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 05, 2026

v1

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?

Summary

Is It Credible?

This article by Williams et al. presents a study within the field of "Computational Criminology," aiming to establish a link between online hate speech and offline hate crime in London. By combining police records, census data, and geotagged Twitter data over an eight-month period, the authors argue that there is a "temporal and spatial association" between the two phenomena that exists independently of major "trigger" events like terror attacks. The central claim is that these findings renew the understanding of hate crime "as a process, rather than as a discrete event," suggesting that online hostility "migrates" to the physical world. The authors further contend that including social media data significantly improves the predictive power of statistical models compared to using census demographics alone.

The credibility of the article's most specific empirical claims is severely compromised by a fundamental mathematical error in the interpretation of the statistical results. In discussing the magnitude of the relationship between online hate speech and offline crime, the authors misinterpret the Incidence Rate Ratio (IRR) derived from their Poisson regression models. For the offense of harassment, they report an IRR of 1.004 and incorrectly state that "an increase of 100 hate tweets would correspond to a 0.4 per cent increase" in the outcome (p. 108). This calculation treats the multiplicative IRR as if it were a linear percentage point addition. The correct calculation for a 100-unit increase with an IRR of 1.004 is a multiplicative increase of approximately 49 percent. Similarly, their claim that 1,000 tweets would yield a 4 percent increase is mathematically impossible under their model; the actual implied increase would be over 5,000 percent. This error indicates that the authors have either drastically underestimated the effect size implied by their own model or that the model itself is producing implausibly large coefficients that went unnoticed due to the calculation

error. Consequently, the article's discussion regarding the practical significance of the findings is unreliable.

Beyond the statistical interpretation, the study's causal and theoretical framing overreaches the limits of the research design. The authors claim their panel models "allow us to determine if online hate speech precedes rather than follows offline hate crime" (p. 108). However, the analysis aggregates data into monthly units. It is methodologically impossible to establish temporal precedence within a concurrent month; an offline crime occurring on the first of the month could easily trigger online hate speech on the thirtieth. The assertion that hate "migrates" from the online to the offline sphere implies a directional causality that the correlational, contemporaneous data cannot support. Furthermore, the study relies on an ecological design that risks a spatial mismatch between theory and measurement. The theoretical framework draws on concepts of online polarization and echo chambers—phenomena that are not geographically bounded—yet the analysis assumes that hate speech produced in a specific London neighborhood is the driver of hate crime in that same neighborhood. This ignores the reality that social media influence is non-local; a perpetrator in London could be radicalized by content produced in New York, which this study's methodology would fail to capture.

Despite these significant issues, the article does contribute to the literature by validating conventional ecological predictors of hate crime. The analysis confirms that factors such as long-term unemployment and the proportion of the population that is Black, Asian, and Minority Ethnic (BAME) are statistically significant predictors of offline hate crime, aligning with prior sociological research. The study also demonstrates that there is a statistical signal connecting the volume of geotagged hate speech to local crime rates, even if the magnitude and causal direction of that signal are misinterpreted. The work represents an ambitious attempt to integrate novel data sources into criminology, but the "predictive" utility of the social media data is overstated given the reliance on simultaneous monthly

data and the profound errors in interpreting the model's coefficients.

The Bottom Line

Williams et al. successfully demonstrate a statistical correlation between the volume of local online hate speech and recorded hate crimes in London, validating traditional demographic predictors in the process. However, the credibility of the study's specific conclusions is undermined by a critical mathematical error that misrepresents the magnitude of the effect, as well as causal claims regarding the "migration" of hate that cannot be substantiated by the monthly aggregate data. Readers should accept the general association but view the specific claims about predictive power and effect size with extreme skepticism.

Potential Issues

Incorrect calculation of practical effect size: The article contains a significant mathematical error in its interpretation of the main finding's magnitude, which leads to a substantial misrepresentation of the practical effect. The authors interpret the Incidence Rate Ratio (IRR) from their Poisson models as if it were linear. For example, for harassment, they report an IRR of 1.004 and state that "an increase of 100 hate tweets would correspond to a 0.4 per cent increase, and an increase of 1,000 tweets would correspond to a 4 per cent increase in racially or religiously aggravated harassment in a given month within a given LSOA" (p. 108). This is incorrect. An IRR is a multiplicative factor on the rate of the outcome. 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% increase in the rate of harassment, not 0.4%. Similarly, their claim that a 1,000-tweet increase corresponds to a 4% increase is also incorrect; the actual implied effect is $(1.004^{1000}) - 1 \approx 53.58$, or a 5,358% increase. This error appears to stem from confusing the percentage increase for a single tweet (approximately 0.4%) with the cumulative effect of 100 or 1,000 tweets. This miscalculation invalidates the article's discussion of the magnitude and practical significance of its findings.

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 "These models therefore 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, a key condition for causality. This claim is not supported by the methodology, which aggregates and links hate tweets and hate crimes within the same month. Using contempo-

raneous variables in a fixed-effects model cannot resolve the issue of simultaneity or reverse causation (e.g., an offline crime event in the first week of a month could trigger online hate speech later in the same month). The article's abstract and discussion frequently use language that implies a directional, causal process, such as concluding that online hate can "migrate to the physical world" (p. 114), which overstates what the correlational, contemporaneous design can demonstrate.

