Table of Contents:

Final Report- Neighborhood Navigator

INTRODUCTION

An interdisciplinary team of social scientists, computer scientists, designers, and researchers from the SAFElab at Columbia University’s School of Social Work, School of Engineering and Applied Science and Data Science Institute, helped to develop a resource called the Neighborhood Navigator under an initiative by the Office of Neighborhood Safety (ONS) within the Mayor's Office of Criminal Justice (MOCJ). The purpose of this initiative was to assess patterns and changes in the sentiment of quality of life, wellbeing, community, and living conditions among residents of New York City. SAFElab received support to carry out this project from the Research and Evaluation Center at John Jay College of Criminal Justice. SAFElab's contributions to the Neighborhood Navigator initiative drew upon community focus groups and one-on-one interviews in concert with artificial intelligence (AI) techniques (e.g. natural language processing [NLP] and computer vision) to provide short-term, recurring feedback on resident sentiment. Over time, greater precision in the AI components could lead to reduced dependence on surveys and more cost-efficient sustainability. The tool will provide policymakers with insight into public sentiment about government work and safety, allowing them to respond accordingly (i.e. through programmatic adjustment or additional services). Here, we present our findings from 24 months of study. This final report covers qualitative findings from interviews and focus groups, natural language analysis of Twitter posts from target site residents, and Computer Vision analysis from the same social media.

ETHICS

This study received approval by the institutional review board (IRB) at Columbia University prior to conducting the interviews, community advisory board meetings and harvesting social media data. Prior to conducting the interviews, we also received informed consent from all participants. During the consent process, we presented the aim of this study, both in writing and verbally, including the right to refuse to participate or to end the interview. We de-identified all interviewee names in this report. In addition, to maintain privacy of social media users, we de-identified social media posts, used a password protected annotation system, and had weekly conversations on the ethics of our work.

Qualitative Research Methods & Design

We conducted a qualitative study through one-on-one interviews (1on1) with 69 residents from 10 out of 15 target communities and community advisory board meetings (CAB) composed of 38 NYCHA residents recruited from 9 target communities (see Table 1). These sites represent the pilot neighborhoods where the Neighborhood Navigator will initially focus. We used a Design Thinking framework[1]—divergent, empathetic, exploratory, and generative approach—for both the 1on1 and CAB to gain insight into community members' lived experiences, pain points, challenges, and opportunities.

Table 1: Target Neighborhoods featuring participation totals: Number of participants per 15 target neighborhoods for 1on1 interviews and CAB meetings.

Group	Community	1on1 Participants	CAB Participants
A	East New York	3	2
A	Queensbridge	7	5
A	Stapleton	3	2
A	East Harlem	5	4
B	Brownsville	8	3
B	Bed-Stuy	1	0
B	Claremont	8	5
B	Red Hook	8	7
C	West Harlem	0	0
C	East Concourse	0	0
C	Crown Heights	0	0
C	Rockaways	21	8
D	Canarsie	0	0
D	Washington Heights	5	2
D	Forest Hills	0	0
	TOTAL	**69**	**38**

[1] Design Thinking is a problem-solving framework that utilizes empathy, divergent-to-convergent ideation and analysis, and rapid prototyping to develop solutions that are human-centered, holistic, inclusive, equitable, and sustainable.

From January of 2021 – December of 2022, SAFElab researchers collaborated with various community-based organizations to recruit residents for qualitative interviews. Additionally, participants were also recruited through a snowball and purposive sampling recruiting technique. Snowball strategies included participants sharing the contact information of other residents in their network with researchers who might be interested in participating in this research. Purposive sampling strategies consisted of formal and informal invitations to participate in the research, such as emails, phone calls, and presentations during community meetings and events.

We conducted 69 in-depth interviews using a semi-structured interview guide with four sections exploring resident's perspectives regarding: (1) life in their neighborhoods (2) the impacts of COVID-19 (3) the impacts of Black Lives Matter and (4) use of social media. The interview guide included questions like: (1) *How has your experience in New York City Housing Authority (NYCHA) changed since you first moved in?* (2) *What emotions has COVID-19 brought up with your family? Your neighborhood? Your neighbors?* and (3) *Can you tell us about your thoughts regarding Black Lives Matter?* MSW and PhD social work students conducted and recorded these virtual semi-structured interviews on Zoom. At the time of the recorded interviews, we obtained verbal consent and collected demographic data, including age, race, gender and number of years living in the community. The interviews lasted between 45 and 60 minutes and we compensated participants with a $50 gift card for sharing their experiences.

The average age of the interviewees was 39-years-old. 73% of the interviewees identified as Black and 14% identified as Latinx. The average number of years participants lived in their communities was 20, with the range of years living in the community being one to 63. 60% of the interviewees identified as female, 35% identified as male and 5% identified as non-binary.

One-on-one Interview ("1on1") Findings

For the 1on1 interviews, we asked residents to share their thoughts and experiences regarding quality of life, the built environment, including housing; community and personal safety and well-being; social media and its role in the community; the Covid-19 pandemic; and Black Lives Matter. We used semi-structured interviewing for the 1on1 sessions and used the following questions to guide the interviews:

1. What emotions has COVID-19 brought up with your family? Your neighborhood? Your neighbors?
2. Can you tell us about your experience(s) at NYCHA? What are some concerns you have from living in NYCHA housing? What are some things that you like about living in NYCHA housing?
3. Do you follow or communicate with people in your community through social media?

4. How has Black Lives Matter impacted you? Your family? Your neighbors and neighborhood?

We identified four central themes in the 1on1 interviews: community resources, livability of the built environment, critical reflections on community dynamics, and community recommendations to reduce violence.

Theme 1: Community Resources

“Community resources” are defined as both existing community assets, goods, and services, as well as those that are not currently accessible, but would enable communities to thrive.

Throughout the interviews, residents were aware of their communities’ strengths. For instance, Joe[2], a 20-year resident of NYCHA in Brooklyn explained that in his community, “The people know what they want and they're working towards it. They might be afraid of change, but they're moving into it.” Residents also identified community resources that needed to be added or improved. Natalia, a 37-year-old resident in Queens, explained residents’ frustration with accessing events and other resources, “I feel like sometimes we have things, but the focal point, it's not us. I don't even go because it's like why bother? I’m probably not going to get anything or I’m not going to be able to take advantage of that resource.”

Theme 2: Livability of the Built Environment

“Livability of the built environment” is defined as the physical spaces (e.g. buildings, infrastructure, playgrounds, and parks) and their impacts on health and safety.[3] Residents suggested improvements, which, if implemented, would support thriving communities.

The interviews reflected that residents enjoy some aspects of their built environments, however, the list of issues negatively impacting the built environments was extensive. Residents identified acute sanitation problems, long waits for maintenance requests, and broken elevators and doors as top priorities. Residents also mentioned gun violence and inadequate policing. Stephanie, age 37, stated, “If we don't pay rent on time they bring it to court, so I can't see through that. You're paying your rent and then they don't want to keep up a standard for you to live in it. It’s terrible to get them to fix the place. Honestly that's the worst part of living here. They don't clean the hallways. I bought a broom and a mop for my hallway. I clean out there because you have someone, a friend visiting you or whatever, and the place is so untidy. It's not presentable for somebody to visit because the hallways are dirty."

Theme 3 Critical Reflections on Community Dynamics

[2] All names used within this report have been changed to protect the confidentiality of our participants

[3] https://www.sciencedirect.com/topics/engineering/built-environment

“Critical reflections of community dynamics” describes how residents experience that their communities are in perpetual exchange with external forces outside of their borders. Since the city is interconnected, ideas, trends, and emotions diffuse across communities. Residents are affected by these exchanges, and often the impact on the communities involved is uneven. An example of inequitable exchange is when outside economic forces threaten community stability. A 23-year-old resident of Far Rockaway expressed uncertainty about changes happening near her community, “I'm kind of nervous because if they're putting all these buildings up and the trend is to change everything, who are they trying to kick out because of it?”

Theme 4 Community Recommendations to Reduce Violence

Residents discussed ways to reduce violence and increase safety in their communities with both direct recommendations to address the built environment as well as more indirect ideas to address systemic social problems.

Direct recommendations aim to decrease violence and increase safety in the community through immediate preventative strategies such as upgrading the physical security measures of the building. An example of a direct recommendation comes from lifelong NYCHA resident, 23-year-old Nicole, “For most of our buildings, the doors are just broken, so people can come in and out as they please. So that's a big issue for me, just security. There're always people that you don't know that aren't supposed to be in your building, and it can be dangerous.” Many residents resonated with the need for better security as they demanded the broken doors and disabled security cameras be repaired.

Additionally, residents also highlighted the need to build connections among neighbors to look after each other and bring back systems in which residents volunteer to take turns monitoring the security of their buildings. Another long-time NYCHA resident, 64-year-old Adam said, “When I was growing up, they had tenant patrol. My mom used to do that. There used to be somebody downstairs sitting at a table, and for anybody that came in, they had to sign the book and put their names down, but now there’s none of that. It's like, come on in. I remember each tenant used to take turns for 24 hours to monitor the building.”

Residents also proposed indirect recommendations to address systemic issues that have a long-term impact on safety and community wellbeing. In particular, residents highlighted the correlation between crime rates, the lack of resources, and issues such as unemployment and unstable housing in their neighborhood. They proposed the need for a wide range of programming for better safety and growth in the community. A lifelong NYCHA resident, Quintin, age 30, expressed that his community needs more affordable housing and programs to get the youth introduced to various careers such construction and culinary arts. He stated,

> *“the biggest concern I have for my neighborhood and that I would like to see change is [the addition of] more affordable housing. I'm talking about people owning their own*

> *houses. There's no program for that. You want to build [construction], culinary arts? Those are the things I would like to see more. See, all these guys need to be doing something, good work, good jobs, and decent jobs. It'll keep these guys, you know, all of us as a community, out of negativity. And this will be a positive and growing neighborhood. But if there's no jobs and no housing, people can't see themselves growing. It's just chaos bro."*

A number of residents echoed these sentiments and shared a strong desire not just for better career development, but also entrepreneurial business ownership opportunities, mentorship and training. Other residents stressed mental and emotional health as a means of reducing violence. Tanya, 39, detailed how young people in her neighborhood need clinical services as they navigate safety concerns in the community,

> *"there's been so much gun violence. So, before the children go to school, as they're waking up, they can hear the gunshots. And then, that's what is on their mind the whole day at school. The kids see all this [violence] there, a lot of them don't have therapists that they can talk to, and a lot of them can't talk to their friend about this because the friend is going through the same thing. So, there's no help, all of this is weighing on the brain. Then they get to school, they have to worry about passing grades and all of that. We need to be helping them more mentally, emotionally. When they are mentally and emotionally good, they can prosper, and they can blossom."*

This holds true for other residents as well, as they also resonated with the need for programs for both the mental and physical wellbeing and development for the youth in the community.

Community Advisory Board Findings

Overview

A series of online Community Advisory Board (CAB) meetings were conducted with 38 NYCHA residents and community members of East New York, Queensbridge, Stapleton, East Harlem, Brownsville, Claremont, Red Hook, Rockaways and Washington Heights, representing 9 of 15 target sites. In total, 38 residents participated in the Community Advisory Board meetings between July 2021 and November 2021. The research team, consisting of MSW and PhD social work students and a user research expert, facilitated the focus groups and each group session lasted 90 minutes.

The average age of focus group participants was 36-years-old. 70% of focus group participants identified as Black and 15% identified as Latinx; 55% identified as female, 40% identified as male; and the average number of years that focus group participants lived in their communities was 36 years.

The CABs explored and gathered additional information from residents regarding community wellbeing, safety, threats, concerns, goals, opportunities, local government, and technology including feedback on the prototype of the Neighborhood Navigator, which was presented by MOCJ partner Velir, a Boston-based design & development agency. We used questions such as (1) *What are some signs that your neighborhood is doing well? (2) How might the city government build greater trust in your neighborhood? (3) What role does technology and social media play in your community?* We also used zoom to conduct the focus groups and compensated participants with a $350 amazon gift card. The compensation was reasonable given that the CAB meetings lasted 90 minutes and took place during weeknights, which may have conflicted with participants' plans or responsibilities. Community residents provided invaluable insight, and their time was honored accordingly. The research team or an outside transcription service transcribed verbatim the audio-recorded qualitative data.

Community Advisory Board #1

Community Advisory Board meeting sessions were held in July and September 2021. Residents were asked to offer insight into community safety, markers of community wellness, trust of local government, and the role and impact of technology, social media, and artificial intelligence within the community.

