
Artificial Intelligence and Inclusion: Formerly Gang-Involved Youth as Domain Experts for Analyzing Unstructured Twitter Data

Social Science Computer Review
1-15

© The Author(s) 2018

Article reuse guidelines:

sagepub.com/journals-permissions

DOI: 10.1177/0894439318788314

journals.sagepub.com/home/ssc



William R. Frey¹, Desmond U. Patton¹, Michael B. Gaskell¹,
and Kyle A. McGregor²

Abstract

Mining social media data for studying the human condition has created new and unique challenges. When analyzing social media data from marginalized communities, algorithms lack the ability to accurately interpret off-line context, which may lead to dangerous assumptions about and implications for marginalized communities. To combat this challenge, we hired formerly gang-involved young people as domain experts for contextualizing social media data in order to create inclusive, community-informed algorithms. Utilizing data from the Gang Intervention and Computer Science Project—a comprehensive analysis of Twitter data from gang-involved youth in Chicago—we describe the process of involving formerly gang-involved young people in developing a new part-of-speech tagger and content classifier for a prototype natural language processing system that detects aggression and loss in Twitter data. We argue that involving young people as domain experts leads to more robust understandings of context, including localized language, culture, and events. These insights could change how data scientists approach the development of corpora and algorithms that affect people in marginalized communities and who to involve in that process. We offer a contextually driven interdisciplinary approach between social work and data science that integrates domain insights into the training of qualitative annotators and the production of algorithms for positive social impact.

Keywords

social media, gang violence, domain experts, artificial intelligence, inclusion, qualitative methods, natural language processing, Big Data, ethics, law enforcement

¹ Columbia University, New York City, NY, USA

² New York University's Langone Health, New York City, NY, USA

Corresponding Author:

William R. Frey, Columbia University, New York, NY 10027, USA.

Email: wf2220@columbia.edu

In the social sciences, Big Data represents an unprecedented opportunity to understand and support segments of the human population that were at one time too difficult to reach through traditional methods (Christians & Chen, 2004). Individuals who are marginalized, stigmatized, or living in areas with high rates of violence may not be accessible for study, as they are too geographically dispersed, experiencing personal anxiety and embarrassment, or making face-to-face contact would pose a significant risk to a participant or researcher's safety. Given the evergrowing popularity of the Internet, smartphones, and social media, these individuals generate data that are able to be acquired and analyzed. As such, various groups including academics, clinicians, nongovernmental organizations, and governmental bodies have seized the opportunity to make use of various forms of data—including data related to natural disaster relief and disease epidemiology, among others—in the service of positive social impact (Decuyper, 2016). In this article, we describe our current research involving unstructured social media data generated by a marginalized, often misunderstood societal group: gang-involved and affiliated young people. We then explain the process for integrating the knowledge of formerly gang-involved individuals to act as domain experts in the area of gang and crew culture.

The use of Big Data requires more than simply collecting large amounts of information. Data must be acquired, stored, extracted, and annotated before analyses can begin to provide meaningful findings to create solutions to social problems (Gandomi & Haider, 2015; Labrinidis & Jagadish, 2012). Further complicating this process is that as much as 80% of data generated by humans are considered “unstructured” (Rizkallah, 2017). Unstructured data, as opposed to structured data, are not easily compatible with existing databases (Che, Safran, & Peng, 2013). Examples of unstructured data include videos, blog posts, and tweets. Khan et al. (2014) describe unstructured data as “human information,” and it is therefore not surprising that these content-based, unstructured data are of high interest. Unstructured data provide a nuanced and rich reflection of the human experience.

Unfortunately, the same richness and nuance that make unstructured data appealing make it difficult to study empirically. Prior to analyzing data for various applications, the data must undergo filtering and classification (Khan et al., 2014). In order to improve efficiency, computational methods are often developed to quickly categorize or otherwise organize unstructured data (Suthaharan, 2014). Although there are many systems for accomplishing this goal, most require some degree of initial training or feedback based on human-generated knowledge. For example, a system that is designed to detect guns in images must first “learn” on a training set of gun images (Seitz, 2016). To have success in this endeavor, this training set must have been created by a human who was able to confirm that the images were of actual guns rather than replicas and other objects that look similar to guns. Therefore, artificial intelligence computational systems for annotating or classifying Big Data are only as accurate as the initial human-classified data set. This initial classification is critical for all future applications of the system, especially given that performance of the system is often measured against a similarly labeled test set.

In the above example of identifying the presence of a gun in an image, categorization may be a relatively straightforward task; however, as is often the case in social science research, the variables of interest pertaining to the unstructured data represent complex societal or human behavior issues. In this situation, the human task of examining the online behavior of another human being and making some sort of judgment regarding a complex behavior is inherently fraught with potential for bias and misunderstanding. Social scientists are trained and prepared to acknowledge bias and to create safeguards to minimize potential for bias in their work. In research settings, it is then considered best practice to articulate this process for minimizing bias in a research report.

Alternatively, commercial or government applications of this type of work have a vested interest in keeping the specific mechanisms for these processes confidential. Commercially, these systems may represent significant investments in time, money, and effort, and it is therefore

reasonable to assume that these would be kept proprietary. In certain government activities like law enforcement, the mechanisms of data analysis are understandably maintained in secrecy, as many policing efforts depend on maintaining covert operations and processes. However, without some degree of understanding about the systems that groups use and how these were developed, we have no way of knowing whether the degree to which bias or misunderstanding has been incorporated in their development.

