**Date: January 25, 2016**

**Subject: Data Collection, Use, and Challenges for NYCHA Projects Concerning Community Sentiment**

**From: Aaron Chalfin, Jason Lerner, Patrick Sharkey, and Thomas Laetsch**

1. Introduction

**This is to outline the various sources of data near use, or being considered for monitoring senti- ment and perceptions of those affected by public housing, particularly residence in these properties.** In the following section, we lay out the specifics of each of the three main resources of data currently in consideration: a cell-phone community survey, administrative data, and online data sources.

2. Data Sources

2.1. Community Survey. We will disseminate a short, cell-phone based, two-cycle community survey amongst residence of NYCHA public housing. This survey will target specific aspects of individual perceptions and sentiment of the aforementioned residents:

• Cycle 1 will be used to gauge: individual satisfaction within the neighborhood; perceived violence in or near the public housing unit; interactions with police including reporting to the agency, or interacting directly with officers; and collective efficacy.

• Cycle 2 will be used to gauge: individual satisfaction within the neighborhood; feelings of victimization or fear created by either previous acts of violence, or the concern for future acts of violence; perceived disorder including cleanliness in and near the housing unit, and potentially deleterious activities within public spaces; and perceptions of police and policing with regards to the public housing community.

2.1.1. Benefits. As these surveys will be responses from the residence of public housing, it will provide a direct insight into the pulse of this community which might otherwise not be readily accessible. The use of cell-phones as the medium to collect survey data streamlines data acquisition, and reduces cost of conducting such a survey. Moreover, the results from these surveys might be valuable in validating sentiment classifiers from less direct methods, such as through inferences made via social media data.

2.1.2. Challenges. To increase response rate and willingness to participate, each survey given will only ask for responses to a small subset of questions comprising the totality of the survey; this subsetting introduces noise within the collected surveys and possible difficulties in analysis of the results which are unknown a priori.

2.2. Administrative Data. Administrative data will be central in moving forward with this effort. Combining several publicly available city-service related data sources, working to understand how to link them spatially and/or temporally and correlating the resulting corpus to our results from other avenues of inquiry, may result in a predictor of several facets of community sentiment based on readily available public data (perhaps in combination with alternate publicly available data from online sources). The sources of administrative data currently in consideration are the following.

2.2.1. 311. Data from 311, a resource for non-emergency services to NYC residents, is made publicly available through the NYC developers portal API system ([https://developer.cityofnewyork.us/api/).](https://developer.cityofnewyork.us/api/) An ostensibly valuable direction to take with this data is to correlate use of and types of non- emergency service calls/responses to perceived neighborhood disorder (such as mentioned in the community survey).

2.2.2. 911.

2.2.3. Arrest and Complaint Records.

2.3. Online Repositories.

2.3.1. Twitter. Twitter is a tempting resource in part because of the ubiquity of its use, in part because it is apparent that sentiment is often expressed within tweets, in part because of the availability of large amounts of Twitter data through Twitter’s API, and in part because there has already been a push by researchers to distil and understand sentiment content from Twitter data.

Nonetheless, the difficulty of the latter task, distilling and understanding sentimental data from Twitter, has been an elusive goal. The work of the Computational Story Lab [3, 5], part of the University of Vermont, has made some strides with their hedonometer which, based on surveys, assigns a “happiness” rating one each word acontextually. Unfortunately, the hedonometer approach appears to only be a statistically useful device for large collections of words, making inferences about collections of words (on the order of 10000 words, in fact) together – this method would not be useful at the level of an individual tweet, nor a small collection of individual tweets. As our goal is to grab spacially and temporally fine (localized) data, we are unlikely to have the ability to retrieve such large bodies of words with which to make inferences. That said, it might be possible to observe some trends in a less localized setting; e.g., time evolution of the hedonometric sentiment for all residence of public housing combined.

Some work which seems better suited to the spatial and temporal localization we seek can be found in [2, 1, 4]. A general theme in this work is to train a tweet sentiment classifier, training on those tweets with emoticons expressing sentiment. In fact, a sentiment map of New York City is offered in [2], created with the output of their sentiment classifier, which purportedly is spatially and temporally localized at levels appropriate for our interests. However, each of these works was done in the Twitter vacuum, without other means to confirm or gauge accuracy of their findings.

What seems to be a prudent course of action is to mine and use Twitter data influenced by the methods of [2, 1, 4] and use the confluence of data sources, particularly the community survey, to test the reliability of the Twitter-based sentimental classifier. As a future goal, given a reliable localized sentiment classifier, we could attempt to find global trends and compare these to the hedonometric-style methods of [3, 5].

2.3.2. Other Possibilities.

• Twitter Survey.

• Facebook. There are Facebook pages set up for several specific NYCHA public housing units; seemingly those participating in posts and/or are members of these pages are persons who are residents of, or closely interactive with residents of these units. The fact that these pages are directed at specific housing units is appealing since posts to these pages give us a timestamped insight into the particular housing units. The challenge is discerning whether these posts are translatable into inferentially meaningful sources of localized sentiment. Within Manhattan, the three most active pages are for the Alfred E Smith Houses, the corresponding Facebook page [here h](https://www.facebook.com/groups/42049093460/)aving 1025 members; the Frederick Douglass Houses, the Facebook page [here h](https://www.facebook.com/groups/104654293240/)aving 1645 members; and the Jacob Riis Houses, the Facebook page [here h](https://www.facebook.com/groups/234063750053/)aving 1694 members. A first step in this direction would be to examine and understand how (if at all possible) to use the Facebook data on these extremely active pages and from there, work to address whether those methods can be used for the less active pages.

• Yelp.

• Google Searches.

• Message Boards.

References

[1] Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, and Rebecca Passonneau. Sentiment analysis of Twitter data. Association for Computational Linguistics, pages 30–38, 2011.

[2] Kz Bertrand, Maya Bialik, and Kawandeep Virdee. Sentiment in New York City: A High Resolution Spatial and Temporal View. arXiv preprint arXiv: . . . , pages 1–12, 2013.

[3] Morgan R. Frank, Lewis Mitchell, Peter Sheridan Dodds, and Christopher M. Danforth. Happiness and the Patterns of Life: A Study of Geolocated Tweets. Scientific Reports, 3:1–9, 2013.

[4] Alec Go, Richa Bhayani, and Lei Huang. Twitter Sentiment Classification using Distant Supervision. Processing, 150(12):1–6, 2009.

[5] Andrew Reagan, Brian Tivnan, Jake Ryland Williams, Christopher M Danforth, and Peter Sheridan Dodds. Benchmarking sentiment analysis methods for large-scale texts: A case for using continuum-scored words and word shift graphs. 2015.