The key themes that emerged from the first CAB included:

1. Relationship with the built environment and infrastructure, including services that provide residents with maintenance and upkeep, safe places to gather outside, and other essential resources.

“Being Outside”
The majority of residents cited that "being outside" was a key indicator of their neighborhood thriving, e.g., kids playing, residents gathering outside, and community events. A key factor that impacts “being outside” is violence, specifically fear of shootings. Some residents stated that police presence can be a deterrent to violence and provide residents with a sense of safety.

Infrastructure
When discussing infrastructure, one resident stated, “we have this major Sandy recovery going on, where they have completely torn up all of the grounds, there is no grass, they removed 457 trees. They created a horrible heat island. There's fencing everywhere, narrow walkways, and very poor lighting; you can't see around the corner... if that isn’t a recipe for a decrease in safety and an increase in crime, I don’t know what is”.

Cleanliness & Upkeep
Cleanliness, maintenance, and upkeep extend beyond the built environment. Residents associated the cleanliness of their environments with how they feel about themselves:

- *“I have to say that cleanliness is next to godliness.”*
- *"You see a whole bunch of garbage on the floor, housing ain't doing their job; the elevator might be broken; might be urine [in the] stairway; and that is devastating because the younger people... they grow up and they may think that's okay; I mean when your room is clean, you feel more calm, simple things."*

Residents also provided feedback on the accountability of the City to provide services related to cleanliness, maintenance, and upkeep:

- *"Why does it take 100 questionnaires and 60 pieces of documents in order to be able to get this [rent assistance during COVID]. You know that these people are fighting to try to keep their apartments, and the money's there.*
- *“Last week NYCHA changed my neighbor’s stove, they left it in a hallway. I literally had to call several people to find it to find out what's going on and then the next thing you know, somebody put a toilet or toilet seat on top of the stove.”*
- On 3-1-1: *"... what we put into it is what we get out of it. 311 complaints are very important. If we don't give a ticket number then we don't get a repeat. You know what I'm saying. And that's the same thing with a sidewalk being fixed, the tree getting pruned, even a manhole down to that point, that you have to have a 311 number."*

2. Availability and awareness of resources for community members, including educational and technology resources

Inequities in Services

Some residents pointed out that the pandemic illustrated how the government treats residents differently. In response to the pandemic, kids were issued tablets and laptops by city government prompting one resident to state:

- *"These kids needed computers way before the pandemic. And they said there was no money to give a child a computer above 96th Street. So I become very, very upset when you can do everything for everyone below 96th Street but cannot do nothing for people in the South Bronx. To not be able to give a child a computer prior to the pandemic and now suddenly it was like the warehouse was making them, they was able to give them, that was very, very wrong."*

Access to Services in a Digital World

Technology is seen in a variety of devices and roles: access, equitable access, internet, wi-fi, tablets, mobile, social media, disaster relief, cableTV, telephony services, etc. As systems that support residents become automated/digital, access to those systems can be challenging for older/senior residents. Seniors face two challenges: (1) having the proper equipment to access the systems (wifi, device); and (2) usability/navigation of systems.

3. Government relations with communities, including themes of community mistrust, cultural misunderstandings between communities and government officials, and the relationship between communities and law enforcement.

Engage Residents Year-round, Not Just Around Election Time

Residents felt that politicians only engage NYCHA residents around election time for votes. One resident stated,*"we don't see them until they want to get voted in. Once they get voted in, that's it. But, come and see how we're living, and then maybe they can understand what we need, you know., Listen to what we say.*"

Police Presence

Some residents felt that police need to strengthen their knowledge and cultural competency of neighborhood norms and activities. One resident explained, *"just because, you know, there might be alcohol, might be people playing dice, things like that, it's not necessarily a dangerous environment. You know what I mean, it's just a part of the culture and part of the things that we do, you know, just enjoy the summer."*

Other residents raised the issues of the potential harm associated with police presence:

- *"Now, they put in more cops and posts around. It is creating like a combative state between the cops and the community because there is a trust issue between Black and Brown individuals dealing with police forces and authority, because the cops, they say* *they're here to protect us, but then again, they're not really protecting; they're doing more harm than good."*

- *"[C]ops bring irritation to my people in my community, you understand? They don't like you already; you see a lot of stuff going on on the TV with a cop and killed another black kid, a cop did this; a cop did that. That's traumatic. That's traumatic for [us], you understand?"*

An Ecosystem Look at Well Being

There are many factors and dynamics that impact residents' well being. Utilizing an ecosystem model helps us understand how these factors are connected and encourage or inhibit well being. This ecosystem includes the community that surrounds residents; the services that support their needs (or lack thereof); and the larger institutions, policies, and cultural and societal elements (stereotypes, ritual, tradition). Exploring the dynamics of these factors can help identify shifts or changes that need to happen to impact well being, and where effort and attention are most needed.

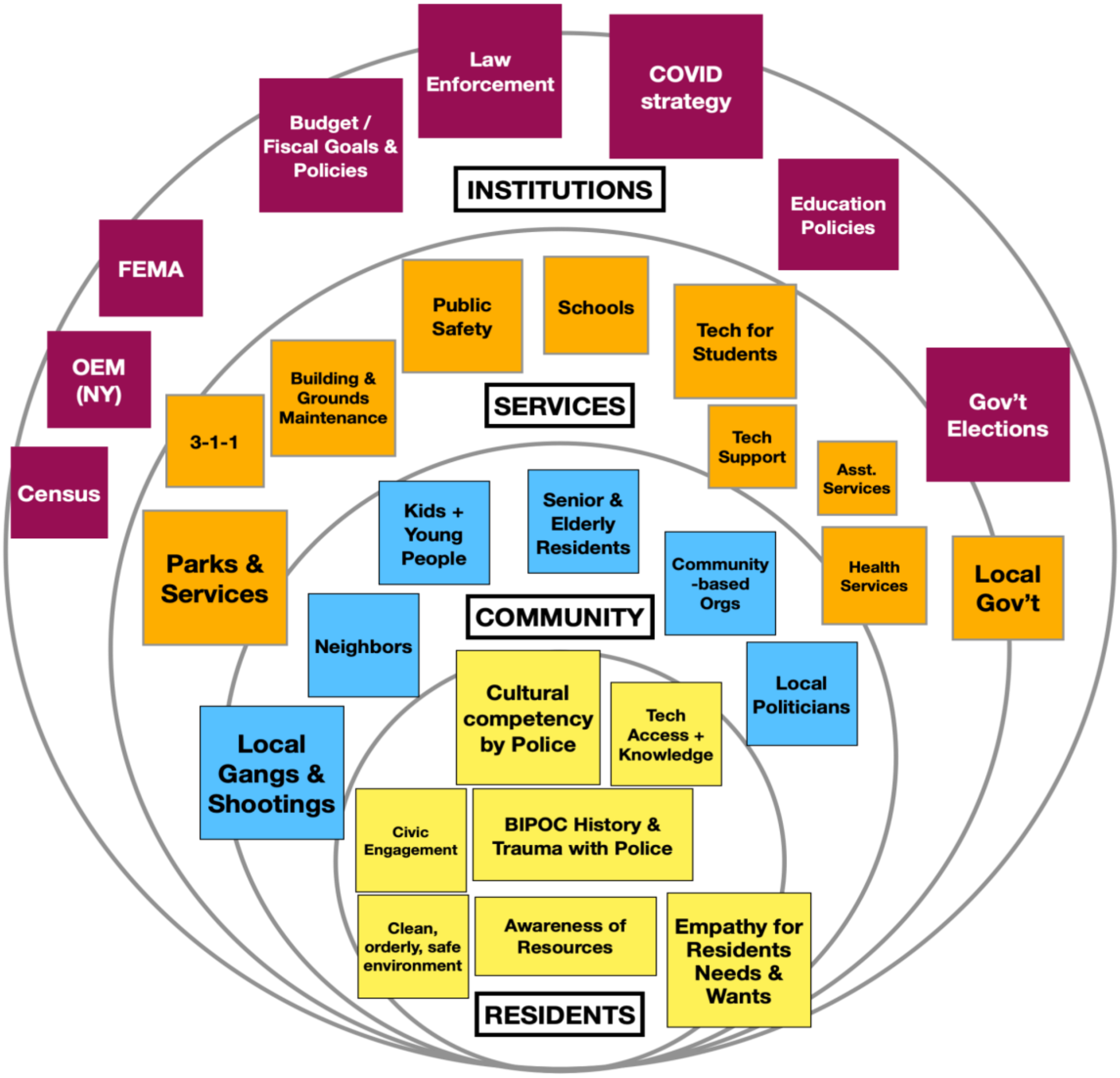


Fig 1: CAB Ecosystem Map: Illustration of factors that impact Residents' well being across three lenses: Community, Services, and Institutions.

RESIDENT - To achieve well being, what does a resident need to know, feel or do? Where is the resident now? How are they feeling?

COMMUNITY - The community includes the household and social networks around a resident. Which individuals or groups play an important role in the resident's life? In what ways do they support or block the resident? What is the power dynamic between this group and the resident? Who has influence or control?

SERVICES - Services include the services and resources available to a resident. Which services does a resident need for well being, safety, growth, etc? For each service consider what kind of access a resident has to this service. Is it good quality? What challenges does the resident face accessing the services? What challenges do the service providers face in delivering the service?

INSTITUTIONS - The systems & policies that influence resident's rights and freedoms make up institutions. What are the rights and freedoms that a resident needs for well being? For each right or freedom consider: Does a resident experience barriers or unequal access to this right or freedom? Which systems and policies support the resident with regards to this right or freedom? Which ones discriminate?

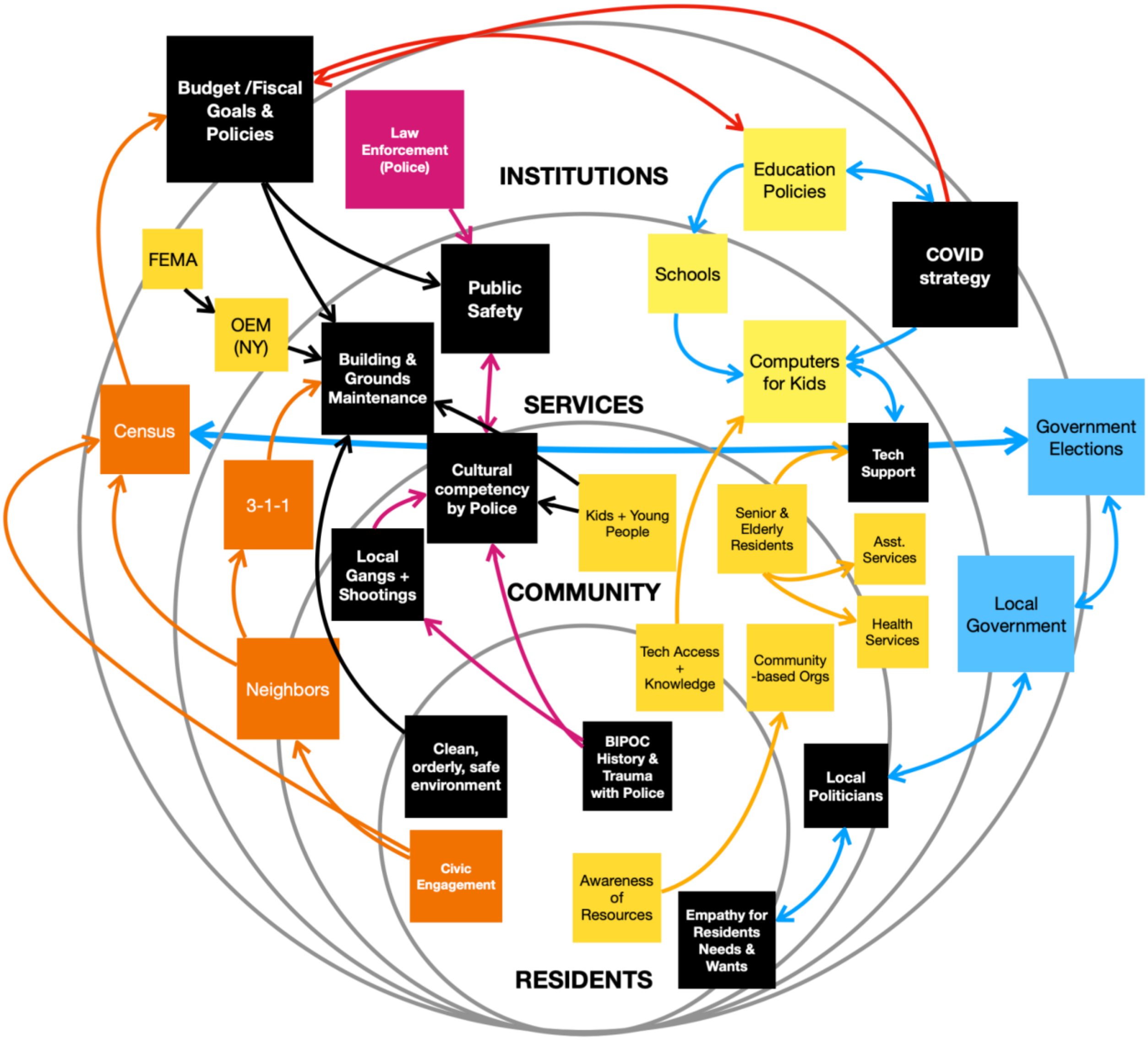


Fig 2 CAB Ecosystem Map: Illustration of factors that impact Residents' well being across three lenses: Community, Services, and Institutions and how they are connected.