One application of unstructured data analysis that is both frequently used and rife with potential bias and misunderstanding is the use of social media surveillance in marginalized communities. Social media surveillance has become common use for police departments, as the International Association of Police Chiefs (2015) suggests that 94.6% of agencies use social media in some capacity. Access to incalculable amounts of social media data and the development of algorithms to sift through posts gives police agencies the ability to scale up and digitize proactive policing strategies. In off-line policing strategies, there are well-founded practices of racial profiling and preemptive criminalization of people of color living in marginalized communities (Barlow & Barlow, 2002; Brunson & Miller, 2006; Carter, 2014; Goffman, 2015; Meehan & Ponder, 2002). When translated into digital spaces, traditional policing strategies may lead to similar consequences for people and communities of color (Hackman, 2015). Some even argue that social media surveillance by policing agencies is the new virtual stop and frisk (Patton et al., 2017).

Policing agencies like the New York Police Department (NYPD) engage in online surveillance of individuals as young as 10 years old, with seemingly no official requirement to notify parents or guardians. It is unclear what restrictions are in place regarding the civil liberties of youth being monitored online (Hackman, 2015). Who are involved in conversations around what constitutes criminal behavior online? Who are being monitored digitally and what factors are considered? How are complex social media posts interpreted and acted upon? These questions are especially relevant when considering communities where gangs and crews are common and corresponding cultural behaviors permeate into the lives of people throughout the community—into the expressions of their lived experiences and language used on social media, blurring the identifying features of gang affiliation and involvement.

In communities where gang involvement and affiliation are prevalent, young people's behavior tends to be influenced by the "code of the street," defined as unwritten rules that dictate norms, territories, and behavior with life and death consequences (Anderson, 1999). Furthermore, it has been found that these codes extend into online spaces like social media, creating a "digital street" (Lane, 2016). Digital language and behaviors adopted by young people through music and interactions with their peers often do not clearly demarcate who are and are not involved and affiliated with gangs through social media. Without a clear understanding of the identifying features of specific gangs and crews, it can be challenging to categorize whether a social media user is gang-involved or affiliated. Even a user's voluntary self-identification as gang-affiliated through social media leaves many unanswered questions around whether or not they are merely posturing.

Gaps left in social media data make interpretation difficult with meaning determination often depending on understandings of community-specific off-line context. Youth across various communities may have vastly different ways of expressing themselves online due to their specific neighborhood context, which may shape why and how they communicate with each other. Insights from one neighborhood may not be generalizable across other groups even when they share similar demographic and social characteristics. If there is an absence of community input when interpreting social media data, there is a dangerous possibility that nuances in language, culture, and context will be misinterpreted. As seen in digital surveillance practices by law enforcement, biases leading to misinterpretations of social media data can wrongfully criminalize youth of color who are also experiencing off-line racial profiling. The question then arises: How do researchers, commercial groups, police, or other entities utilize these data in ways that minimize the potential for bias and

misinterpretation? We argue that giving voice to youth of color with lived experience in gang and crew culture decreases bias in our computational systems, improves the accuracy of unstructured data categorization, and greatly enhances our ability to accomplish our overarching goal of preventing violence.

Twitter, Gang Violence, and Natural Language Processing

The Gang Intervention and Computer Science Project is a research study between faculty and students at the School of Social Work and the Data Science Institute at Columbia University. The project uses publicly available Twitter data from gang-involved and affiliated youth in Chicago to better understand the factors and conditions that shape aggressive and threatening communication online known as Internet Banging (Patton, Eschmann, & Butler, 2013), determine the pathways from online communication to off-line violence, and develop computational tools for preventing violence instigated through social media communication. We use a mixed-method process which includes a deep read textual analysis of Twitter communication informing a set of machine learning algorithms that detect and predict aggression and loss in Twitter data (Blevins et al., 2016; Patton, Eschmann, Elsaesser, & Bocanegra, 2016).

Part-of-Speech Tagger

In our prior research, traditional part-of-speech tagging applications (e.g., Manning et al., 2014) did not work for analyzing Twitter posts by gang-involved youth from Chicago (Blevins et al., 2016; Patton, McKeown, Rambow, & Macbeth, 2016). Tweets in our corpus contain unique, localized language with nonconventional spelling and grammar that is different from standard English. To address these challenges, we developed a new part-of-speech tagger informed by interpretations from our domain experts. Our tagger performed with an accuracy of 89.8% on the heldout development set, considerably higher than accuracy achieved by the Carnegie Mellon University Twitter Part-Of-Speech Tagger (Owuputi et al., 2013), which had an accuracy of 81.5%.

Classifier

Domain expert interpretations of Twitter posts informed our thematic analysis, which produced qualitative themes used as features in the development of a classifier. Initially, 26 unique themes were generated from domain expert interpretations and thematic analysis. These themes were collapsed down to three for classification: *aggression*, *loss*, and *other* due to their role in potentially leading to off-line violence. “Aggression” was defined as insults, threats, mentioning of physical violence, and wanting retribution. “Loss” was defined as expressions of grief, trauma, and sadness, as well as mentioning someone’s death or incarceration. “Other” contained a variety of themes, from status-seeking and mood, to health and authenticity. Posts identified as *other* did not contain features involved in *aggression* and *loss*. The classifier used natural language processing to automatically analyze the text and emojis to classify a Twitter post.

