Predicting Community-Level Criminal Behaviors by Estimating Human Attitudes from Social Media

Application for Presidential Fellowship in Data Science

Mohammad Al Boni
ma2sm@virginia.edu
Advisor: Dr.Matthew M. Gerber
Department of System and Information Engineering

Jordan Axt
jra3ee@virgnia.edu
Advisor: Dr. Brian A. Nosek
Department of Psychology

Lin Gong
lg5bt@virginia.edu
Advisor: Dr.Hongning Wang
Department of Computer Science

Background

Crimes that target individuals based on race or sexual orientation remain a common problem. In 2014, there were over 6,000 hate crimes in America. Half of these were racially motivated, and a fifth were based on sexual orientation (Federal Bureau of Investigation, 2015). In many cases, perpetrators of hate crimes have expressed their attitudes on social media. For example, Craig Stephen Hicks, who in 2015 killed three Muslim students in Chapel Hill, North Carolina, frequently voiced anti-religious attitudes online (Mohney, 2015). Another recent tragedy involving the death of 14 innocent civilians in San Bernardino, California shed light on similar behaviors. Tashfeen Malik, one of the shooters, had posted several inflammatory comments on social media prior to moving to the US. These comments were missed during the background checks by the immigration officials (Apuzzo, Schmidt & Preston, 2015). Detecting such online behavior from criminals early and effectively might then save lives.

Attitudes are a key factor in understanding behavior (Ajzen & Fishbein, 1977), and while we cannot retroactively measure the attitudes of a specific individual, we can use existing data to estimate the attitudes of a community. We plan to use communal language on Twitter to first measure an area's attitudes, and then investigate how area-level attitudes relate to crime. If we can leverage social media data to estimate the attitudes of a community, we can more effectively predict criminal behaviors and take the necessary steps to better prevent such tragedies.

While most psychological research focuses on how an individual's subjective construal of the environment affects attitudes, beliefs and behavior (Wilson, 2011), this more individual-centered framework overlooks the role of more objective influence of one's surroundings. Through a greater focus on the objective environment, we adopt a socioecological approach (Oishi, 2014) to explore how individual attitudes and behavior are influenced by social and natural habitats. This approach allows researchers to better understand how broader, contextual forces are related to individual psychological processes.

Our fellowship will explore one salient aspect of the social environment: language. Specifically, we focus on Twitter language to see how the words used within a community relate to attitudes. This project draws off previous work illustrating how language on social media is an effective predictor of community-level outcomes. In one study, the proportion of tweets containing negative words (e.g., "hate") within a county was associated with greater prevalence of atherosclerotic heart disease (AHD; Eichstaedt et al., 2015). In fact, Twitter language alone predicted AHD just as well as county-level demographics like age, income and education. Such results are striking considering that the people expressing negative sentiment on Twitter, where the median age is 31, were unlikely to be the people contracting AHD. Rather, Twitter language may capture a more communal factor - such as greater stress - that is associated with both younger residents using more negative words and older residents contracting more AHD. Using a similar approach, we will first explore how Twitter language predicts community-level attitudes, and then test the relationship between communal attitudes and criminal behavior.

Objectives

To complete our analyses, we will first construct a dataset of Twitter language, community attitudes and criminal behaviors. This will involve collecting a considerable amount of Twitter data, racial and sexual attitudes data as well as crime data. Once we have these data available, we will build a predictive model by first analyzing the correlation between Twitter language and attitudes. Next, we will extend this analysis to other outcomes by testing the relationship between attitudes and criminal behavior.

The scope and potential impact of this project requires expertise in systems engineering, computer science, and psychology, making it an ideal fit for this collaborative fellowship.

Methodology

Acquiring Large-Scale Data

We will first investigate whether the prevalence of anti-gay and anti-Black language on Twitter is related to an area's racial and sexual attitudes. This analysis will help validate our measure of area-level attitudes, which we will then use to predict criminal behavior. Twitter has made a random sample of geotagged tweets available for research, and we will collect a corpus of tweets through the API provided by Twitter featured by location search.