Community Advisory Board #2 (November 2021)

An early prototype of one part of the Neighborhood Navigator—a website that showed various metrics to track and index a neighborhood's safety and well being—was presented to community members for feedback at the second CAB meetings. This included discussing some naming options to brand the tool (Neighborhood Navigator).

In the meetings, the Community Advisory Boards also explored the role and impact of the Neighborhood Navigator within the NYCHA communities. They explored how the Neighborhood Navigator might increase safety and inform community improvements; and how it could be used to build trust with communities. We also asked how members would like to contribute information to NYCHA to keep developing the Neighborhood Navigator.

Community members felt the Neighborhood Navigator could be used to better understand what is happening in a community or neighborhood, but that it is not a replacement for on the ground presence from the mayor's office and other government agencies. Furthermore, feedback suggested that the Neighborhood Navigator will be trusted more if changes related to information from the tool are implemented in a timely manner. Community members felt that delays in actually implementing improvements would further erode the relationship with community members as their needs would continue to be ignored. Community members suggested establishing metrics and scoring to track changes and accountability, and to direct and prioritize policies and investment in the built environment of NYCHA.

Reaction to the website demo varied. Some community members asked “is this only going to be a website or is it going to be an app?" and "how is this meant to help the individual who might be struggling in their community?" It was not clear from the demo how community members would input data. Regarding the usability, others thought the user interface design and information architecture were easy to understand and navigate. Some commented on the "localization" of the images being used, which made it feel more representative of their neighborhoods.

Other comments and feedback included:

Neighborhood Navigator Tool was perceived as performative-only

Some residents expressed doubts that information collected from the tool would result in actual, tangible changes based on residents’ feedback:

- "it just seems like even though we had these complaints, and things like that, we're still not being heard. So this is a good way to probably, you know, generate things to be changed, but it's like, we the people that's living in these communities, they say they want to come and help but they don't really do much."
- "Are they really going to use this information?"

Empirical Proof

Community members expressed that they would like to use the Neighborhood Navigator Tool and see results to better understand how it might actually be helpful to them.

Holistic + Inclusive Input & Co-design Approach

Community members felt that to get the most out of any initiative, input must reflect all generations that reside in NYCHA (youth to senior/elderly). Members also expressed concern that technology-based solutions need to also be usable by its most senior or elderly residents. One resident explained, *"my concern is more for the elderly—and for them to have input on what's being done and what changes would they like to see in their developments?"*

Suspicious of motivation

Some residents expressed suspicion of the government or research team's motivations for using the tool, including concerns for how the data and feedback could impact existing residents:

- "Okay, so, here's my issue. Um, let's say this service helps in some type of way, I do not believe that it's really for us. I feel like if [city gov] come[s] into the community and start doing all of this, they're gonna put better schools, and this and this, and all of a sudden people are losing their apartments; the rate of homelessness is going up. 'Why is the rate of homelessness going up?' Oh, those who was the people that was in the houses, the housing developments, that we done improve, now they don't have them anymore. It's like, come on."
- "Gentrification. They want to beautify our neighborhoods, make it look all good. They want to build our neighborhoods up, but at the same token, they're pushing the people that there out because most of the people that live in our neighborhoods is a low income. And, you know, black people, people of color and you want to raise our rent, you want to put these new things here, and that makes it cost more so, how was this going to benefit our people in our community?"

Conclusion for 1on1 Interviews and Community Advisory Board

The 1on1 interviews and CAB meetings reinforced that for an initiative like the Neighborhood Navigator to succeed, it has to be co-designed by community members who are the experts that understand the culture, needs, pain points, challenges, opportunities, and overall ecosystems of their respective communities. While technology (social media, artificial intelligence, etc,) can have a meaningful impact for community members in their built environments, any effort by government agencies and elected officials must: (1) center the voices of those who are directly impacted by the outcomes of the design process; (2) prioritize design's impact on the community over the intentions of the designer (or political expediency); and (3) look for what is already working at the community-level before seeking new design solutions.[4]

[4] Design Justice Network Principles: https://designjustice.org/read-the-principles

In addition to collecting qualitative data through interviews and meetings, our team has conducted computer vision and Natural Language Processing (NLP) research to further inform the Neighborhood Navigator. The following sections describe how we designed and implemented these methods to analyze resident's sentiments about their quality of life and communities through their social media activity.

Computer Vision Methods and Design

Overview

Users express their opinions in multiple ways online. Some users write text describing their feelings or experiences, while others share images to illustrate the points they make in tweets. We wanted to assess whether the themes we described above were also expressed on social media. We also wanted to discover the degree to which these themes were conveyed by text or images separately. While many methods exist for predicting the sentiment (positive or negative feeling), relatively few methods have studied predicting sentiment from multimodal social media posts (i.e. using *both* the image and the text of a single post at the same time to make a prediction).

We believe that leveraging a multimodal sentiment prediction approach is important for several reasons. First, we have evidence that leveraging either modality alone is insufficient to truly understand our dataset. For example, we have seen many examples where users employ sarcasm, expressing a seemingly positive sentiment within the text (e.g. "great job at repairs"), but when the image is considered with the text, it is clear that damage remains and that the user is actually expressing a negative opinion. Second, by training a multimodal sentiment prediction model, our experiments show that our system outperforms existing methods for sentiment prediction that rely on only one modality (text-only or image-only). We thus have focused on building a multimodal sentiment prediction system which can be applied to multimodal (image+text) or unimodal (text or image) data.

In the Methods and Design section, we first describe the dataset we have collected, our automatic training data labeling approach, our model designs, and our data annotation efforts. We then discuss our experimental results that validate our model choice. Next, we use our model to discover statistically significant shifts in sentiment and sentiment trends. Finally, we analyze our model's performance in more detail.

Research Methods

Data collection

In order to analyze the sentiments of by New York City residents, it was necessary for us to assemble a large dataset for analysis and model training. Our group inspected a variety of social media platforms for suitable data. Specifically, we inspected Facebook, Instagram, Twitter, TikTok, and Snapchat. Facebook was appealing because groups had been created for a number of different housing sites. We thus did an initial data harvest from Facebook. However, post content was almost entirely about housing issues at the various sites (e.g. maintenance issues affecting various buildings, sanitation conditions, etc.). This, coupled with relatively sparse data, led us to look at other alternatives.

While we found a few Instagram and TikTok posts relevant to this research, we ultimately concluded that there were too few relevant posts for meaningful analysis. We also briefly collaborated with a research scientist at Snapchat in assessing the feasibility of analyzing publicly posted snap stories. While this data was appealing because of its geolocation data, we ultimately found very few relevant snaps on this platform. Twitter was most appealing to us because users appeared to post about a wide variety of topics, the posts featured images, text, and sometimes video, the data was publicly posted, and data could be readily obtained by scraping.

After choosing to focus on Twitter posts, our group employed a snowball sampling approach to harvest relevant data. First, we identified around 10 thousand unique users who had sent a tweet tagging the @NYCHA or @CrimJusticeNYC Twitter handles within a four-year period. After we obtained this list of users who had engaged with these official government social media handles, we harvested all publicly available tweets (as well as associated images) from those users within the past 10 years. Each tweet contains rich metadata, such as the time and date the tweet was posted, the number of retweets, hashtags, etc. In all, we obtained around 12 million text-only tweets and around 500,000 multimodal tweets (tweets containing both images and text). For comparison, the data we harvested from Facebook and Instagram pages associated with particular NYCHA housing sites is much smaller and consists of only a few thousand posts. Because this data was too sparse to draw significant conclusions, all of our analysis is conducted on our large-scale Twitter dataset.

Automatic data labeling

Training modern, state-of-the-art multimodal models requires a large amount of training data. We explored relevant literature to search for a suitable training dataset consisting of multimodal social media posts annotated with the sentiment of the post, which we could use for training. However, we were unable to find any dataset that met our model's training needs in terms of the number of samples or the type of training data (social media posts). We thus created our own *automatically* labeled dataset to train our model. Because of the large amount of data needed for training, acquiring human annotations on enough training data is not practical. We thus obtained multimodal sentiment annotations on our dataset automatically for free. To do so, we leveraged an existing text-only sentiment model trained on Twitter data to predict the sentiment of each tweet within our dataset, using only the text.

Next, we trained a state-of-the-art image classification model using an existing dataset created by our group containing adjectives and noun phrases which are predictive of sentiment ("dirty house”). Once we trained this model, we applied it on all of the images within our Twitter dataset to obtain predicted adjective-noun pairs. We then feed these sentiment-invoking adjective noun-pairs into the same text-sentiment prediction model that we used to predict the sentiment from the Tweet text to obtain a sentiment prediction from the image. We then obtained a text sentiment prediction and an image sentiment prediction. We looked at the cases where the sentiment prediction of the image and the text both agree and use these examples to train our multimodal sentiment model.

Multimodal sentiment model

We trained a multimodal sentiment prediction model using the automatically labeled data described above. Our model is a modified version of ViLT (Kim, et al. 2021) trained to predict the sentiment label of an image and text post. To ensure that the model looks at both modalities, we performed random masking of both the image and the text (essentially hiding portions of the image or text during training), which forced the model to look at both modalities in order to make correct predictions. Because our training data is automatically labeled and the label may be incorrect in some cases, we employed a number of techniques for training on this type of data (called weakly-supervised training) to minimize the impact of incorrect training labels.

Data annotation efforts

In order to evaluate how accurate our system is at making sentiment predictions from social media data within our dataset, we needed some human annotated data to compare our system’s predictions against. We created two annotation interfaces. In our earlier work, we collected several hundred annotations on a set of tweets that mainly focused on maintenance issues in NYCHA buildings. However, this set of tweets only covers a few narrow topics of interest. Our large-scale tweet dataset contains tweets expressing sentiments towards a much broader set of topics.

In order to assess our system’s performance at predicting sentiment on these topics (many of which are more abstract and complex than the maintenance issues tweets), we created a second, improved annotation tool. The new annotation tool allows us to capture much finer-grained annotations with significantly lower annotator effort. The annotator is asked to specify the topics the tweet expresses a sentiment towards, identify image regions corresponding to the sentiment being expressed for each topic, identify words in the tweet text corresponding to each topic, and specify the sentiment being expressed towards each topic. These fine-grained annotations are important in order to assess how human-like our sentiment prediction model’s inference process is. We recruited a number of annotators for this task, including a community member living in NYCHA housing. This has the added benefit of bringing community-knowledge into our work in order to properly interpret tweets which may reference events happening within the community, which use local dialect, etc.

Experimental Results

Text-only	**Image-only**	**Multimodal (Ours)**
68.X%	64.X%	72.X%

Table 1: Sentiment model accuracy on human-annotated data. The model is trained on a three-way classification task of predicting whether the post is positive, negative, or neutral.

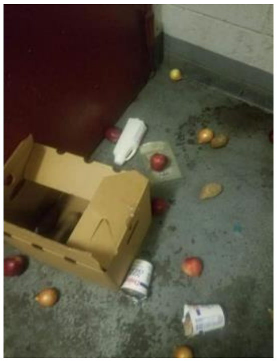
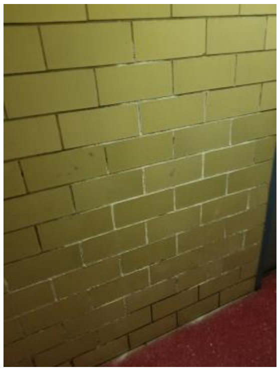
 “NYCHA residents showing their appreciation for where they live and also for the food they take from donation”	 “Gotta love #Manhattanville @NYCHA @NYCMayor #NYC #newyorktough #nycha”
Text prediction: Positive Multimodal prediction: Negative	Text prediction: Positive Multimodal prediction: Negative

Figure 1: The importance of multimodality in understanding social media posts. In both cases, the text expresses a positive sentiment, but when the image is considered, the sentiment is actually negative due to sarcasm.