There are some community organizations using social media in their intervention and outreach work, such as the E-Responder program (Citizens Crime Commission, 2017). However, the number of social media posts produced daily by any given community makes the use of social media data in urgent violence interruption and intervention unmanageable. Our plan is to develop and refine automatic classification of social media posts that range from trauma and grief responses to more direct threats, used for community intervention from professionals, such as social workers and violence outreach workers.

Data

We created an unstructured social media corpus in February 2017. This data set contains the last 200 tweets of 279 users from neighborhoods in Chicago with high rates of violence who suggest on Twitter that they have a connection, affiliation, or engagement with a local Chicago gang or crew. These users were chosen based on their communication with a seed user, Gakirah Barnes, and the top 14 communicators in her Twitter network.¹ Gakirah was a self-identified gang member in Chicago before her death in April 2014. An additional 214 users were collected using snowball sampling techniques (Atkinson & Flint, 2001). Traditionally, snowball sampling has been used to recruit hard-to-reach research subjects, where one subject provides the name of another subject, and so on. We adapted this approach for social media data by looking for clues and references from one social media user to find another social media user in the network who may be displaying similar behaviors or gang affiliations as the first user.

Method

We developed a multistep process that integrates domain insights from formerly gang-involved youth to inform the master of social work (MSW) student annotators process, which provides the training data for our natural language processing analysis. Our process includes (1) identifying, onboarding, and integrating domain experts; (2) initial domain expert interpretations of Twitter data; (3) training and assessing social work student annotator quality; and (4) iterative domain expert involvement and reconciliation of student annotator disagreement (Figure 1).

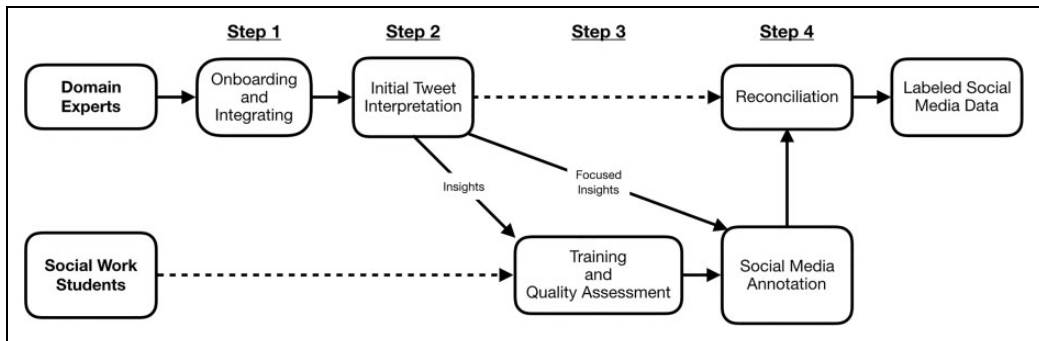


Figure 1. Overview of qualitative annotation, labeling, and reconciliation process.

Step 1: Identifying, Onboarding, and Integrating Domain Experts

We selected two young men (Black and Latino) aged 18 years and older who live in Chicago neighborhoods with high rates of community violence to work as domain experts. We partnered with a community-based organization in Chicago that offers violence prevention programming across the city. The executive director, a social worker, and seasoned violence prevention expert identified our domain experts based on their prior gang involvement, exposure to violence, and willingness to participate in the research process. Our domain experts were both in the process of rebuilding their lives after leaving gangs, and one had recently relocated to another community during this process. They had recently completed a violence prevention program aimed at minimizing the risk for incarceration and gang involvement through a curriculum consisting of support around belonging, positive identity development, and community engagement. Our domain experts had current experience using social media to connect with their peers.

Once selected, domain experts were formally hired by Columbia University's School of Social Work, provided tablet computers and keyboards, and trained on the process of interpreting tweets in

Chicago by an MSW-trained social worker who is a researcher with 8 years of youth development experience. Eliciting insights from formerly gang-involved young people is not as easy as it may seem. These young people did not see their knowledge as domain expertise, only as subconscious choices made to navigate their physical, social, and digital environments and nothing more. Without first unlocking the understanding that their interpretations of social media posts contained knowledge unknown to us as researchers, our domain experts had difficulty deciphering what was and was not common knowledge, often leaving out vital domain expertise from their interpretations. Domain experts hand-annotated tweets with people unfamiliar of their domain for the purpose of uncovering their contextually specific insights. Once the young people were able to recognize that they had domain expertise, we provided them with a test set of tweets to practice the process of recording their interpretations before we had them interpret the official data set.

Step 2: Initial Domain Expert Tweet Interpretation

Domain experts initially hand-annotated 185 tweets remotely through a password-protected excel spreadsheet. They provided insights around meaning, emoji and hashtag usage, and relevant off-line context to determine intracommunity perceptions of social media posts. Domain experts interpreted tweets from other gang-involved and affiliated young people utilizing their lived experiences as a framework for making meaning of each post (Table 1). Most tweets in our corpus included language that is specific to a Chicago neighborhood or the city as a whole, such as words referencing geographic locations and institutions, off-line events and people, and emojis. A gang may have an emoji or set of emojis they use to represent their affiliation.

