content produced globally, not just by tweets sent from their immediate physical vicinity. While the authors attempt to control for this by removing outliers like Heathrow Airport, the fundamental mismatch between the borderless nature of online "echo chambers" and the rigid administrative boundaries of LSOAs remains. The study effectively shows that areas with high Twitter activity and high crime rates overlap, but the inference that local tweets are a valid proxy for the specific local tensions driving those

crimes is tenuous.

The Bottom Line

The study provides evidence of a correlation between the volume of geotagged hate speech and police-recorded hate crime in London, confirming that these phenomena often co-occur in the same geographic areas. However, the headline claims regarding the predictive power and causal direction of online hate speech are not credible. The interpretation of the findings is compromised by critical internal inconsistencies regarding effect sizes, and the monthly aggregation of data prevents the establishment of temporal precedence.

Potential Issues

Internal inconsistency in effect size interpretation: The article contains a significant contradiction between its statistical tables and its textual interpretation of the findings. The authors interpret the Incidence Rate Ratio (IRR) from their count models as if it were linear or vastly smaller than reported. For example, for harassment, they report an IRR of ~1.004 in Table 2 but state in the text that "an increase of 100 hate tweets would correspond to a 0.4 per cent increase" in crime (p. 108). This is mathematically incorrect based on the reported IRR. An IRR is a multiplicative factor. The correct calculation for a 100-unit increase in the predictor is $(IRR^{100}) - 1$. In this case, $(1.004^{100}) - 1 \approx 0.49$, which corresponds to a 49 percent increase in the rate of harassment, not 0.4 percent. Similarly, their claim that a 1,000-tweet increase corresponds to a 4 percent increase is also incorrect; the actual implied effect is $(1.004^{1000}) - 1 \approx 53.58$, or a 5,358 percent increase. This error suggests the authors have either misreported the coefficient in the table or fundamentally misunderstood the non-linear nature of their model, invalidating the article's discussion of practical significance.

Contradictory causal claims and unsubstantiated assertion of temporal precedence: The article makes conflicting statements regarding its ability to draw causal conclusions. The authors correctly state that making strong causal claims "would stretch the data beyond their limits" due to the ecological study design (p. 107). However, they also claim that their panel models "allow us to determine if online hate speech precedes rather than follows offline hate crime" (p. 108), a statement that implies the establishment of temporal precedence. This claim is not supported by the methodology, which aggregates and links hate tweets and hate crimes within the same month. While the authors use daily data for descriptive graphs, the inferential models use monthly aggregates. Using contemporaneous variables in a fixed-effects model cannot resolve the issue of simultaneity or reverse causation,

such as an offline crime event in the first week of a month triggering online hate speech later in the same month.

Ecological fallacy and mismatch between data and theory: The study's design is ecological, analyzing correlations between aggregate data at the Lower Layer Super Output Area (LSOA) level. However, the theoretical framework and conclusions are largely based on individual-level psychological and behavioral processes, such as "political polarization," "echo chambers," and how individuals are "influenced by social media communications" (pp. 97–98). The analysis can show that geographic areas with more hate tweets also tend to have more hate crimes, but it cannot establish that the individuals posting or being exposed to the online hate are in any way connected to the individuals perpetrating the offline crimes. This represents a risk of committing the ecological fallacy by drawing inferences about individual behavior from group-level data.

Spatial mismatch of causal mechanism and measurement unit: The study's design appears to be misaligned with the nature of social media by assuming that the production and impact of online hate speech are spatially contained within small administrative boundaries (LSOAs). The analysis correlates tweets geotagged within an LSOA with crimes occurring in the same LSOA. However, the influence of online content is not constrained by geography; a user in one LSOA can be influenced by content created anywhere in the world. The study measures where hate speech is *produced*, not where it is *consumed*. The authors attempt to address this by framing the production of hate tweets as a proxy for local "inter-group racial and/or religious tension" (p. 101), but this operationalization is still limited. The geotagged location of a tweet may not represent the user's community of residence, creating a potential spatial mismatch between the measured online speech and the local community where offline crimes are recorded.