Sentiment model performance

In order to assess whether our sentiment model is outperforming existing literature, we compared it to a number of state-of-the-art methods which use only a single modality. Specifically, we compare our method to a recent state-of-the-art text-only model trained on ~58 million tweets and fine tuned for sentiment prediction on the TweetEval benchmark[5]. For the image-only model, we train a vision model on the SentiBank[6] benchmark developed by our group, a set of images crawled from FlickR and labeled with sentiment scores. We evaluate the accuracy of each method at predicting the sentiment label of our annotated sentiment data and present our results in Table 1 above.

[5] Barbieri, Francesco, et al. "TweetEval: Unified Benchmark and Comparative Evaluation for Tweet Classification." *Findings of the Association for Computational Linguistics: EMNLP 2020*.

[6] Borth, Damian, et al. "Sentibank: large-scale ontology and classifiers for detecting sentiment and emotions in visual content." *Proceedings of the 21st ACM international conference on Multimedia*. 2013.

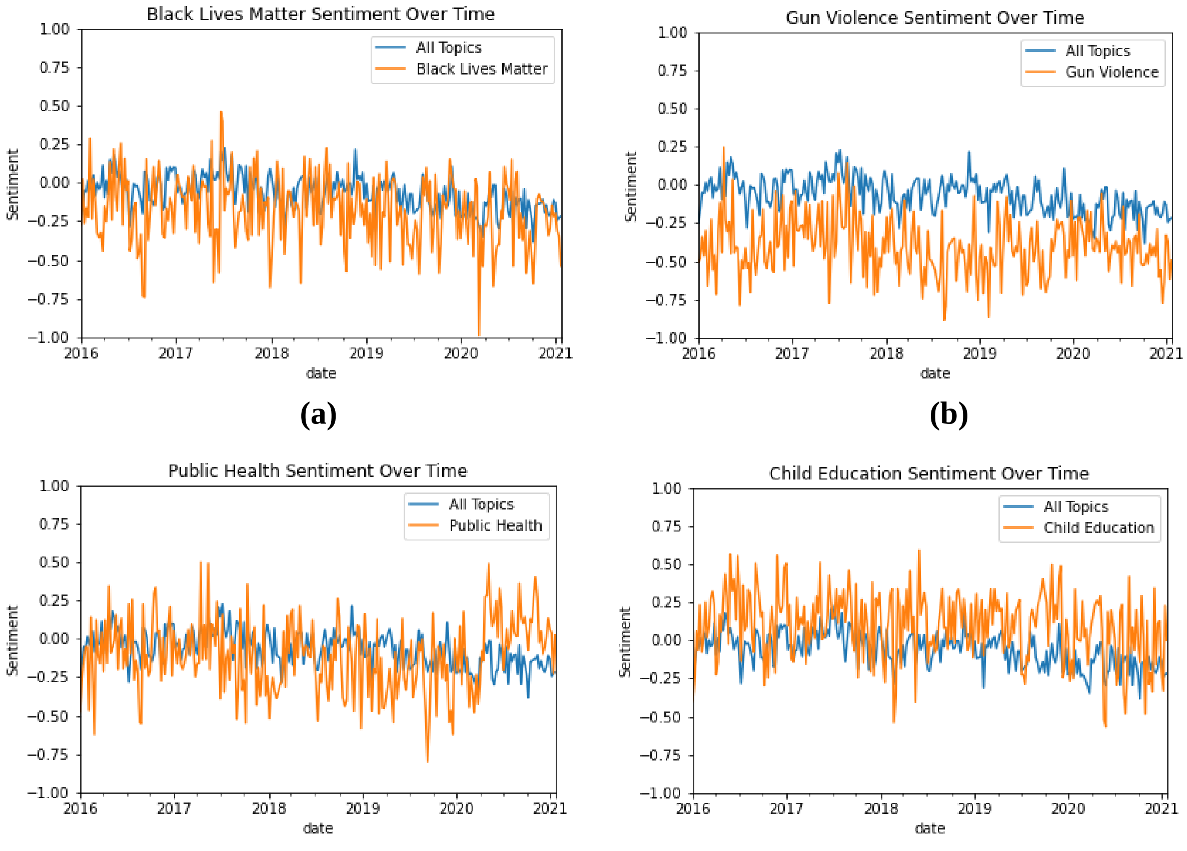
We observed that our multimodal sentiment model outperforms the best-performing single modal baseline by around 4%. While a 72% accuracy may appear low, we analyzed the model's failure cases. Specifically, the vast majority of errors are "soft" errors, where a sample is confused with a compatible label. For example, a sample labeled "neutral" may be confused with positive or negative and vice versa. These types of errors are often subjective in nature. For example, an annotator labels a factual report of a negative event as neutral or vice versa. In contrast, we observed very few "hard" errors, where positive or negative samples are confused with each other. In Figure 1, we showcase examples of the importance of the multimodal model. We observed two posts that our best-performing single modal model (text-only) got incorrect because it only considered the text. However, when the image is considered, one understands the text is actually sarcastic and the sentiment is negative. There are a number of examples like this within our dataset and our model is better able to handle such cases.

Sentiment trends

After validating the performance of our model on our annotated dataset, we applied our model to the dataset we harvested. We assigned the positive sentiment label +1, the negative sentiment label -1, and the neutral sentiment 0. By assigning a numeric value to each class, we were able to calculate the average sentiment score across time periods in order to discover sentiment trends. We then used the timestamp of each post to bin the posts by week. Thus, we calculated the average sentiment per week across the entire dataset and plotted the results over time. This allowed us to compute a "baseline" sentiment in the dataset overall across time for statistical comparison purposes. We observed, for example, that the overall sentiment of the dataset tends to center around 0 (i.e. neutral). We also observed a decay in overall sentiment starting around the beginning of 2019, i.e. the posts within the dataset tended to express a more negative sentiment overall. While the overall sentiment trend is interesting, drawing finer-grained conclusions is difficult because it encompasses posts about a variety of different issues within our dataset.

In order to draw finer-conclusions, we leveraged the topic information of the post. Specifically, we used two types of topics: 1) we used the indicators predicted by the natural language processing team (described in the following section), and 2) we performed topic discovery within our dataset to discover topics beyond those covered by our indicators. In order to discover new topics not covered by the indicators, we clustered the tweets and used an automatic cluster naming technique[7] to arrive at a set of topics.

[7] Grootendorst, Maarten. "BERTopic: Neural topic modeling with a class-based TF-IDF procedure." *arXiv preprint arXiv:2203.05794* (2022).



(a)

(b)

(c)

(d)

Figure 2: A sample of our topic-specific sentiment trends. We compare each topic's sentiment trend line to the overall sentiment trend within the dataset to discover trends which are statistically significant.

We next looked at trend lines for each topic. Specifically, we computed a trend line for each topic by taking the average sentiment score for each week for all posts labeled with that topic. We then compared these sentiment trend lines to the overall sentiment trend. We include a sample of these results in Figure 2. In Figure 2(a), we see the sentiment trend line for posts labeled "Black Lives Matter". From this, we observe that the most negative sentiment in five years occurred in early 2020. This period of particularly negative sentiment coincided with the protests in response to the killing of George Floyd and the Breonna Taylor and Ahmaud Arbery killings. In Figure 2(b), we show the trend line for "gun violence". We observe that posts labeled gun violence tended to have the most negative sentiment on average of any topic we studied, with the majority of the gun violence trendline being significantly lower than the overall sentiment trend line. On only a few occasions does the trend line ever cross into the positive territory, and when it does, it is only temporary. We observed that public sentiment became particularly negative regarding gun violence after early 2018, which coincided with the Stoneman Douglas High School shooting.

In Figure 2(c), we show the sentiment trend line for public health tweets. Surprisingly, we observe that the average sentiment for public health tweets *increased significantly* slightly after the onset of the pandemic. One explanation of this was that public sentiment was, on average, supportive of public health during this time by, for example, encouraging masking, supporting healthcare workers. Finally, in Figure 2(d), we show the sentiment of tweets towards childhood education. We observed that sentiment towards this topic tends to be more positive than the overall dataset, but, after the onset of the COVID-19 pandemic, the sentiment tended to become progressively more negative than typical, achieving the lowest average sentiment across five years in mid-2020.

Topic	**t-statistic**	**p-value**
Black Lives Matter	9.202	**1.106e-17**
Building Complaints	12.730	**2.878e-29**
Child Education	-13.852	**3.619e-33**
Community Safety	-5.665	**3.828e-08**
Domestic Violence	-0.438	0.662
Gun Violence	31.566	**1.332e-91**
Homeless	13.490	**6.665e-32**
Local Businesses	-9.464	**1.729e-18**
Nonprofit Organizations	-33.497	**4.724e-97**
Public Health	3.000	**2.957e-03**

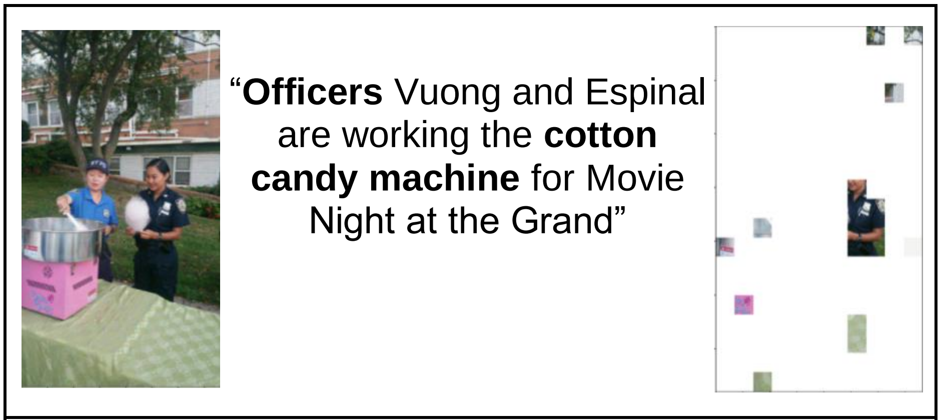
Table 2: Statistical significance of topic-specific trend lines (across all time periods) vs overall trend line for a set of topics within our dataset.

In order to assess whether topic-specific sentiment trend lines were statistically significantly different from the overall sentiment trend line, we performed a paired t-test, comparing the average sentiment of the entire dataset for each time period with the sentiment for each trend line. We show our results in Table 2 above. One interesting observation is that all topic-specific trend lines significantly differed from the overall trend line, except for domestic violence. One possible explanation of this is that posts labeled domestic violence are not as accurately identified, resulting in a set of posts labeled domestic violence that does not substantially differ from the overall trend within the dataset. Another possibility is that particular cases of domestic violence are not discussed as frequently online, meaning many domestic violence tweets are pointing to social resources (and thus labeled neutral) rather than expressing a clear sentiment. Similar to our visual analysis above, from the t-statistics, we can observe that gun violence tends to have a much lower sentiment overall than the overall trend line (31.566), while nonprofit organizations tend to have a much higher sentiment than the overall trend line on average (-33.497).

We include sentiment trend plots and t-statistics for each trend found within our dataset as supplementary material to this report. We also include sentiment trend plots and statistical tests for each indicator topic (e.g. economic readiness) in our supplementary material. In addition, we perform *fine-grained* statistical tests, for each time period. That is, for each topic and for each week of data, we test whether the sentiment for that time period is statistically significantly different from the overall sentiment for that week. This allows us to more precisely identify when significant shifts in sentiment occur.

Explainability

In order to assess the robustness of our sentiment model and mitigate potential biases, we wanted to understand what cues our model has learned to use to make its predictions. By understanding what portions of the data our model relied on to make its predictions, we are better able to assess whether 1) our model is paying attention to image regions that are not semantically sensible (i.e. background, blank spaces, etc.), which would indicate a failure to learn a generalizable and robust model; or 2) our model is regularly focusing on concerning areas of the image (i.e. skin color). To this end, we leveraged explainability techniques that visualize which areas of the image our model relied on most when making its predictions.