Table 1. Examples of Key Insights Domain Experts Add to the Analysis of Social Media Data.

Insight	Example
Language	Referring to marijuana as the name of someone a person has allegedly killed to bolster their reputation
Emojis	The use of the clapping hands emoji (👏) to reference shooting someone
Song lyrics	Knowledge of lyrics from nationally and locally known musical artists. Song lyrics may seem to have one meaning, if not registered as such
Behavioral/temporal cues	A user in an image “throwing up” or “throwing down” a hand gesture can alter meaning from support to disrespect
People	Recognizing a name being mentioned as someone who was recently killed
Neighborhood references	A user mentioned where a party was occurring by referencing a specific community landmark, unknown to people outside of the neighborhood
Gang/crew knowledge	Understanding the contentious relationship between two gangs and the actions were taken to cause the current strife

Although large amounts of available data provide unprecedented access to the online behavior of this marginalized group, the lack of highly contextualized and nuanced understandings of gang cultures drastically limits the potential for comprehension without domain expertise or extensive contextual training. To address this issue, we integrated the use of domain experts in this area to inform our research. For example, a domain expert interpreted the following tweet (Figure 2).

if yu get caught jus never snitch 🤨One Day
Yu Coming Home 🏠

Figure 2. Example of tweet interpreted by a domain expert.

This person is saying that if you were to get caught doing an unlawful crime, there is no point in telling on the next man because eventually you will be released from jail (and face the repercussions of having told). This person is speaking from prior knowledge which is why he uses the 100 emoji.

The domain expert suggested that the tweet describes being caught while committing a crime, even though it is not mentioned in the tweet. He also stated that the use of the 100 emoji in this context means that the user has prior experience with being incarcerated and not sharing information with law enforcement—*snitching* [👉]. Interpretations like this one were used to train social work student annotators in the SAFElab.

Interpretations from domain experts, at times, were in conflict with one another because each expert had their own set of unique experiences. Our utilization of domain expert insights is enhanced when provided a variety of perspectives for the same social media post. When annotating a social media post, we are not looking for singular meaning of intention, or right and wrong translations. Instead, we seek to understand the complexity of digital communication and behavior through various domain expert experiences and perspectives.

Step 3: Training and Assessing Social Work Annotator Quality

Our social work student annotators are trained to analyze social media posts using domain expert interpretations and immersion in the local context of Chicago neighborhoods from which the social media data derive. Student annotator interpretations were compared to ones provided by our domain experts to validate their training and understandings of the data. In addition, trainings included other local sources of Chicago domain knowledge, such as Chicago as a context, physical locations, gang and crew territories, news articles, YouTube videos, and immersion in Twitter accounts of gang-involved youth. Web-based resources were also utilized, including but not limited to Twitter Advanced Search and Hipwiki, in order to find contextually specific information. While formerly gang-involved and affiliated young people have various forms of expertise around localized language, culture, and events, social work annotators have their own expertise. They are able to thematically analyze social media data and consider its pertinence to social work, mental health, and violence prevention.

After the training, social work annotators were given a practice set of 100 Twitter posts and 20 social media accounts of gang-involved young people to immerse themselves in the local context of the social media data. Deep immersion in the data and practice annotating social media posts led our annotators to better understand the language, emojis, and hashtags used by gang-involved young people. Annotations were completed in a web-based visual and textual analysis system, which we created for the purpose of this study. Our annotators were trained using the Contextual Analysis of Social Media (CASM) approach to qualitatively analyze the textual and contextual features of a social media post in-depth. This approach requires our annotators to complete a baseline interpretation of the post without considering context, in order to uncover any assumptions they may have. Next, annotators analyze text, emojis, and hashtags, researching any words they may not know using web-based resources. Then, our annotators analyze the author of the post, their digital peer network, any off-line events being referenced, virality (how far the post is traveling within the digital network), and any engagement others have with the original post. Finally, they assess their baseline, comparing their interpretations before and after they examine the contextual features in the post. This assessment includes an explanation of the contextual features and evidence they have found to support their interpretations.

After social work students completed their practice annotations, a subject expert inspected their annotations to provide feedback and reconcile any mistakes using domain expert interpretations to determine accuracy. For example, one of our annotators may be misinterpreting a certain word or was

unable to recognize a Chicago-specific event mentioned. Once it was determined that our social work annotators grasped the domain-specific context and the contextual features relevant to our social media corpus, they were given the official Twitter corpus to annotate. Each qualitative annotator interprets 100 tweets every week. Each tweet is coded by two different annotators and reconciled by either a domain expert or a SAFElab subject expert who breaks any disagreements in the interpretations.

Step 4: Iterative Domain Expert Involvement and Reconciliation

Our social work annotators relied heavily on iterative insights from domain experts. The neighborhood climate constantly shifts, especially with regard to conflicts between crews and gangs. Without iterative insights from domain experts, it is challenging to glean the complexity of gang conflicts through social media data alone. Domain experts first provided their insights through broad interpretations of a widespread set of social media posts from gang-involved young people. As they began to reach qualitative saturation and a minimal amount of new information surfaced about language and community context, domain experts provided more intermittent, focused insights regarding deeper contextual information, nuanced language, and emoji usage as mistakes and questions arose from our social work annotators. Throughout our annotation process, domain experts continued coding a random sample of posts and reconciled any disagreements our social work annotators had in their interpretations. When assessing annotator disagreement, we asked domain experts to view both of our annotators interpretations and choose which interpretation was more accurate based on their lived experiences and expertise. Once reconciled, interpreted tweets were used to enhance our part-of-speech tagger and classifier.

