specific, and therefore likely to impact more individuals who are geographically proximate to the crime location.

(1) Hence we first categorized all our tweets into different Census block groups using the geolocation information in tweets. Then we computed a mean measure of NA, anger, anxiety, sadness over all tweets in each block group.

(2) Next, for each block group, we calculated the historical crime statistics, using the crime data we collected between Jan 1-Oct 9, 2014. We divided the number of crimes by the population of the block group provided by Census in order to generate the crime per capita for the block group for each crime category.

(3) Finally, we also compiled a set of demographic and socio-economic status (SES) variables per block group, as given by Census data. Our motivation for computing the demographic and SES variables was that it is known these variables can differ widely across different census block groups. We used the following variables – proportion of males in population, median age, median income, proportion of population who are high school graduates, and proportion of population who have a bachelor's degree.

Following this, we framed a regression task, where our **dependent variable** was the mean LIWC NA, anger, anxiety, or sadness per block group, and **independent variables** were different types of crimes that happened before the tweet time period i.e., Jan 1-Oct 9, 2014, and the demographic and SES variables in the same block group. Specifically we used a two phase linear regression – first we built a "baseline model" for each LIWC category where only the demographics and SES variables were used. As a next step, we build a "crime model" where we used both demographics and SES and crime data as independent variables. Note that, for analysis we only included those block groups entirely within the Atlanta jurisdiction map ($N$=235). Block groups were excluded if they had no census population or no population of age eligible for one or more of SES independent variables utilized in the analysis.

## Results on Long-term Impact of Crime

**NA-Crime Model.** From Table 1, we observe that adding crime related variables to the baseline model (that uses only demographics and SES variables) results in improvement of model fit for NA. The negative of the log likelihood of the crime model decreases (Likelihood ratio=20) and the adjusted $R^2$ reduces by 29%. Based on a chi-square test ($\chi 2$ ($N$=235)=10, $p$<.0001) we find this difference to be significant over the baseline model. We find all crime types, except robbery to have positive $t$-statistic values – this indicates that increase in homicide, aggravated assault, larceny and burglary are associated with heightened NA in twitter posts for the block groups we examine.

| Variables | Baseline *t*-stat | | Crime *t*-stat | |
|---|---|---|---|---|
| (Intercept) | 7.988 | *** | 8.676 | *** |
| male ratio | 0.959 | | 0.077 | |
| median age | -0.966 | | -1.350 | ** |
| income | -3.352 | *** | -1.718 | *** |
| prop high school | -0.878 | | -0.429 | |
| prop bachelors | -2.475 | ** | -1.484 | * |
| prop under poverty | 0.145 | | 1.766 | ** |
| homicides | | | 1.501 | ** |
| robbery | | | -1.713 | *** |
| assault | | | 2.372 | *** |
| larceny | | | 1.569 | * |
| burglary | | | 0.297 | |
| df | 228 | | 223 | |
| Log likelihood | -730.9 | | -740.9 | |
| Likelihood ratio wrt baseline | | | 20 | |
| $p$ | 4.81E-22 | | 4.02E-23 | |
| adj. $R^2$ | 0.322 | | 0.416 | |
| * p < 0.05 ** p < 0.01 *** p < 0.001 | | | | |

**Table 1. Linear regression model results with LIWC Negative emotion in Census block groups as the dependent variable. Baseline model includes only demographic and SES variables. Crime model includes historical crime data per capita along with demographic and SES as the independent variables.**

**Anger-Crime Model.** Next we observe that addition of crime variables in predicting anger also results in an improved model fit, compared to the baseline model (Table 2). The negative of the log likelihood decreases (Likelihood ratio=20.6) here and the adjusted $R^2$ decreases by 8.4%. A chi-square test ($\chi 2$ ($N$=235)=10.3, $p$<.0001) shows the improvement in model fit over baseline to be significant. The strongest crime variables here are again robbery and assault, followed by larceny and homicide.

| Variables | Baseline *t*-stat | | Crime *t*-stat | |
|---|---|---|---|---|
| (Intercept) | 7.663 | *** | 5.854 | *** |
| male ratio | 1.136 | | 0.328 | |
| median age | -1.584 | ** | -1.992 | *** |
| prop bachelors | -3.554 | *** | -1.844 | ** |
| prop under poverty | 1.388 | ** | 1.531 | ** |
| homicides | | | 0.688 | * |
| robbery | | | -2.009 | *** |
| assault | | | 3.395 | *** |
| larceny | | | 1.441 | ** |
| df | 228 | | 223 | |
| Log likelihood | -860.4 | | -870.7 | |
| Likelihood ratio wrt baseline | | | 20.6 | |
| $p$ | 4.75E-25 | | 1.06E-25 | |
| adj. $R^2$ | 0.41 | | 0.448 | |
| * p < 0.05 ** p < 0.01 *** p < 0.001 | | | | |

**Table 2. Linear regression model results with LIWC *Anger* in Census block groups as the dependent variable. Only significant variables are shown.**

**Anxiety-Crime Model.** Our third model examines the predictability of LIWC anxiety words in tweets using crime variables (Table 3). The performance of this model is relatively worse compared to the previous two models. Here, we still see slight decrease in negative of the log likelihood of the crime model over the baseline (Likelihood ratio=10.4); the adjusted $R^2$ also shows some reduction (13.5%). A chi-square goodness of fit test however does not yield significance here ($\chi 2$ ($N$=235)=5.2, $p$=.13).