“**Officers** Vuong and Espinal are working the **cotton candy machine** for Movie Night at the Grand”

Predicted label: positive

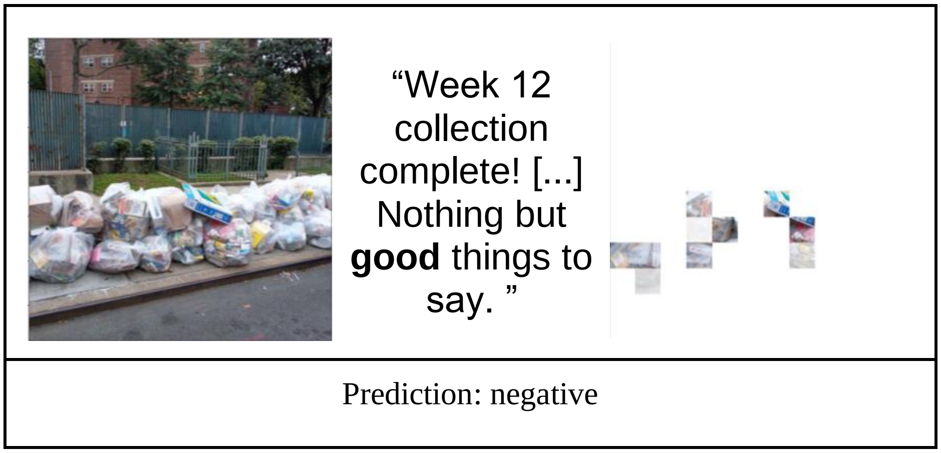

"Week 12 collection complete! [...] Nothing but **good** things to say. "
Prediction: negative

Figure 3: Multimodal sentiment explanations produced by our model. We visualize the words (in bold) and patches our model paid most attention to in making its predictions.

In Figure 3, we show a few example multimodal explanations produced which explain our model's predictions. We observed that while our model does pay attention at times to facial regions, it appears more focused on facial expressions than the race of the individual. For example, in the first post, we see the model focuses on the fact that the person is wearing a police uniform and only the part of the individual's face showing the expression (i.e. the smile), rather than the entire face. Textually, the model focuses on the words "officers" and "cotton candy machine" to inform its prediction.

In the second post, our model incorrectly predicts that the post is negative, perhaps believing it is sarcastic. The model focuses on the word "good" in the text, but focuses on the bags of garbage in the image. From this, the model concludes the post is conveying a negative sentiment, when in actuality the poster is praising garbage service. Such mistakes (between negative and positive and vice versa) tend to be rare within our dataset, but our explanations help understand the model's reasoning.

Concluding remarks

We showcase a few key results in this text, but include complete results, figures, and statistical tests (both overall and by week) for all topics and indicators in our dataset. We include example multimodal social media posts for each topic in our supplemental materials provided with this report. We also explored an extensive number of techniques beyond those discussed here.

Briefly, we explored methods for detecting text in images and leveraging it as an additional feature for sentiment prediction. In practice, we found that the text in the images was usually fairly unhelpful and contained content that was distracting to our model (e.g. street signs, text on vehicles, ads, etc.). We also explored an approach to complement our sentiment model which explicitly detected facial expressions in the images, such as anger, fear, disgust, happiness, etc.

We discovered that while indeed smiling faces tended to indicate positive sentiment, the other facial expressions were not strongly predictive of the overall sentiment of the post and caused worse performance. Finally, we explored techniques for predicting the “intent” of the post. We wanted to explore whether certain categories of posts (e.g. building maintenance, gun violence, etc.) intended to provoke the reader into action, persuade them into taking a position, etc. We observed promising results in this direction, but our method requires further refinement and validation before relying on its predictions.

Natural Language Processing Methods and Design

We aim to understand how residents of different New York City housing communities , neighborhoods and boroughs are feeling about life in the city, specifically related to the three central themes - community resources, livability of the built environment and community dynamics - focusing on fine grained aspects like public safety, housing, transit and other quality of life indicators. To do so, we used Twitter data Since Twitter is predominantly text, we explored many Natural Language Processing (NLP) techniques to: (1) automatically discover ways to associate users on social media with different communities or neighborhoods; and (2) cluster social media within the MOCJ wellness-indicator hierarchy.

In this section, we describe these Machine Learning models and techniques developed to collect tweets from users, discover their communities and assign them their relevant indicator topics. The models developed include a geotagger and a hierarchical indicator discovery (clustering) system.

The dataset for this tool included more than 12 million tweets from 3000 users and was collected by snowball sampling tweets from users who mentioned NYCHA at least once between 2016 and 2020. Further research and analysis were done on this dataset as described in the following sections.

Geotagging

To discover trends and patterns across communities, we need to know the targeted location for each tweet. However, it may be impossible to discover this information using just a particular tweet, especially if it does not contain any geospatial information. For example, a tweet that says “Don’t you want to keep tenants safe! I live in Baruch Housing and I just passed 3 workers walking around with no masks on!” explicitly states the targeted community (Baruch), whereas a tweet that states “Why would you shut down the water in our houses? At a time like this and with no warning!” does not. In this case, if both the tweets are posted by the same user, we can claim, with a good degree of confidence, that the second tweet also targets the Baruch community.

Therefore, we associated users on Twitter (from within our corpus) with the 15 target areas provided by MOCJ and tagged all tweets from the same user to be targeting their assigned community. We did this in two stages - first, we assigned users to NYCHA communities, then we assigned these communities to target sites (Fig. 1). As there are almost 291 discovered NYCHA communities that are more geographically-specific than the target sites, this allows easier and more precise location identification (especially as they are also more commonly mentioned in the tweets than the sites).

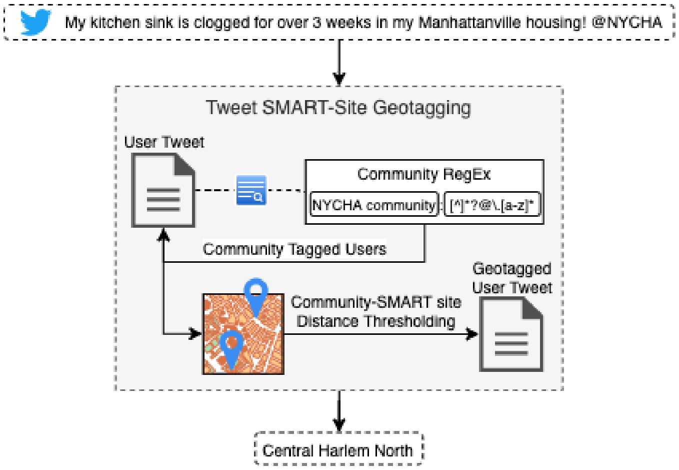

My kitchen sink is clogged for over 3 weeks in my Manhattanville housing! @NYCHA
Tweet SMART-Site Geotagging
User Tweet
Community RegEx
NYCHA community
[^]*?@\.[a-z]*
Community Tagged Users
Geotagged User Tweet
Community-SMART site Distance Thresholding
Central Harlem North

Fig. 1: Geotagging - discovering target sites, given an user tweet.

To perform the community tagging, we used regular-expressions (rules to search for particular predefined patterns in text). We identified and extracted key terms related to the 291 communities from the tweet content or hashtags including their names, abbreviations, cross streets and addresses. This was followed by aggregation of confidence scores (assigned by counting the number of string groups matched per user and taking their majority voting) to determine the community that was the most strongly associated with the user.

We then mapped the communities (and in-turn the users) to the target areas using their geographical locations by first matching their respective boroughs (as the 15 sites are well distributed among the boroughs), then finding the distances of communities from all sites within the borough, assigning them to the nearest site. To limit the number of communities being mapped, we empirically determined distance thresholds by looking at the histogram plots of the number of communities within a particular distance from sites in each borough (Fig. 2).

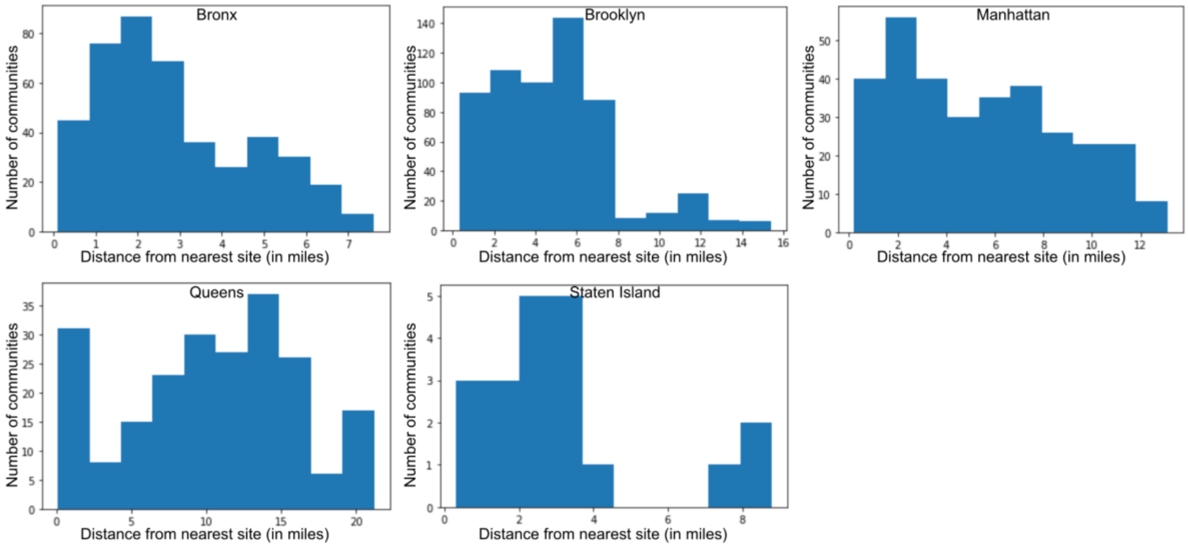


Fig 2: NYCHA community to target site mapping - Borough distance histogram

Hierarchical Indicator Discovery

In order to provide short-term, recurring feedback, and provide policymakers with insight into public sentiment, it is crucial to understand what the public is talking about. To do this, we study the three central themes - community resources, livability of the built environment, and community dynamics - while focusing on fine grained aspects like public safety, housing, transit and other quality of life indicators. Specifically, we look at the NIS Safety and Thriving Indicator Hierarchy, provided by MOCJ. The hierarchy is depicted in Fig. 3 where three of the domains are framed through an economic lens, and the other three encompass the areas that residents identify as important to community safety and thriving. The hierarchy consists of three levels of granularity.