Case Examples

Language and Emojis

After reviewing interpretations from domain experts, our annotators began to recognize patterns in language and emoji use in social media data that had not been interpreted by our domain experts. They were able to determine meaning through deep immersion in language, context, and building an awareness of community culture over time. For example, the following tweet was analyzed by our social work annotators but has not been interpreted by our domain experts (Figure 3).

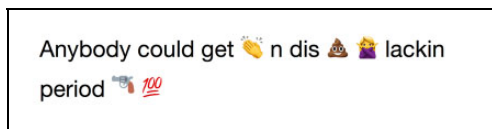


Figure 3. Example of tweet interpreted by a social work student annotator.

Our annotator noticed that some words have been omitted and replaced with emojis. Through previous domain expert interpretations, our annotator learned that one of the emojis is replacing a word (🤝), *clapped*. The word *clap* is a verb meaning *to shoot* (with a gun). When the other two emojis replacing words were analyzed (🗑️ and 📺), an initial interpretation of the tweet is:

Anybody could get clapped [🤝] in this shit [🗑️], no [📺] lacking period 🗡️💯

Next, some words and phrases needed to be translated to understand their contextual meaning within this specific tweet (*in this shit* and *no lacking period*). Domain experts often interpreted *in this shit* to mean *in the street life, in gang activity, and gang violence*. When interpreting *no lacking*


period, the term *lacking* must first be understood. Our domain experts shared with us that *lacking* refers to *being caught off-guard, unprepared, or unaware* by opposing gang members. When a gun is mentioned or a 🔫 emoji is present, *lacking* can be used to reference being caught unarmed, which is the case in this post. Based on our analysis, our qualitative annotator suggested this final interpretation of the Twitter post:

Anybody could be shot in these streets and in this gang life, do not get caught unprepared without a gun on you, period.

Domain expert insights provided annotators with an alternative meaning of words and emoji use in order to accurately interpret the meaning of this post. Without training our annotators on insights from domain experts, they may have trouble recognizing local community context and nuances within social media data.

Song Lyrics

Some social media posts in our data set contained seemingly aggressive and threatening language that may be nothing more than a young person posting lyrics to a song they liked or were listening to (Figure 4).




aint kill yo mans & ion kno ya homie

Figure 4. Example of tweet involving song lyrics.

At first glance, this post may seem to be communicating that this user stated they were not involved in the death of someone (*aint kill yo mans*) and doesn't know them (*ion kno ya homie*). It would be relatively logical to assume that this post is referencing off-line violence. However, our domain experts noted that these are lyrics from a song by a Chicago rapper, Lil Durk. Song lyrics posted on Twitter have been a consistent meaningful theme in our corpus when determining meaning. Without this vital insight by our domain experts, many of our classifications would be inaccurate, leading to tweets being incorrectly coded as *aggression* when they are actually song lyrics in the absence of any other contextual information to suggest conflict.

Referencing Off-Line Context

Another thematic insight offered by our domain experts involved the use of off-line context to reference and brag about previous events of violence. In the following post, domain experts pointed out a specific way gang-involved young people insult or *diss* other gangs and brag about their past acts of violence (Figure 5).



Jay Smokin thinkin bout DMoney 🤔👉👈

Figure 5. Example of tweet referencing off-line context.

The author of this post is sharing that they are “*Jay Smokin*,” with the word “*smokin*” to describe their current activity of marijuana use. Jay is the name of a rival gang member whom this user has

allegedly killed. In many tweets from this user's Twitter feed, they referenced weed by the name *Jay*. Our domain experts informed us that this is a way to *diss* (disrespect) rival gang members and bolster one's own reputation as a tough and violent person. Gang-involved youth are able to iteratively reference and remind others of their violent acts through disrespecting a rival gang. Additionally, this user is referencing a past event in their life, by expressing sadness around their friend "*DMoney*" being shot and killed by the Chicago Police Department (CPD). This user is simultaneously referencing a person in a rival gang that they have allegedly killed (*Jay Smokin*) and expressing grief around their friend being killed by CPD (*thinkin bout DMoney* 🥺🙏🔪). This post highlights various ways that off-line context is brought into social media by gang-involved young people. We relied heavily on insights from domain experts to understand meaning and learn the various ways off-line context is used in our Twitter corpus, which informed the way we trained our social work annotators.

Discussion

This article describes the importance of involving formerly gang-involved youth as domain experts for machine learning. Our multistep process leverages the insights of domain experts to train social work annotators and create an inclusive, community-informed algorithm aimed at predicting and detecting potential acts of gang violence. Insights from domain experts were used to enhance the interpretation of unstructured data sets and elucidate misinterpretations of social media posts through reconciliation and feedback loops with social work student annotators. These feedback loops led to more precise understandings of social media data and the development of comprehensive computational tools used to support positive social change in marginalized communities and violence prevention work.

