A Review of "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

January 14, 2026

v1

isitcredible.com

Disclaimer

This report was generated by Reviewer 2, an automated system that uses large language models to assess academic texts. It has been read and approved by a human editor on behalf of The Catalogue of Errors Ltd. The report's goal is to facilitate the discovery of knowledge by identifying errors in the existing literature. Comments can be made here. Any errors 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., and 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, pp. 93–117.

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

Abstract Summary: This article uses Computational Criminology to link police crime, census, and Twitter data to establish a temporal and spatial association between online hate speech targeting race and religion and offline racially and religiously aggravated crimes in London over an eight-month period. The findings renew the understanding of hate crime as a process, rather than a discrete event, for the digital age.

Key Methodology: Computational Criminology, linking police crime, census, and Twitter data, negative binomial regression, random- and fixed-effects Poisson panel models, supervised machine learning classification (Support Vector Machine with Bag of Words).

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?

The study by Williams et al. presents a compelling dataset and an ambitious attempt to link the digital sphere with physical reality. By combining machine-learning classification of Twitter data with police-recorded crime statistics across London, the authors provide strong evidence for a general association between online hate speech and offline racially and religiously aggravated offenses. The use of fixed-effects Poisson models to control for unobserved heterogeneity is a methodological strength that lends weight to their core claim: that online hate is not merely a reflection of offline events, but a predictor of them. However, while the core correlation appears robust, the credibility of the article is significantly undermined by a series of interpretative errors and clerical inconsistencies. The authors appear to misunderstand the mathematical properties of their own statistical models, leading them to drastically understate the magnitude of the effects they have discovered.

The most damaging issue lies in the narrative interpretation of the Poisson regression results. In their Random Effects model for harassment, the authors report a coefficient of roughly 0.004. They interpret this to mean that an increase of 100 hate tweets corresponds to a mere 0.4 percent increase in crime. This interpretation reveals a fundamental confusion regarding the model. A coefficient of 0.004 in a Poisson model implies a multiplicative effect; an increase of 100 units should result in an increase of approximately 50 percent in the outcome variable, not 0.4 percent. By confusing the decimal coefficient with a percentage and applying linear scaling to a non-linear model, the authors have understated the strength of the association by two orders of magnitude. While this error paradoxically means the actual relationship is much stronger than claimed, it casts doubt on the rigorousness of the quantitative interpretation and the review process.

Furthermore, the article struggles to maintain a consistent position on causality and temporal precedence. Early in the methodology, the authors correctly acknowledge that their ecological design and monthly aggregation prevent them from determining whether online hate speech precedes offline crime. Yet, in the discussion of their results, they reverse this

stance, claiming that their models "allow us to determine if online hate speech precedes rather than follows offline hate crime." This claim is methodologically unsupported. The use of contemporaneous monthly data—where tweets in a given month predict crime in that same month—cannot establish temporal order. The aggregation of data to the monthly level likely smooths out the high-frequency dynamics of "trigger" events, obscuring the very mechanisms the authors aim to illuminate.

Finally, the presentation of statistical significance is marred by clerical inconsistencies. In the fixed-effects model for criminal damage, the interaction term between BAME population and hate tweets is marked as statistically significant, yet the reported t-statistic (derived from the coefficient and standard error) is approximately 1.0, which is far below the threshold for significance. Similarly, the reported range of variance explained does not match the data in the tables. These errors, while perhaps clerical in origin, suggest a lack of precision in the final reporting. The study succeeds in demonstrating a correlation, but the specific quantification of that relationship and the claims regarding its directionality are unreliable.

The Bottom Line

Williams et al. successfully demonstrate a statistical link between the volume of anti-Black and anti-Muslim tweets and the frequency of offline hate crimes in London. However, the study is severely compromised by a major mathematical error in the narrative interpretation of the regression coefficients, understating the impact of online hate speech by a factor of roughly 100. Additionally, the article makes contradictory claims regarding its ability to determine whether online speech causes offline crime, overstating the power of its methodology. While the general correlation is likely valid, the specific numbers and causal claims presented in the article should be viewed with extreme skepticism.

Specific Issues

Clerical error in narrative interpretation of Incident Rate Ratios: There is a fundamental error in the narrative interpretation of the magnitude of the effect of online hate speech on page 108. In the Random Effects Poisson model for harassment (Table 2, p. 110), the coefficient for "Hate Tweets" is 0.00404, yielding an Incident Rate Ratio (IRR) of 1.00405. The authors state that "an increase of 100 hate tweets would correspond to a 0.4 per cent increase." This is mathematically incorrect for two reasons. First, an IRR of 1.004 represents a 0.4 percent increase per single unit, not 0.004 percent as the authors imply. Second, Poisson models are multiplicative, not linear. An increase of 100 tweets corresponds to raising the IRR to the power of 100 (1.00405^{100}), which results in an increase of approximately 50 percent, not 0.4 percent. Consequently, the authors have understated the effect size by two orders of magnitude.