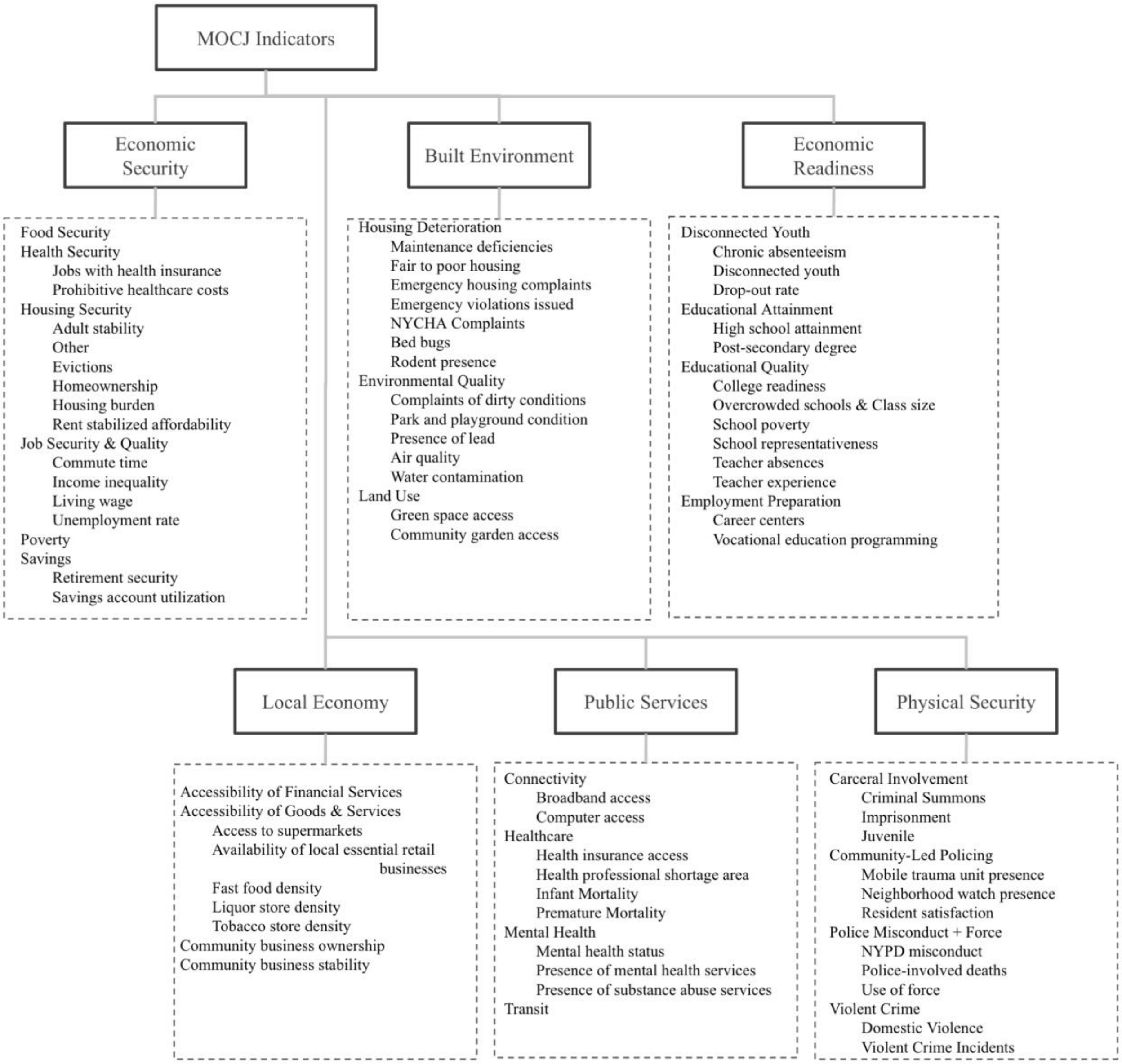


Fig. 3: MOCJ Indicator Hierarchy - 3 levels of granularity

To automatically discover and cluster tweets within the various indicators in the hierarchy, we studied Seeded Hierarchical Clustering (SHC) and designed a novel algorithm to solve it. In SHC, the aim is to automatically fit unlabeled data to predefined hierarchies using only a small set of labeled examples. Generally, working with expert-crafted predefined hierarchies, such as this, is challenging for NLP algorithms. This is because these hierarchies are hand-crafted by domain experts to explore particular areas of focus, and this causes them to be *unbalanced* (with subtopics that over or under represent one aspect of their parent topic) or *incomplete* (with subtopics that are only partially enumerated). Moreover, while working with diverse corpora (e.g.: tweets used in our study), the hierarchy may not fully *represent* all tweets. Additionally, annotating sufficient tweets for each topic in the hierarchy to train a model may be expensive.

These constraints and challenges make SHC a difficult task, and thus we proposed a novel weakly-supervised algorithm, HierSeed, which uses a user-defined topic hierarchy and only a few labeled examples to assign tweets from the much larger unannotated corpus to individual topics (Fig. 4). We also showed that it is highly efficient and outperforms all existing solutions on three real-world benchmark datasets from different domains. We additionally evaluated HierSeed on a randomly sampled set of tweets manually annotated with MOCJ indicators, which provided us with further evidence that it works well for this specific application.

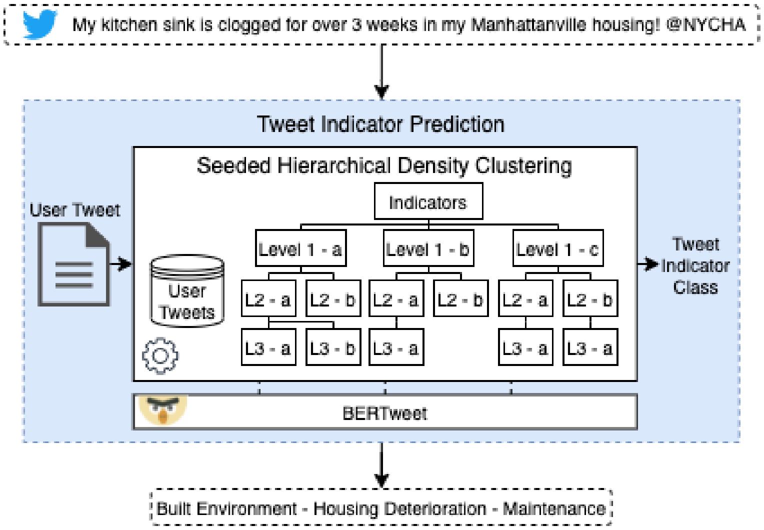

My kitchen sink is clogged for over 3 weeks in my Manhattanville housing! @NYCHA
Tweet Indicator Prediction
Seeded Hierarchical Density Clustering
User Tweet
Indicators
Level 1 - a
Level 1 - b
Level 1 - c
User Tweets
L2 - a
L2 - b
L2 - a
L2 - b
L2 - a
L2 - b
L3 - a
L3 - b
L3 - a
L3 - a
L3 - a
Tweet Indicator Class
BERTweet
Built Environment - Housing Deterioration - Maintenance

Fig. 4: Architecture of the SHC pipeline using our algorithm, HierSeed.

Weakly-supervised methods are better suited for SHC than unsupervised or supervised techniques. Though unsupervised techniques automatically discover latent structure within a text corpus without any supervision, they are difficult to integrate with user-defined hierarchies. Supervised methods avoid this issue, but are usually very data intensive. HierSeed makes use of the best of both worlds by incorporating weak supervision using only a few example tweets per topic, and discovering the structure of the corpus in an unsupervised manner, in order to match it as closely as possible to the structure of the hierarchy.

The technique starts off by using a language model designed for social media (BERTweet) to embed the tweets into a coordinate system such that each unique tweet has a distinct position within the high-dimensional space, and similar tweets are closer to each other than to the dissimilar ones. Then, we created three to five representative tweets for each indicator topic within the hierarchy. For example, for the Economic Security indicator, we created three tweets for the Savings subtopic at level two, and three to fourtweets for each subcategory at level three, as shown below:

Economic Security -- *(L1)*

…

Savings -- *(L2)*

- "I spend what I earn in a day. No idea how to save up for my future."
- "Finally, some great news surrounding our student debt crisis. Biden administration cancels $500 million in student debt"
- "I don’t have a health savings account but I think they are a great investment tool for folks that are generally healthy and have access to one."

    Retirement Security -- *(L3)*

    - "I wish I could save up just enough to sustain myself when I'm old. Better hopes for the future."
    - "Association delegate on his well deserved retirement after 32 years!!! Wishing you a happy and healthy retirement!!!"
    - "Retirement security ‘is shakier than ever’ and ‘Americans are not saving enough’ for old age"

    Savings Account Utilization -- *(L3)*

    - "My bank account is in two figures and I'm literally broke."
    - "Personal savings rate dipped to 9.4% in June, down from 10.3% in May … now at lowest since February 2020"
    - "I signed up for family dollar savings"
    - "Thought to myself how sad it is that even if you got out of debt & started saving, a typical savings account will only give you less that 2% return "

…

Next, these representative tweets were also embedded into the high-dimensional space to obtain numeric representations, and each topic/subtopic’s initial representation was computed by taking the average of the tweets belonging to it. Now, the idea was to assign unlabeled tweets that are closer to a topic in this space to that particular topic. Since the topics are arranged hierarchically, we first assigned tweets to the top-level topics, followed by distributing them among the subtopics based on their distances.

However, we observed that the lower-level topics/tweets (i.e.at level three) are more specific and cohesive. Higher level topics (i.e. at level two) on the other hand are more generalized. This led to problems since the top levels are not representative enough and many tweets that were actually close to subtopics ended up not being assigned to them as the parent topic was far away. We solved this by updating and setting the parent topic representation to a weighted average of - itself, the centroid of the children subtopics, and the center of the Largest Empty Sphere (LES) within the subtopics (see Fig. 5). The centroid captures the density information by going closer to the closely packed subtopics. The LES captures the sparsity information by searching for the largest sphere containing no subtopics within itself. This is desirable as it is as far as possible from all subtopics, but not too far from any particular subtopic.

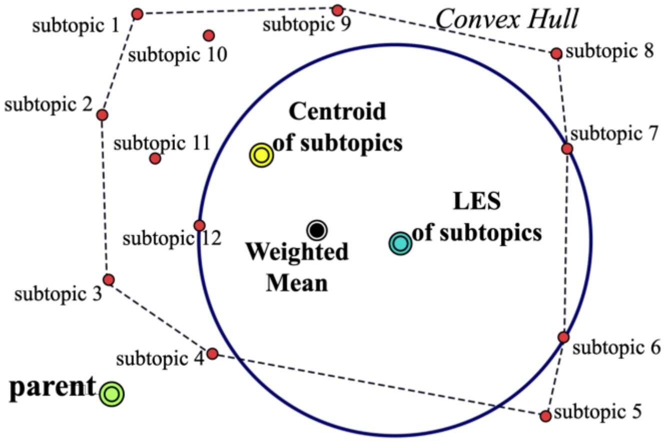
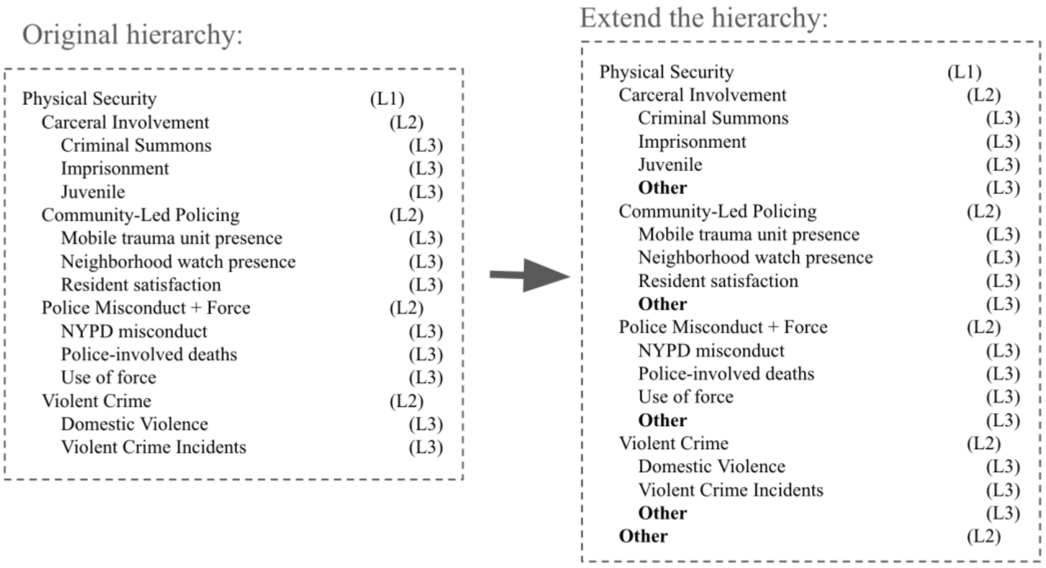


Fig. 5: Updating the topic representation using the centroid and LES of its subtopics.

Another problem is the incompleteness of the hierarchy. It may be the case that not all possible subtopics are listed for a topic. We solved this by introducing an “Other” category at each level to make up for the remaining topics. For example, consider the Physical Security indicator. We extend each of its levels with an “Other” topic as shown below:



Original hierarchy:

Physical Security	(L1)
Carceral Involvement	(L2)
Criminal Summons	(L3)
Imprisonment	(L3)
Juvenile	(L3)
Community-Led Policing	(L2)
Mobile trauma unit presence	(L3)
Neighborhood watch presence	(L3)
Resident satisfaction	(L3)
Police Misconduct + Force	(L2)
NYPD misconduct	(L3)
Police-involved deaths	(L3)
Use of force	(L3)
Violent Crime	(L2)
Domestic Violence	(L3)
Violent Crime Incidents	(L3)

Extend the hierarchy:

Physical Security	(L1)
Carceral Involvement	(L2)
Criminal Summons	(L3)
Imprisonment	(L3)
Juvenile	(L3)
Other	(L3)
Community-Led Policing	(L2)
Mobile trauma unit presence	(L3)
Neighborhood watch presence	(L3)
Resident satisfaction	(L3)
Other	(L3)
Police Misconduct + Force	(L2)
NYPD misconduct	(L3)
Police-involved deaths	(L3)
Use of force	(L3)
Other	(L3)
Violent Crime	(L2)
Domestic Violence	(L3)
Violent Crime Incidents	(L3)
Other	(L3)
Other	(L2)

Once the initial tweets were assigned to each topic at level three, we calculated the average of the assigned tweet representations and set that to be the topic’s representation. This also caused a propagation of updates to the parent topics. Using these new representations, we rediscovered the set of tweets that were closest to each topic. This process (clustering) was repeated either until there were no more changes or a stopping criterion was reached. This produced the indicators (if any) that were most representative of each tweet within the corpus.