Our experience of incorporating formerly gang-involved youth as domain experts has highlighted several important points for both social scientists and data scientists. First, as Big Data is increasingly utilized for positive social change, there are prime opportunities for mutually beneficial collaborations between social science and data science researchers. Given that unstructured data must be organized or classified by human intellect in some capacity, machine learning processes require more than just large amounts of data and immense computing power. As applications of machine learning move toward achieving societal or human behavior change, the data in need of classification will increasingly involve information reflecting the human experience and will deal with complex societal issues (e.g., preventing gang violence). Interdisciplinary approaches become necessary to understand the context of social media data and recognize social processes involved in the interpretation of that data. The ability to interpret social media data may be limited through traditional methods of observation and pattern analysis (Quan-Haase & McCay-Peet, 2017). Through bypassing walls that have traditionally separated various academic disciplines, both social science and data science researchers stand to benefit.

Incorporating machine learning and computational techniques greatly enhances the work of social scientists. Through these partnerships, social scientists can better understand groups, individuals, or even individual behaviors that would be difficult to study through traditional methods. For example, before the Internet, communications of escalating tensions between rival gangs primarily occurred in face-to-face exchanges. To have examined these interactions, a researcher would have had to be physically present for these communications or gather recollections of someone involved after the fact. Today, these communications unfold through public postings of gang-involved youth on Twitter (Patton et al., 2013). The advantages go beyond the study of certain people or behaviors and extend to the ability to deliver—or provide a pathway to—individual social services. An example of this work was described in Facebook's recent announcement of using natural language processing to identify suicidal individuals and provide an automated response that includes

recommendations for social services (e.g., suicide hotlines; Constine, 2017). Our system of accurately classifying a tweet through domain expert-informed computational methods will ultimately allow us to reach and provide resources to many more people than any group of individual clinicians manually scouring the Internet, or the punitive surveillance methods of law enforcement agencies.

Data scientists also stand to benefit greatly from qualitative contextual analysis approaches (e.g., CASM). As the amount of data collected and available computing power grow exponentially (Talluri, 2016), knowledge of how to categorize human information may be a limiting factor in the efficiency or effectiveness of machine learning. In future studies using Big Data for the public good, it may be the research team with the most nuanced understanding of societal and behavioral processes—not the team with the fastest processors—that is most positively impactful. Data scientists face practical challenges of identifying, building relationships with, and working alongside people from marginalized groups. Social scientists who have community building and outreach experience can aid in the inclusion of domain experts from marginalized groups in data science research. Additionally, social scientists are ideal candidates for assisting in the process of understanding and classifying data accurately by way of their training and subject matter expertise. Using this training, along with the process described in this article, social scientists are well positioned to assist in efforts to recognize and understand the impacts of biases that may enter into machine learning systems.

Partnerships and collaborations across disciplines take on even more importance as there is an increasing expectation that artificial intelligence (AI) systems be able to explain how they have come to specific decisions (Doshi-Velez et al., 2017). The development of future AI systems may place significant emphasis on incorporating a discrete process for discerning why a system took a specific action. Calls for algorithmic transparency must stretch beyond development and utilization to the impacts of biases and inaccuracies on the intended purpose of these algorithms. Processes like the one we have provided hope to broach this challenge by highlighting the complexity of the problem and ways we can begin to strengthen the connection between communities, social science, and data science to build impactful solutions together.

Our discussion of the methods described in this article would not be complete without acknowledging the larger context of predicting gang violence through social media and the related ethical concerns, especially regarding who can access our data and methodological approaches. We have taken the position that sharing our data and methods with community-based violence prevention organizations, as opposed to law enforcement agencies, maximizes the potential benefits and minimizes risks to young people and communities. This decision rests largely on empirical evidence that certain on- and off-line police practices (e.g., social media surveillance, selective policing, and traditional “stop and frisk”) result in both disproportionately high rates of contact with law enforcement and higher rates of incarceration for people of color (American Civil Liberties Union, 2018). Given that our data and methods could be used in a similar way by law enforcement, we choose to instead partner with community-based violence prevention organizations who work toward resolving conflict in ways that do not involve criminalization and incarceration. Additionally, community organizations are only able to carry out their work through trusting relationships with the community. This includes the understanding that community interactions with the organization will not result in police intervention or arrest. Since our broader work relies on partnerships with community organizations, such as bolstering violence interruption practices and identifying domain experts, sharing information with law enforcement may also prevent us from partnering with these groups in the future and undermines the trust these groups have built over many years with the communities they serve.

Future Directions

Our process relies on collaboration with community organizations working to prevent violence. For this purpose, we need to evaluate our process with other groups of people in communities with high

rates of violence. For example, violence prevention workers have been found to have valuable insights regarding activity of young people on social media, including knowledge around taunting, posturing, and boasting about violent events on social media (Patton, Eschmann, et al., 2016). Capturing insights from violence prevention workers may enhance our ability to prevent violence before it happens.