Interpretation of practical significance based on unrealistic scenarios: The article's discussion of the practical importance of its findings may be misleading because it is

anchored to scenarios that are extreme outliers in the data. The authors interpret the effect size based on a hypothetical increase of 100 or 1,000 hate tweets within an LSOA in a single month (p. 108). However, the study's own descriptive statistics show that the average LSOA-month had a mean of 8 hate tweets with a standard deviation of 15.8 (p. 101). An increase of 100 tweets represents an event more than six standard deviations above the mean. Basing the general interpretation of the effect size on such rare occurrences may inflate the perceived importance of the relationship.

Findings potentially driven by outlier removal: The robustness of the study's conclusions is uncertain due to the influential effect of removing a small number of outliers. The authors report removing four "influential points" (outliers), stating that their inclusion "did change the magnitude of effects, standard errors and significance levels for some variables and model fit" (p. 104). While they describe the qualitative impact, the article does not present the full results of the analysis with the outliers included. This omission prevents readers from assessing the fragility of the findings and understanding how sensitive the reported statistical associations are to the presence of these few, albeit atypical, areas.

Unexplained sample size discrepancy: There is a notable discrepancy in the reported sample size. The methods section states the study includes 4,720 LSOAs (p. 101), but the first regression table reports a sample size of N = 4,270 (p. 107). This represents a loss of 450 units (approximately 9.5 percent of the sample) that is not explicitly accounted for in the text. While this may be due to missing census data or zero-population areas, the lack of explanation for such a significant drop in sample size raises questions about data completeness.

Ambiguity in model specification: The article provides a contradictory description of its fixed-effects models. In one section, the authors state that the models are based on "within-borough variation" (p. 102), while in another section they state they are based on "within-LSOA variation" (p. 108). These are distinct specifications—one controlling for borough-level heterogeneity and the other for LSOA-level heterogene-

ity. This ambiguity makes it difficult to precisely understand the controls applied in the final analysis.

Omission of plausible time-varying confounders: The study's models do not control for observable, time-varying local factors that could create a spurious correlation between online hate speech and offline hate crime. For instance, an offline event such as a far-right march or a leafleting campaign within an LSOA could plausibly increase both the perpetration of offline hate crimes and the volume of online hate speech from local individuals. In such a scenario, the online speech would be a symptom of offline organizing rather than an independent predictor.

Conceptual mismatch between measured construct and theoretical concept: There is a notable gap between the theoretical concept of "hate speech" and how it was operationalized. The machine learning classifier was trained to identify text that is "offensive or antagonistic in terms of race, ethnicity or religion" (p. 101). The authors are transparent about this, stating that their measure is not designed to correspond to the legal definition of a hate crime but is instead a measure of "online inter-group racial and/or religious tension" (p. 101). While this is a reasonable operational choice, it means the study is correlating a broad measure of online "tension" with a narrow, legally defined category of offline crime.

Methodological transparency issues: Several aspects of the study's methodology lack the detail required for full critical evaluation or replication. First, the description of the machine learning classifier is high-level; it omits crucial details about the feature engineering process and the sampling strategy for selecting the 2,000 tweets used for training the model (p. 101). Second, the study uses data collected between August 2013 and August 2014, and while the authors acknowledge this was before major changes in platform policies (p. 114), the six-year gap between data collection and publication may limit the external validity of the findings.

Future Research

Correcting statistical interpretation: Future work must accurately calculate the cumulative effect of count variables in non-linear models. Researchers should re-evaluate the magnitude of the association between hate speech and crime using the correct exponential transformation of Incidence Rate Ratios to determine if the implied effect sizes remain plausible or if they suggest omitted variable bias.

High-frequency temporal analysis: To genuinely test whether online hate precedes offline crime, future studies should utilize daily or weekly data rather than monthly aggregates in their regression models. By employing Granger causality tests or similar time-series methods with lagged variables, researchers could determine if spikes in online hate speech provide a leading indicator for offline violence or if they occur simultaneously.

Network-based spatial analysis: Future research should move beyond administrative geographic boundaries (like LSOAs) which may not reflect online social dynamics. Investigations could employ social network analysis to map the "location" of hate speech based on user connectivity and interaction rather than physical geotags, thereby testing the influence of non-local digital communities on local offender behavior.

isitcredible.com