Contradictory claims regarding temporal precedence: The article presents conflicting statements about its ability to establish causality. On page 107, the authors correctly note, "We cannot say if online hate speech precedes rather than follows offline hate crime." However, on page 108, they contradict this by claiming the models "allow us to determine if online hate speech precedes rather than follows offline hate crime." Given that the data is aggregated to the monthly level and the models use contemporaneous regressors (tweets in month t predicting crime in month t), the methodology cannot statistically distinguish precedence. The claim on page 108 is unsupported by the study design.

Contradiction between significance marker and test statistics: In Table 2 (p. 109), the interaction term (Prop. BAME × Hate Tweets) for the criminal damage model is marked with an asterisk, indicating statistical significance ($p < 0.05$). However, the reported coefficient is 0.00003 and the standard error is 0.00003, yielding a t-statistic of 1.0. A t-statistic of 1.0 is not statistically significant. This contradicts the article's claim on page 111 that the interaction term was significant for "all" hate crime categories.

Clerical errors in reporting variance and fit statistics: There are multiple minor discrepancies in the reporting of model fit. The article claims the increase in variance explained (Ad-

justed R^2) ranges "between 13 per cent and 30 per cent" (p. 111). However, the data in Table 2 shows the increase for criminal damage is only 11.25 percent (0.1367 minus 0.0242), which falls outside the claimed range. Additionally, the authors report the classifier F-measure as 0.771 (p. 101), whereas the calculation based on the reported precision (0.89) and recall (0.69) should be approximately 0.777. Furthermore, the authors report Adjusted R^2 values for Random Effects models only (p. 110), despite identifying Fixed Effects as the more robust test, a limitation necessitated by the statistical framework but which creates a disconnect between the "best" models and the reported fit metrics.

Methodological limitations regarding aggregation and proxies: The study relies on monthly aggregation of data (p. 101), a temporal resolution that may be too coarse to capture the immediate "trigger" effects of online hate speech, which often occur over hours or days. Additionally, the authors use the total count of geo-coded tweets as a proxy for population density (p. 111) and exclude religious demographic variables due to multicollinearity (p. 102). While these are defensible choices supported by cited literature, they introduce constraints on the precision of the findings. The authors also acknowledge that they removed four influential outlier LSOAs, including Heathrow Airport and Westminster (p. 104), which limits the generalizability of the findings to typical residential areas.

Ecological and data quality limitations: The authors acknowledge several limitations inherent to their data sources. The machine learning classifier has a recall of 0.69 (p. 101), meaning roughly 31 percent of hateful tweets were missed, likely leading to attenuation bias. The study also lacks explicit validation of the spatial accuracy of the "geo-coded" tweets (p. 101), leaving it unclear whether these are precise GPS coordinates or broader bounding boxes. Finally, the authors engage in speculative interpretation of the inverted U-shape relationship regarding unemployment (p. 106), suggesting specific perpetrator demographics that cannot be verified with the aggregate ecological data used. The authors do, however, transparently acknowledge the "ecological fallacy" risk (p. 107) and the inability to observe individual-level mechanisms.

Future Research

Correction of effect size estimation: Future work must re-evaluate the magnitude of the relationship between online hate speech and offline crime using the correct mathematical interpretation of Poisson regression coefficients. Researchers should calculate the predicted probabilities and incident rate ratios using the exponential function proper to the model, rather than the linear approximation used in this article. This is essential to accurately quantify the risk posed by online hate speech.

High-frequency temporal analysis: To genuinely test the hypothesis that online hate speech *precedes* offline crime, future studies should move beyond monthly aggregation. Research should utilize daily or weekly data to employ time-series methods such as Granger causality tests or vector autoregression. This would allow for the detection of lead-lag relationships and better isolate the impact of online triggers from simultaneous events.

Individual-level mechanism verification: To address the ecological limitations acknowledged in the article, future research should attempt to link data at a more granular level. This could involve qualitative forensic analysis of specific cases where online threats culminated in offline violence, or the use of victimization surveys that capture the digital experiences of victims prior to physical incidents. This would help validate the "process" theory of hate crime by observing the transition from online to offline hostility at the individual rather than the neighborhood level.

www.isitcredible.com