Additional research is needed to explore the various ways of collecting domain expert insights. What are the differences between interpretations of social media data by focus groups of domain experts as opposed to individuals? How would focus group interpretation change the insights domain experts contributed? Would we be able to triangulate meaning from a collective of domain expert perspectives? Furthermore, changing the ways that domain experts are involved may influence the direction and scope of our work, embodying a more collaborative and participatory research approach. In our current work, insights from domain experts will inform digital intervention and prevention strategies used in their own neighborhoods. Additional work is needed to identify strategies to support the integration of domain experts in all aspects of our work. Finally, we need to evaluate whether our process involving domain experts can be used in research addressing other pressing social issues beyond violence prevention.

Limitations

Our current work has a few limitations that are relevant for future research. First, our corpus contains posts from only one social media platform, Twitter. Due to a variety of factors like privacy settings on Facebook and the short-lasting nature of posts made on Snapchat, it is challenging to observe interactions between users and construct a thematic understanding of social behaviors on these platforms. Second, each domain expert had a different relationship to their community and local gangs and crews. Involving only two domain experts in our research may result in certain community-level understandings not being represented. We are considering a focus group model for involving domain expert insights in our future work, which may lead to a more representative and collective understanding of social media posts and the community as a whole. Finally, our methodology continued to evolve throughout the process, making it challenging to retrain our domain experts on new annotation systems and tasks, which may have hindered our ability to collect focused insights around the evolving domain-specific context. Capturing the contextual evolution of the domain is of the utmost importance in order to make informed and relevant interpretations of social media data.

Interpretations, whether from a domain expert or a trained social work annotator, may vary based on what is informing their knowledge of the domain. They may be based on biases, prior trauma, upbringing, or a variety of other circumstances. Through our work, it seems that there may not be one *right* answer regarding the meaning of a social media post. Domain experts may not have the same interpretations and insights due to each expert experiencing their neighborhood and lives differently. This makes it challenging to measure domain expertise and the expansive or narrow nature of insights from domain experts. However, the complexity of social media communication can only be uncovered through the involvement of people who have knowledge of the localized language, culture, and changing nature of community climate.

Conclusion

Domain experts must be involved in the interpretation of unstructured data, solution creation, and many other aspects of the research process. This goes beyond harvesting and capturing domain expertise. The involvement of domain experts in various areas of social and data science research, including mechanisms for accountability and ethically sound research practices, is a critical piece of

truly creating algorithms trained to support and protect marginalized youth and communities. If the gap between people who create algorithms and people who experience the direct impacts of them persists, we will likely continue to reinforce the very social inequities we hope to ameliorate.

Appendix



Figure A1. Examples of other tweets in our corpus.

Authors' Note

We would like to express our gratitude to our domain experts for providing their insights and sharing their experiences. We would also like to thank Eddie Bocanegra, Meg Helder, Owen Rambow, and Kathleen McKeown for their support and input on this study, including feedback from the anonymous reviewers. All tweets shared in this article (other than tweets from Gakirah Barnes) have been deidentified and made unsearchable. Due to the nature of the data and the vulnerable and marginalized population from which the data originates, data will only be made available through an application process and the signing of a memorandum of understanding (MOU) through the SAFElab at Columbia University (dp2787@columbia.edu).

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

Note

1. Top communicators were statistically calculated by most mentions and replies to Gakirah Barnes.

References

- Atkinson, R., & Flint, J. (2001). Accessing hidden and hard-to-reach populations: Snowball research strategies. *Social Research Update*, 33, 1–4.
- American Civil Liberties Union. (n.d.). Retrieved March 25, 2018, from <https://www.aclu.org/issues/racial-justice/race-and-criminal-justice>