In order to verify the effectiveness of our algorithm and compare it with existing methods, we performed two sets of experiments - (1) evaluated HierSeed on three benchmark datasets and compared performance with other methods, (2) evaluated HierSeed on a manually annotated test set of NYCHA tweets using the MOCJ indicator hierarchy.

For the benchmark evaluation, we used three publicly available datasets: RCV1-V2[8], NYTimes (NYT)[9], and Web-of-Science (WOS)[10]. RCV1-V2 and NYT are news categorization corpora while WOS includes categorization of published scientific paper abstracts. RCV1 has 103 topics, NYT has 166, and WOS has 141. For each, we created a seed document set for training by randomly sampling 4 documents per topic. We also compared our method with eight other unsupervised and supervised techniques. The metrics used were B-cubed F1 score[11], and V-Measure[12]. The values are reported in Table 1.

Method		**B-cubed F1**			**V-Measure**		
Type	**Model**	**WOS**	**NYT**	**RCV1**	**WOS**	**NYT**	**RCV1**
Weakly Supervised	**HierSeed**	**0.7131**	**0.6173**	**0.6546**	**0.7661**	**0.534**	**0.4815**
	JoSH	0.594	0.4692	0.5366	0.5927	0.4447	0.3591
Supervised	WeSHClass	0.642	0.5008	0.6034	0.5984	0.4461	0.4289
	HDLTex	0.2349	0.484	0.4231	0.0793	0.2253	0.1614
	HiAGM	0.4044	0.4065	0.4567	0.3467	0.2264	0.2754
	HiLAP-RL	0.1858	0.3451	0.4279	0.1383	0.1098	0.1602
	HFT(M)	0.3055	0.4079	0.5041	0.2729	0.1562	0.3422
Unsupervised	HLDA	0.1972	0.3811	0.3873	0.1984	0.1781	0.2336
	TSNTM	0.2262	0.3349	0.3726	0.1916	0.2052	0.3026

Table 1: B-cubed F1 and V-Measure on the WOS, NYT, RCV1 datasets.

We see that HierSeed outperformed all baselines on the SHC task with the best score for the WOS corpus, which we hypothesized was due to its simpler taxonomy compared to NYT and RCV1. Thus, we see the advantages of weakly-supervised approaches, and especially HierSeed, which can both adhere to a predefined structure (i.e., a predefined taxonomy), and make good use of the much easier to obtain unlabeled corpus, in an unsupervised manner.

[8] David D Lewis, Yiming Yang, Tony Russell-Rose, and Fan Li. 2004. Rcv1: A new benchmark collection for text categorization research. Journal of machine 654 learning research, 5(Apr):361–397.

[9] Evan Sandhaus. 2008. The new york times annotated corpus. Linguistic Data Consortium, Philadelphia, 6(12):e26752.

[10] Kamran Kowsari, Donald E Brown, Mojtaba Heidarysafa, Kiana Jafari Meimandi, Matthew S Gerber, and Laura E Barnes. 2017. Hdltex: Hierarchical deep learning for text classification. In 2017 16th IEEE international conference on machine learning and applications (ICMLA), pages 364–371. IEEE.

[11] Amit Bagga and Breck Baldwin. 1998. Algorithms for scoring coreference chains. In The first international conference on language resources and evaluation workshop on linguistics coreference, volume 1, pages 563–566. Citeseer

[12] Andrew Rosenberg and Julia Hirschberg. 2007. V measure: A conditional entropy-based external cluster evaluation measure. In Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL), pages 410–420.

Next, we evaluated HierSeed on the NYCHA tweets corpus using the MOCJ indicators. To do so, we first manually labeled around 1000 randomly sampled tweets from the corpus using the indicators. As a baseline, we used the clustering algorithm K-NN to compare our model with. In addition to the metrics in Table 1, we also report the Homogeneity, Completeness [13], and the B-cubed Precision, Recall[14]. The level-wise results are shown below.

Level	**Method**	**Homogeneity**	**Completeness**	**V-Measure**	**Precision**	**Recall**	**B-cubed F1-Score**
Level 1	K-NN	0.0611	0.0409	0.049	0.8302	0.4103	0.5492
	HierSeed	**0.0804**	**0.0491**	**0.0618**	**0.8311**	**0.5266**	**0.6447**
Level 2	K-NN	0.1962	0.2215	0.2081	0.3779	**0.4172**	0.3966
	HierSeed	**0.2855**	**0.3134**	**0.2988**	**0.8252**	0.3919	**0.5313**
Level 3	K-NN	0.2059	0.2123	0.2091	0.4869	**0.4378**	0.461
	HierSeed	**0.3792**	**0.2518**	**0.3026**	**0.8282**	0.4139	**0.552**

Table 2: Comparison of K-NN versus our technique, using 6 metrics at each level.

We also performed various ablation experiments where we evaluated the usefulness of using the LES method to update topic representations by comparing HierSeed's performance without using LES. We also evaluated the performance of our method with and without extending the hierarchy with the "Other" category. We observed that in both cases performing the operation gives better results, which showed the effectiveness of HierSeed in completing the SHC task.

Trends and Patterns

Analysis of the social media posts from twitter, coupled with geographical tagging and identification of socio-economic wellness-indicators led us to finding some interesting patterns and discussion dynamics regarding the quality of life, wellbeing, community, and living conditions among residents of New York City. We hypothesized and verified that amount of discussion is generally a good proxy for public concern about that indicator – as more discussion about a public issue (especially on twitter) tends to skew towards complaints or raising issues. Most trends observed in the data were what we expected to see, corroborated by contemporaneous news articles and incidents; others revealed more unexpected patterns.

We produced frequency histograms of the raw tweet counts for each MOCJ indicator, per month and for each target site by aggregating the number of tweets by users associated with that target site, posted in that month, and about that indicator. To account for a natural growth of Twitter

[13] Andrew Rosenberg and Julia Hirschberg. 2007. V measure: A conditional entropy-based external cluster evaluation measure. In Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL), pages 410–420.

[14] Amit Bagga and Breck Baldwin. 1998. Algorithms for scoring coreference chains. In The first international conference on language resources and evaluation workshop on linguistics coreference, volume 1, pages 563–566. Citeseer

usage and the expected increase in mentions for all indicators over time (due to individuals tweeting more), we normalized these counts by dividing each by the total number of tweets by users associated with that target area within each month, across all indicators. This made the month-to-month values comparable and more suitable for pattern and trend detection. Finally, we manually inspected each plot and flagged the readily observable patterns (consisting of peaks and valleys - over time). Some of the observations are as follows:

- Economic Readiness - Disconnected Youth:
  - Chronic Absenteeism

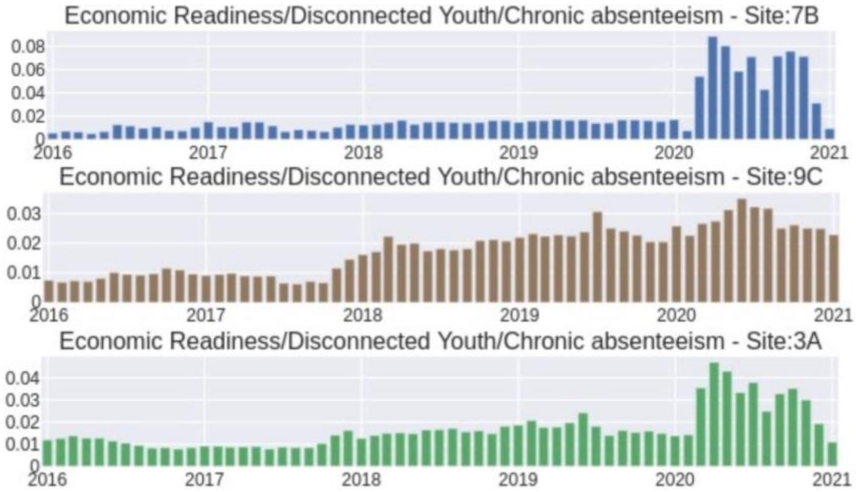


Fig. 5: Frequency of NYCHA tweets over time for the Economic Readiness - Disconnected Youth - Chronic Absenteeism indicator, shown for three target sites.

    - A clear spike is observed in this indicator around May-July 2020 (in Fig. 5). This may be due to the new regulations, school closures, and commencement of online classes during the onset of the COVID-19 pandemic around the same time.
    - Some of the most relevant tweets identified around this spike include, “Many teachers in nyc schools tested positive and this continues to be a huge issue, and now schools are not safe either”, and “if these were children living anywhere else, the public outcry would be all - consuming. Given the gross negligence & lack of response to the pandemic, why isn’t <school> on the worst list?”.
    - This trend is consistent across most target sites during the same time period (three sites shown in Fig. 5).
  - Dropout Rates

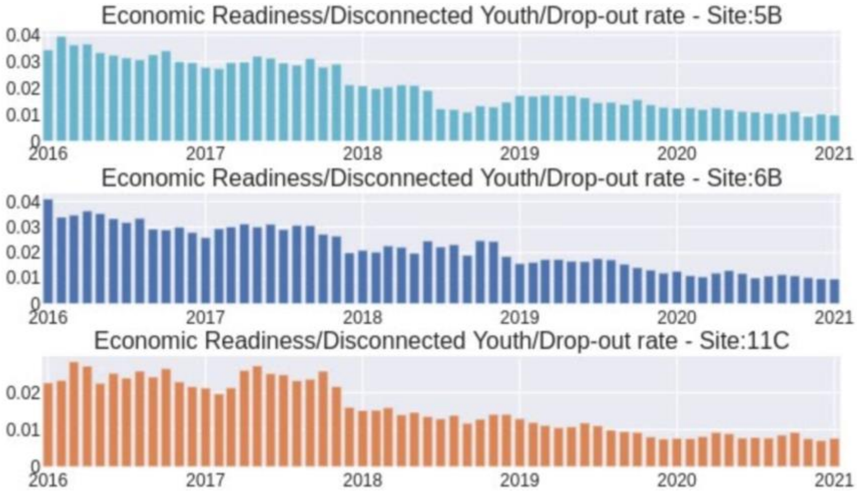


Fig. 6: Frequency of NYCHA tweets over time for the Economic Readiness - Disconnected Youth - Dropout Rate indicator, shown for three target sites.

- There is a general drop in the tweets concerned about dropout from high school or post-secondary degree attainments over the years 2016-2021. This trend is also consistent across almost all the sites (see Fig. 6).

- Economic Readiness - Employment Preparation
  - Career Centers

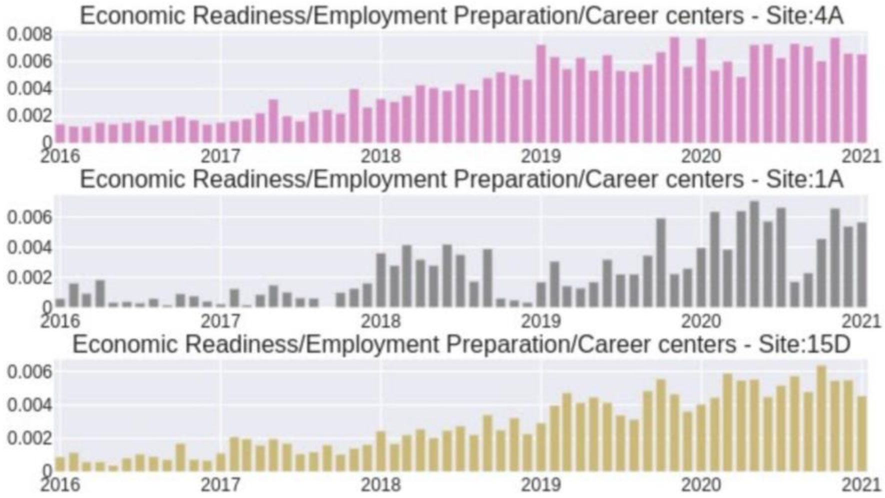


Fig. 7: Tweet frequency about Career Centers, shown for three target sites.