| Variables | Baseline *t*-stat | | Crime *t*-stat | |
|---|---|---|---|---|
| (Intercept) | 5.182 | *** | 4.121 | *** |
| median age | 1.838 | ** | 1.693 | ** |
| income | -2.053 | ** | -2.552 | *** |
| prop under poverty | 1.276 | * | 1.793 | ** |
| robbery | | | -1.907 | *** |
| df | 228 | | 223 | |
| Log likelihood | -1028.5 | | -1033.7 | |
| Likelihood ratio wrt baseline | | | 10.4 | |
| *p* | 1.24E-07 | | 2.08E-07 | |
| adj. $R^2$ | 0.148 | | 0.168 | |
| * p < 0.05 ** p < 0.01 *** p < 0.001 | | | | |

**Table 3. Stepwise regression model results with LIWC *Anxiety* in Census block groups as the dependent variable. Only significant variables are shown.**

We do not include detailed model results for sadness for space limitations. The performance of this model is the poorest among the four in terms of model fit (adjusted $R^2$=.044). We also find only marginal improvement in the crime model over the baseline model that uses only demographic and SES variables – a chi-square goodness of fit test does not show significance ($\chi 2$ ($N$=235)=3.7, $p$=.21).

Together, the four models indicate that while there is variability in *what* psychological attributes in tweets may be predicted using crime data, for NA and anger we see notable ability of historical crime to explain variance. Among the different crime variables, robbery shows significance consistently for the first three models – NA, anger, anxiety. Aggravated assault is highly significant for NA and anger. Surprisingly, we found homicide to be significant only marginally for our models; it is significant only for NA, anger, and sadness. To summarize, the regression models indicate that historical incidence of homicide, robbery, larceny, and burglary are important predictors of the psychological content of tweets in a future period.

## Discussion and Conclusion

In our first study, our results did not indicate the presence of any short-term impact of crime on social media derived psychological attributes. Better user filtering may help to focus on user accounts which are particularly likely to be affected by crime events. We may also wish to focus on a more homogeneous set of neighborhoods, such as residential areas, since likely, both the distribution of crimes and of Twitter content differs substantially in downtown areas. Additionally, through the permutation test we did not account for the presence of cascading effects of crimes, or crimes that impact communities differentially. Gauging how residents of these areas learn about the crime events and their (media) exposure will help us better interpret the findings. In the future, we are also interested in analysis of additional variables that may explain the observed effects.

On the other hand, in the next part of our study corresponding to determining the impact of long-term crime, our results showed that historical incidence of criminal activity is related to social media in the same area at a later time. Perhaps expected, data also showed that demographic characteristics, in particular income and education, have a large effect on the psychological measures derived from Twitter. However, the predictivity of crime variables remained significant even after controlling for SES variables.

Overall the findings of our historical crime study align with those in the literature (Skogan 1981) – we show long-term exposure to crime to be associated with heightened negative emotion and anger. We conjecture that perhaps in a city like Atlanta, crime impacts individuals, however only when it is beyond a certain exposure level. Hence we find that while crime was not correlated with social media affect immediately, it was so when we considered a longer-period of crime incidents. While a causal relationship may not be directly inferred, we hope our findings to shed light on the potential and limitations of using social media as a lens to understand urban well-being longitudinally.


## Acknowledgement

This research is supported through a National Institutes of Health grant R01 GM112697-01.


## References


Chainey, S., Tompson, L., & Uhlig, S. (2008). The utility of hotspot mapping for predicting spatial patterns of crime. *Security Journal*, 21(1), 4-28.

Coppersmith, G., Harman, C., & Dredze, M. (2014). Measuring post traumatic stress disorder in Twitter. In *Proc. ICWSM*.

Covington, J., & Taylor, R. (1991). Fear of crime in urban residential neighborhoods. *The Sociological Quarterly*.

De Choudhury, M., Monroy-Hernández, A., & Mark, G. (2014). Narco emotions: affect and desensitization in social media during the mexican drug war. In *Proc. CHI*.

Gerber, M. S. (2014). Predicting crime using Twitter and kernel density estimation. *Decision Support Systems*, 61, 115-125.

Glasgow, K., Fink, C., & Boyd-Graber, J. (2014). Our grief is unspeakable: Measuring the community impact of a tragedy. In *Proc. ICWSM*.

Kim, J., Cha, M., & Sandholm, T. (2014). SocRoutes: safe routes based on tweet sentiments. In *Proc. WWW Companion*.

Lee, L. (1999). Measures of Distributional Similarity. *Proc. ACL*.

Pacione, M. (2003). Urban environmental quality and human wellbeing—a social geographical perspective. *Landscape and Urban Planning*, 65(1), 19-30.

Pennebaker, J. W.; Mehl, M. R.; and Niederhoffer, K. G. (2003). Psychological aspects of natural language use: Our words, our selves. *Annual Rev. Psychology* 54(1):547–577.

Ross, C. E., Reynolds, J. R., & Geis, K. J. (2000). The contingent meaning of neighborhood stability for residents' psychological well-being. *American Sociological Review*, 581-597.

Skogan, W. G., & Maxfield, M. G. (1981). Coping with crime: Individual and neighborhood reactions (p. 272). Sage Publ.

Wang, X., & Brown, D. E. (2011). The spatio-temporal generalized additive model for criminal incidents. In *Proc. ISI*, 2011.

Wang, X., Brown, D. & Gerber, M. (2012). Spatio-temporal modeling of criminal incidents using geographic, demographic, and Twitter-derived information. In *Proc. 2012 IEEE International Conference on Intelligence and Security Informatics*.