- Anderson, E. (1999). *Code of the street* (pp. 107–141). New York, NY: Norton.
- Barlow, D. E., & Barlow, M. H. (2002). Racial profiling: A survey of African American police officers. *Police Quarterly*, 5, 334–358.
- Blevins, T., Kwiatkowski, R., Macbeth, J., McKeown, K., Patton, D., & Rambow, O. (2016). Automatically processing Tweets from gang-involved youth: Towards detecting loss and aggression. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers* (pp. 2196–2206). Osaka, Japan.
- Brunson, R. K., & Miller, J. (2006). Young black men and urban policing in the United States. *British Journal of Criminology*, 46, 613–640.
- Carter, W. M. Jr. (2014). Thirteenth amendment and constitutional change. *New York University Review of Law & Social Change*, 38, 583.
- Che, D., Safran, M., & Peng, Z. (2013). From big data to big data mining: Challenges, issues, and opportunities. In B. Hong, X. Meng, L. Chen, W. Winiwarter, & W. Song (Eds.), *International conference on database systems for advanced applications* (pp. 1–15). Berlin, Heidelberg: Springer.
- Christians, C. G., & Chen, S.-L. S. (2004). Introduction: Technological environments and the evolution of social research methods. In M. D. Johns, S. L. S. Chen, & G. J. Hall (Eds.), *Online social research. Methods, issues, & ethics* (pp. 15–23). New York, NY: Peter Lang.
- Citizens Crime Commission of New York City Researching Inequity in Society Ecologically Team. (2017). E-responder: A brief about preventing real world violence using digital intervention. Retrieved March 29, 2018, from <http://www.nycrimecommission.org/pdfs/e-responder-brief-1.pdf>
- Constine, J. (2017, November 27). Facebook rolls out AI to detect suicidal posts before they're reported. Retrieved November 29, 2017, from <https://techcrunch.com/2017/11/27/facebook-ai-suicide-prevention/>
- Decuyper, A. (2016). On the research for big data uses for public good purposes. Opportunities and challenges. *Réseaux Sociaux et Territoires*, 30, 305–314.
- Doshi-Velez, F., Kortz, M., Budish, R., Bavitz, C., Gershman, S., O'Brien, D., . . . Wood, A. (2017). Accountability of AI under the law: The role of explanation. arXiv preprint arXiv: 1711.01134.
- Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, 35, 137–144.
- Goffman, A. (2015). *On the run: Fugitive life in an American city*. London, England: Picador.
- Hackman, R. (2015, April 23). Is the online surveillance of black teenagers the new stop-and-frisk? Retrieved February 15, 2016, from <http://www.theguardian.com/us-news/2015/apr/23/online-surveillance-black-teenagers-new-stop-and-frisk>
- International Association of Chiefs of Police. (2015). 2015 Social media survey results. Retrieved from <http://www.socialmedia.org/Portals/1/documents/FULL%202015%20Social%20Media%20Survey%20Results.pdf>
- Khan, N., Yaqoob, I., Hashem, I. A. T., Inayat, Z., Mahmoud Ali, W. K., Alam, M., . . . Gani, A. (2014). Big data: Survey, technologies, opportunities, and challenges. *The Scientific World Journal*, 2014, 1–18.
- Labrinidis, A., & Jagadish, H. V. (2012). Challenges and opportunities with big data. *Proceedings of the VLDB Endowment*, 5, 2032–2033.
- Lane, J. (2016). The digital street: An ethnographic study of networked street life in Harlem. *American Behavioral Scientist*, 60, 43–58.
- Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J. R., Bethard, S., & McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations* (pp. 55–60). Retrieved from <http://www.anthology.aclweb.org/P/P14/P14-5010.pdf>
- Meehan, A. J., & Ponder, M. C. (2002). Race and place: The ecology of racial profiling African American motorists. *Justice Quarterly*, 19, 399–430.
- Owuputi, O., O'Connor, B., Dyer, C., Gimpel, K., Schneider, N., & Smith, N. A. (2013). Improved part-of-speech tagging for online conversational text with word clusters. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*

- (pp. 380–391). Stroudsburg, PA: Association for Computational Linguistics. Retrieved from <http://aclweb.org/anthology/N/N13/N13-1000.pdf>
- Patton, D. U., Brunton, D. W., Dixon, A., Miller, R. J., Leonard, P., & Hackman, R. (2017). Stop and frisk online: Theorizing everyday racism in digital policing in the use of social media for identification of criminal conduct and associations. *Social Media + Society*, 3, 1–10.
- Patton, D. U., Eschmann, R. D., & Butler, D. A. (2013). Internet banging: New trends in social media, gang violence, masculinity and hip hop. *Computers in Human Behavior*, 29, A54–A59. doi:10.1016/j.chb.2012.12.035
- Patton, D. U., Eschmann, R. D., Elsaesser, C., & Bocanegra, E. (2016). Sticks, stones and Facebook accounts: What violence outreach workers know about social media and urban-based gang violence in Chicago. *Computers in Human Behavior*, 65, 591–600.
- Patton, D. U., McKeown, K., Rambow, O., & Macbeth, J. (2016). Using natural language processing and qualitative analysis to intervene in gang violence: A collaboration between social work researchers and data scientists. *Bloomberg Data for Good Exchange Conference*, (pp. 1–5).
- Quan-Haase, A., & McCay-Peet, L. (2017). Building interdisciplinary social media research teams: Benefits, challenges, and policy frameworks. In L. Sloan & A. Quan-Haase (Eds.), *Handbook of social media research methods* (pp. 41–56). London, England: Sage.
- Rizkallah, J. (2017, October 12). The big (unstructured) data problem. Retrieved November 16, 2017, from <https://www.forbes.com/sites/forbestechcouncil/2017/06/05/the-big-unstructured-data-problem/#1bdc09dc493a>
- Seitz, J. (2016, January 11). Automatically finding weapons in social media images part 1. Retrieved November 16, 2017, from <https://www.bellingcat.com/resources/2016/01/11/automatically-finding-weapons-in-social-media-images-part-1/>
- Suthaharan, S. (2014). Big data classification: Problems and challenges in network intrusion prediction with machine learning. *ACM SIGMETRICS Performance Evaluation Review*, 41, 70–73.
- Talluri, S. (2016). Big data using cloud technologies. *Global Journal of Computer Science and Technology*, 16. Retrieved from <https://computerresearch.org/index.php/computer/article/view/1395>

Author Biographies

William R. Frey is a doctoral student, qualitative researcher, and digital ethnographer in the SAFElab at Columbia University's School of Social Work. He holds an MSW from University of Michigan's School of Social Work.

Desmond U. Patton is an associate professor and director of the SAFElab at Columbia University's School of Social Work and a fellow at the Berkman Klein Center for Internet and Society at Harvard University. He is an expert in social media communication and youth violence.

Michael B. Gaskell is a postdoctoral research scientist in the SAFElab at Columbia University's School of Social Work. He holds a PhD in clinical psychology from Xavier University.

Kyle A. McGregor is an assistant professor in the Department of Child and Adolescent Psychology and Department of Population Health at New York University's Langone Health. He is also the director of Pediatric Mental Health Ethics.