Our Twitter analyses will focus on hate speech concerning two groups: gay people (measured through use of the words "dyke" and "faggot") and Black people (measured through use of the word "nigger"). Both Black and non-Black people use the word "nigger" on social media frequently and with differing intents; to account for this disparity, initial analyses will look at total uses of the word and follow-up analyses may code for writer's intentions.

We will test whether the frequency of anti-Black and anti-gay words in an area is related to community-level estimates of racial and sexual attitudes. We will use measures of both *explicit* and *implicit* attitudes. Explicit attitudes refer to people's felt, conscious preferences, and can be measured through asking people whether they prefer White or Black people, for example. Conversely, implicit attitudes are attitudes that may not be endorsed or felt consciously but are still present in mind (Greenwald & Banaji, 1995) and are partly shaped by exposure to cultural messages about group status (Axt, Ebersole & Nosek, 2014).

In the most common measure of implicit attitudes, the Implicit Association Test (IAT; Greenwald, McGhee & Schwartz, 1998) participants categorize words or images one at a time as fast as possible using two keys. In an IAT measuring attitudes towards White and Black people, participants complete some blocks where they categorize images of Black people and positive words using the same key and images of White people and negative words using the other key. In other blocks, the pairing is reversed, and participants categorize images of White people and positive words with one key and images of Black people and negative words with the other key. The speed at which participants complete these different blocks is used to infer the strength of associations between the categories and attributes. Being faster at blocks that pair White people with positive words (and Black people with negative words) compared to blocks that pair White people with negative words (and Black people with positive words) indicates more positive implicit attitudes for White compared to Black people. Using both explicit and implicit attitude measures allows us to investigate whether an area's Twitter language is more related to endorsed, explicit attitudes or potentially unendorsed but still impactful implicit attitudes.

Estimates of area-level racial and sexual attitudes will come from Project Implicit, an online laboratory where participants volunteer to complete IATs about several topics. Sexuality attitudes will come more than 500,000 U.S. participants between 2006 and 2015. Race attitudes will come from more than 1,000,000 U.S. participants between 2002 and 2015. Participants also provide their zip code, meaning we can create estimates of an area's overall implicit and explicit attitudes. The size of these datasets allows us to conduct a well-powered test of the relationship between Twitter language and attitudes, with 558 counties having at least 10 participants in the sexuality dataset and 815 counties in the racial dataset. We selected three counties (Los Angeles, CA; Cook, IL and Harris, TX) in the datasets and collected a sample of tweets from these areas over 60 days. We collected at least 100 million tweets from each county, and are confident we can collect comparably large numbers of tweets from the remaining counties.

Finally, county-level crime behavior will be measured through police reports, which record both the location and type of crime committed. For example, Cook County, IL provides incident-level detail for more than 6,000,000 crimes since 2001. We will use similar police reports to aggregate crime data for the remaining counties included in our analysis.

Incorporating Attitudes as Latent Variables for Predictive Modeling

In our first analysis, we will treat implicit and explicit racial or sexual attitudes, represented by the IAT score, as our response variable, and we will build regression models relating tweets' textual content to county-level attitudes. Then, we will test whether Twitter language is a better predictor of implicit and explicit attitudes than county-level demographic variables such as age, education, income, and political orientation. These analyses can reveal how Twitter language may be a useful proxy for area-level attitudes that is even more effective than county-level demographics. Note, we will initially only consider tweets with anti-Black and anti-gay language, but then we will locate what words in general best predict county-level attitudes. Herein, we will expand our analysis to include all tweets and extract important features using feature selection methods, i.e., information gain, mutual information (Yang & Pedersen, 1997).