- Over the years, from 2016-2021, the number of mentions of career centers in tweets steadily increased (Fig. 7). This may signal the development of more professional training centers that have been created over the years.
- To investigate the seasonality observed in the plot, we selected tweets prom each of the local peaks:

– "My son seems busy – glad to see they are coming up with multiple career and development centers in our community to prepare them for the job market" - Feb 2017; and

– "<user> is happy to announce: <user> of 'The 10Ks of #PersonalBranding' presents @ our fabulous new-year event, Jan. 10, 2019. Join us 8a-12p for networking breakfast, career networking w/sponsors #leadership #diversity #Latinx" – Dec 2018; "Attended a monthly virtual chapter mtg w/ <user>, joined the #career and #selfimprovement Community of Practice. I mentioned everyone should connect with ea. other on LinkedIn" – Mar 2020.

- Economic Security - Health Security
    - ○ Healthcare Costs

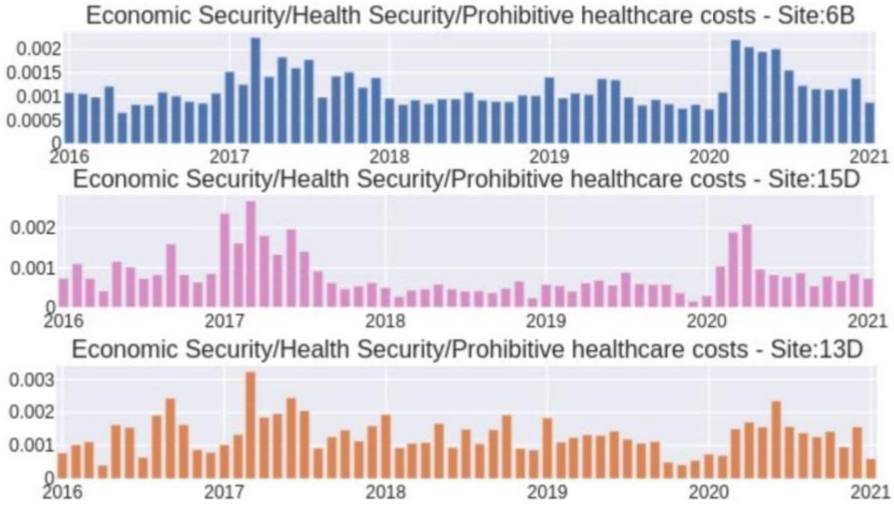


Fig. 8: Tweets about rising or prohibitive healthcare costs over time, shown for three sites.

- ■ A spike around mid-2020 is observed (see Fig. 8). This is also seen across all the target areas with clear boundaries. The observation may be the result of increased hospitalizations, facilities shortages, and high medical care requirements during the pandemic.
- ■ Some tweets from these local spikes included, "Any bill bipartisan or not which doesn't include $1200 for Americans who had to suffer during covid at the result of an incompetent gov't is dead on arrival!" – May 2020; and "Enough is enough <user> residents have been living in inhumane conditions pre - covid19. now with this crisis we need our rent canceled. living conditions have become grave. <user> always had his mind set on privatization instead of the livelihood!" – Jun 2020.

        - We also observed a clear spike in 2017, consistently across all sites. On further investigation, this seems to be consistent with various reports which state that in 2017, per-person spending reached a new all-time high[15].
        - Observed tweets for this included,“Medicare doesn't even pay for all the costs as it's written today so putting everyone on medicare won't absolutely protect ppl from going broke because of healthcare” – Mar 2017; “Obamacare rate on mental care - if you're not on medicaid you can't afford it” – Dec 2016; and “The problem with <user> health plan is that the ppl would get taken advantage of managing their own healthcare money” – Jan 2017.
- Physical Security - Carceral Involvement
    - Criminal Summons

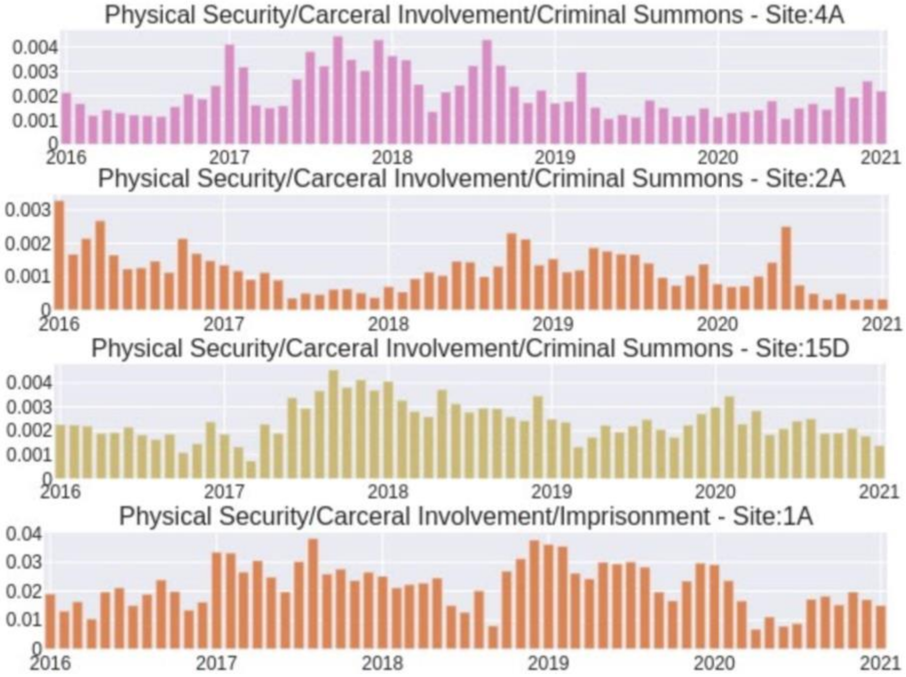


Fig. 9: Frequency of tweets related to Physical Security - Carceral Involvement - Criminal Summons and Imprisonment, shown across various target sites.

    - Imprisonment
        - For both of the above indicators, the trend is relatively stable over the years, except for a spike in tweets about criminal summons in mid-2017-18 (Fig. 9). Further analysis of the tweets showed that most posts were positive discussions on summons, imprisonment, and better feelings of security among the residents. Various crime reports also show that criminal activity in New York State plummeted in 2018[16].

[15] HCCI’s 2017 Health Care Cost and Utilization Report: https://healthcostinstitute.org/annual-reports/2017-health-care-cost-and-utilization-report

[16] New York State Crime Report, 2018 by Division of Criminal Justice Services: https://www.criminaljustice.ny.gov/crimnet/ojsa/Crime-in-NYS-2018.pdf

        - Some tweets assigned to this indicator, consisting of positive sentiments and discussions include,“<user> thank you for your passion and tireless work to improve nycha communities, and for inspiring us all !!”; “FINALLY! the new york state assembly standing committee on housing will hold a virtual public hearing on nycha’s ‘blueprint for change’ proposal. right now <user> is testifying. we will have a full report of what happened”; and“<url> exclusive one on one interview with lawyer who just won a bronx mom and daughter a million dollar lawsuit against nycha.”
        - We also observed a new consistent increasing trend in the year 2020, which plateaus. Tweets such as, “Drive up shootings? Keep in mind the state courts are closed, covid-19 is ravaging our jails and they had to slow down piling people up in the jails and release repeat offenders, because of covid” suggest that this may again be a result of the general hardships faced during the pandemic in 2020, leading to increased criminal activity.

- Physical Security - Police Misconduct + Force

    - NYPD Misconduct

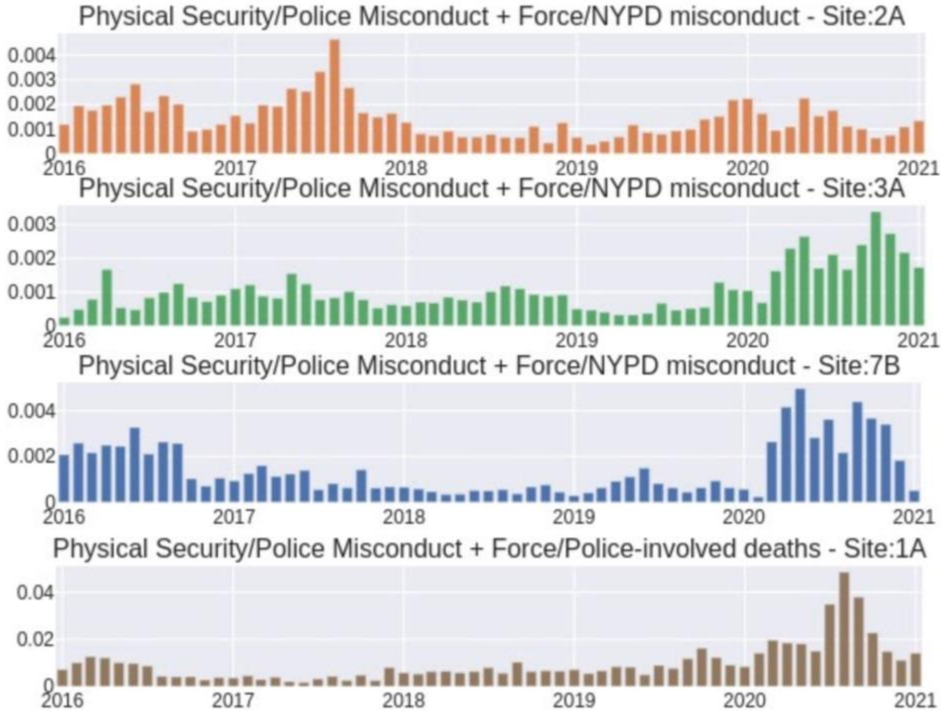


Fig. 10: Frequency of NYCHA tweets over time for the Physical Security - Police Misconduct + Force - NYPD misconduct and Involved Death indicators, shown for various target sites.

        - Some posts highlighting police misconduct, as tweeted by NYC residents included – “Is the mayor serious!! You think you have worked with #nypd to stop the brutality of black and brown people? Either you are out of your mind or your a bold face liar.”; “<user>, 20, was viciously assaulted by an NYPD officer while at a protest. That officer is now facing charges and will be held accountable thanks to their attorney”;

> "Im sure cop <user> got what he deserved. I stand with the civilians that are just trying to protest cops played us too many times and have done us dirty."

- Police involved Deaths
    - There is a consistent observed spike in the frequency of tweets concerning police misconduct and police involved deaths around May to July 2020.
    - The observation is consistent across most of the target sites.
    - This may be due to the citywide protests following the killing of George Floyd[17].
    - Relevant tweets identified include,
        - "Shocked to hear the #NYPD chief say - It's a great thing that large crowds are assembling (to protest) for this cause. #RIPGeorgeFloyd";
        - "People have explained over and over again that no one means take all the funds smh. Got the nerve to have George Floyd and BLM in the bios yet can't understand why defunding is necessary";
        - "My Bronx community is out here marching for black lives, defunding the police, police free schools, and removing the foot of white supremacy in our borough"; and
        - "If I were an alien, casually keeping track of US politics from afar, I would see last night compared to #BLM and assume that the government finally got around to defunding the police!"

NLP Conclusion

The proposed NLP techniques to automatically discover social media discussions from residents of various NYC neighborhoods and identify their concerns, coupled with this preliminary investigation and analysis, demonstrated the usefulness of the tools. They allow a way to passively monitor public concerns about the MOCJ indicators, and reveal both obvious and inconspicuous trends. Our proposed models - Geotagger and HierSeed - can allow policy makers to develop trend plots, detect spikes and dive deeper, in real time, to see the constituent tweets or posts and better understand the voices of those being directly impacted by the outcomes of such policies.

SAFElab Conclusions

This report provides a unique opportunity to gain community recommendations and broad insight into the perceptions of New York City residents through qualitative interviews, focus groups, and NLP and computer vision analysis tools. The study's design included insights and input from MOCJ, social work researchers, computer scientists, and local New York City residents. Through this process, researchers were able to understand attitudes of low-income New York City community members regarding community well-being, social media usage, as well as the health

[17] Murder of George Floyd: https://en.wikipedia.org/wiki/Murder_of_George_Floyd

and living disparities during the co-occurring pandemics of COVID-19 and anti-Black racism.. We identified key factors of residents' lived experiences and recommendations for community- and government-level interventions for these intersecting social crises by working alongside the aforementioned communities. We expanded on the current scholarship and supported ongoing conversations about the importance of community engagement and collaboration when addressing social issues marginalized communities face. Future studies and reports should look into and potentially replicate similar analyses of different social issues, emphasizing community collaboration and social media data.