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. The authors acknowledge this limitation late in the article, stating that "the individual level mechanisms responsible for the link between online and offline hate incidents remain to be established by more forensic and possibly qualitative work" (p. 114), but the narrative throughout the article strongly implies an individual-level causal story that the data cannot directly support.

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*. While the authors frame the production of hate tweets as a proxy for local "inter-group racial and/or religious tension" (p. 101), this operationalization is still limited. The geotagged location of a tweet may not represent the user's community of residence, as people tweet from work, commercial centers, or transit hubs. While the authors remove influential outliers like Heathrow Airport (p. 104), this is presented as a statistical procedure to handle influential data points rather than an explicit strategy to address the theoretical problem of non-local influence, which remains a limitation of the design.

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 involving large increases in hate tweets. 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 five standard deviations above the mean. While the authors argue such spikes are "not fanciful" in the context of "trigger events" (p. 111), and the maximum observed value was 522, basing the general interpretation of the effect size on such rare occurrences may inflate the perceived importance of the relationship. This issue is compounded by the mathematical error in calculating the effect size, which together create a distorted picture of the findings' practical significance.

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), which represent less than 0.1% of the 4,720 LSOAs, stating that their inclusion "did change the magnitude of effects, standard errors and significance levels for some variables and model fit" (p. 104). While removing influential outliers can be a valid methodological step, 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.

High measurement error from the hate speech classifier: The study's key independent variable, the count of hate tweets, is measured with a substantial degree of error. The machine learning classifier used to identify hate speech had a "retrieval" (recall) of 0.69, meaning it failed to identify 31% of the tweets that human coders had labeled as hateful (p. 101). This high false-negative rate indicates that the "Hate Tweets" variable is a systematic undercount of the phenomenon. While measurement error in an independent variable often biases coefficients toward zero, making the estimates conservative, the article does not fully discuss the potential impact of this significant error on the regression results. The authors do acknowledge that "algorithmic classification of hate speech is not perfect" (p. 113), but they do not connect this to potential bias in the estimated coefficients. This introduces a degree of uncertainty into the precision of the reported coefficients and the validity of the statistical associations.

Unaddressed selection biases and mischaracterization of data limitations: The article claims its social media data mitigates certain biases by not being "produced by the police," meaning it is "immune from inherent biases normally present in the official data generation process" (p. 113). While true that the data avoids police recording bias, this claim may be misleading as it overlooks the severe and different biases inherent in the social media data itself. The study relies on a small, nonrandom subset of the population: Twitter users who choose to geotag their posts. This group is subject to significant demographic and behavioral self-selection biases that are not fully addressed. The authors do acknowledge some of these biases elsewhere (p. 114), but the claim of "immunity" overstates the objectivity of the data source.

Omission of plausible time-varying confounders: The study's models do not con-

trol 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. The authors acknowledge the limitation that they were unable to observe "sub-LSOA factors" (p. 107), but the absence of controls for such potential confounders limits the confidence that can be placed in the estimated relationship.

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 "Ours is a measure of online inter-group racial and/ or religious tension, akin to offline community tensions that are routinely picked up by neighborhood policing teams" (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. This conceptual slippage weakens the claim of a direct link between like-for-like phenomena (online hate predicting offline hate).

Methodological transparency and presentation issues: Several aspects of the study's methodology and presentation 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, the specific validation method used to generate the performance metrics, 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 that "The data used in this study were collected at a time

before the social media giants introduced strict hate speech policies" (p. 114), the six-year gap between data collection and publication may limit the external validity of the findings. Third, there is an unexplained discrepancy in the sample size: the methods section states the study includes 4,720 LSOAs, but the first regression table reports a sample size of N = 4,270, a loss of 9.5% of the units that is not accounted for (pp. 102, 107). Finally, the article provides a contradictory description of its fixed-effects models, stating in one section that they are based on "within-borough variation" (p. 102) and in another on "within-LSOA variation" (p. 108), creating ambiguity about the model specification.

Future Research

Granular temporal analysis: Future work should utilize daily or weekly time-series data rather than monthly aggregations. This would allow for the application of Granger causality tests or similar time-lagged regression techniques to rigorously assess whether spikes in online hate speech genuinely precede offline incidents, thereby addressing the issue of simultaneity and temporal precedence.

Network-based exposure measurement: Researchers should move beyond the geographic production of tweets (geotags) and focus on the geographic consumption of content. By analyzing follower networks or estimating the location of users exposed to hate speech, future models could test whether the *audience* of hate speech is located in the areas where offline crimes occur, correcting the spatial mismatch inherent in production-based measures.

Corrected effect size estimation: Subsequent studies should replicate the Poisson regression analysis ensuring the correct mathematical interpretation of Incidence Rate Ratios. This is essential to determine whether the relationship between social media posts and crime rates is of a magnitude that is practically significant for policy interventions, or if the statistical significance observed in this study was driven by the large sample size rather than a substantial effect.

isitcredible.com