Next, we will use the best performing models from the first analysis to predict the IAT scores concentrations of various regions, and then, use them to predict crime. We will build classification models relating a set of features describing a geo-spatial point to a response variable which indicates the odds of observing future crimes at that point. We will also test our models on various crime types, yet we will focus a great deal of attention on certain crime types that have potential relation to anti-Black and anti-gay attitudes specifically. If high anti-Black and anti-gay attitudes are indicative of greater intergroup mistrust, then they may be more related to certain crimes than others. For instance, implicit and explicit attitudes may be better predictors of crimes that require a victim (e.g., assault versus narcotics). As for evaluation, we will generate surveillance plots for the study regions (Gerber, 2014), and compare our models to a standard baseline including only the historical crime density estimate as a predictor for future crimes.

Impact

We will test whether anti-Black and anti-gay language on Twitter can be used to predict area-level attitudes, and then analyze whether such attitudes are related to criminal behavior. These analyses will combine very large datasets of tweets, attitude measures, and police reports. This project will have a large and interdisciplinary impact. Results will be published in both psychological and computer science journals. Moreover, the work will introduce psychologists to new or underused analyses, and further highlight the use of social media data for research. The project will also expose computer scientists and systems engineers to existing attitude datasets as a resource and allow for additional analyses to new outcomes (e.g., the prevalence of mental health disorders).

Our project has both practical and theoretical applications. Practically, this work will allow us to use Twitter language to estimate attitudes in areas with very little or no existing data, improving knowledge of how such attitudes vary around both the country and the world. In addition, our results could indicate whether attitudes may be used to predict where certain types of crimes are more likely to take place, an insight that could help save lives. Finally, this work can inform whether communal language is more related to explicit versus implicit attitudes, and identify what types of behavior are most related to highly negative intergroup attitudes. By taking a broader approach and focusing on the general prevalence of words and attitudes within a community, researchers can better understand what factors shape individual behavior.

References

Ajzen, I., & Fishbein, M. (1977). Attitude– behavior relations: A theoretical analysis and review of empirical research. *Psychological Bulletin*, *84*, 888–918.

Apuzzo, M., Schmidt, M., & Preston, J. (2015, December 12). U.S. visa process missed San Bernardino wife's online jealousy. *The New York Times*. Retrieved from http://www.nytimes.com/2015/12/13/us/san-bernardino-attacks-us-visa-process-tashfeen-maliks-remarks-on-social-media-about-jihad-were-missed.html

Axt, J. R., Ebersole, C. R., & Nosek, B. A. (2014). The rules of implicit evaluation by race, religion, and age. *Psychological Science*, *25*, 1804-1815.

Eichstaedt, J. C., Schwartz, H. A., Kern, M. L., Park, G., Labarthe, D. R., Merchant, R. M., ... & Weeg, C. (2015). Psychological language on Twitter predicts county-level heart disease mortality. *Psychological Science*, *26*, 159-169.

Federal Bureau of Investigation (2015). Hate crime statistics 2014. Retrieved from https://www.fbi.gov/about-us/cjis/ucr/hate-crime/2014

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

Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: attitudes, self-esteem, and stereotypes. *Psychological Review*, *102*, 4.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. (1998). Measuring individual differences in implicit cognition: The Implicit Association Test. *Journal of Personality and Social Psychology*, *74*, 1464-1480.

Mohney, G. (2015, February 13). Chapel Hill shooting: Social media provides a glimpse of alleged killer. *ABC News*. Retrieved from http://abcnews.go.com/US/chapel-hill-shooting-alleged-shooter-long-frustrated-parking/story?id=28948627

Oishi, S. (2014). Socioecological psychology. *Annual Review of Psychology*, *65*, 581-609.

Wilson T.D. (2011). *Redirect: The Surprising New Science of Psychological Change*. New York: Little, Brown.

Yang, Y., & Pedersen, J. O. (1997). A comparative study on feature selection in text categorization. *ICML*, *97*, 412-420.