The most shocking example occurred a week ago when the extremist group al-Shabab live-tweeted about the mall siege in Kenya, defending the mass killing, threatening more violence and taunting the military.  Criminals and gang members are advertising their wares, flaunting their exploits and recruiting new members in 140 characters or less. Extremists spread their propaganda via video. Gangs post their colors, signs and rap songs to showcase their criminal enterprises. Prostitutes and drug dealers troll for new customers. Teens trash a former NFL player's house and brag about it with photos on Twitter. But while Twitter can serve as a valuable recruitment and communications tool, it also can be a double-edged sword: Public boasting about illegal deeds can serve as a road map for law enforcement officials and lead to arrests. Twitter helped spur a 30% growth in online forums for hate and terrorism over the past year.

**INTERNET-BANGING AMPLIFIED**

A March study by Arizona State University criminologist Scott Decker found that nearly 20% of gang members reported that their gang had a website or social networking page and 50% said that their gang posts video online. gangs involved in drug dealing use Twitter, but because police know the corners and other spots where transactions generally take place, gang members will tweet out an address. He says the context of the tweet is unclear to a lay person, but the person on the receiving end understands the message.

**DECIPHERING THE CODE**

Rob D'Ovidio, a Drexel University criminologist, says gang members use code to boast about their deeds. For example, he says, they use "biscuit" or "clickety" for a gun, "food," "sea shells" or "gas" for bullets and "rock to sleep early" for murder. He says street gangs are crafty in their online recruitment techniques. The gangs associate their group with popular music that has a violent message or a message that portrays ethnic oppression, which leads youngsters to believe they have something in common with the gang, he says.

**Literature Survey**

<https://pdfs.semanticscholar.org/b1cd/a279336ff6072d80fb259f5126dc8ca8eb9a.pdf>

<https://www.rsaconference.com/writable/presentations/file_upload/cct2-w05_a_framework_for_analyzing_twitter_to_detect_community_crime_activity_final.pdf>

<https://www.vice.com/en_us/article/53npza/the-haunting-social-media-trail-left-by-a-teen-gang-member>

Tweets, Gangs and Guns: A Snapshot of Gang Communications in Detroit

[Desmond U. Patton](https://www.ncbi.nlm.nih.gov/pubmed/?term=Patton%20DU%5BAuthor%5D&cauthor=true&cauthor_uid=28810937), [Sadiq Patel](https://www.ncbi.nlm.nih.gov/pubmed/?term=Patel%20S%5BAuthor%5D&cauthor=true&cauthor_uid=28810937), [Jun Sung Hong](https://www.ncbi.nlm.nih.gov/pubmed/?term=Hong%20JS%5BAuthor%5D&cauthor=true&cauthor_uid=28810937), [Megan Ranney](https://www.ncbi.nlm.nih.gov/pubmed/?term=Ranney%20M%5BAuthor%5D&cauthor=true&cauthor_uid=28810937), [Marie Crandal](https://www.ncbi.nlm.nih.gov/pubmed/?term=Crandal%20M%5BAuthor%5D&cauthor=true&cauthor_uid=28810937), and [Lyle Dungy](https://www.ncbi.nlm.nih.gov/pubmed/?term=Dungy%20L%5BAuthor%5D&cauthor=true&cauthor_uid=28810937)

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5640454/>

Violence and crime-related communications fell into four main categories: (1) Beefing (267,221 tweets); (2) Grief (79,971 tweets); (3) Guns (3,551 tweets); and (4) Substance use and distribution (47,638 tweets). Patterns in violent and criminal communication that may be helpful in predicting future gang activities were identified, which has implications for violence-prevention research, practice, and policy.

Determining the prevalence of gang communication on social media requires an in-depth understanding of the socio-linguistic variations of Standard English. As such, we take an in-depth, qualitative approach to identifying specific keywords, phrases and content that would most likely represent gang communication. Online images glorify gang culture and possibly encourage gang participation offline. However, gang behavior is generally more symbolic than instrumental, suggesting that posts are often used for marketing gang activity, as opposed to recruitment.  Gang presence on social media was most closely linked to promoting gang lifestyle or street culture through individual displays. Twitter profiles tend to depict a way of life, rather than promoting criminal activities. Many gang members in this study posted pictures of money, guns, women, parties, drugs, and alcohol. The authors noted that some individuals did promote their gang affiliations by praising achievements or expressing their allegiance.

In this study, the DCC, in collaboration with the study authors, performed an advanced keyword search of 8.5 million publicly available tweets using an extensive five step process, which included: (1) identifying key terms and phrases that may be associated with gang culture and behavior, (2) cross referencing the list of key terms with websites for additional context, (3) refining the list of keywords and phrases for a deeper focus on Detroit-based gangs, 4) identifying Twitter users based on key terms, and (5) refining that list of users based on key terms and DCC intelligence.

**\*\*\* FINDING STREET GANG MEMBER PROFILES ON TWITTER**

<https://etd.ohiolink.edu/!etd.send_file?accession=wright1516054679956178&disposition=inline>

CHALLENGE: Many Twitter profile classifiers search for contextual clues in tweets and profile descriptions, but gang member profiles use a rapidly changing lexicon of keywords and phrases that often have only a local, geographic context. Given the very local, rapidly changing lexicon of gang members on social media, building a database of keywords, phrases, and other identifiers to find gang members nationally is not feasible.

**Detecting Gang-Involved Escalation on Social Media Using Context**

<https://aclweb.org/anthology/D18-1005>

**Automatically Processing Tweets from Gang-Involved Youth: Towards Detecting Loss and Aggression**

<https://www.aclweb.org/anthology/C16-1207>

**FRAMEWORK FOR ANALYZING TWITTER TO DETECT COMMUNITY SUSPICIOUS CRIME ACTIVITY Safaa.S Al Dhanhani**

<https://pdfs.semanticscholar.org/b1cd/a279336ff6072d80fb259f5126dc8ca8eb9a.pdf>

Investigating crimes based on statistical and graph analysis on Twitter data. Our solution supports an investigative processes composed of the following phases (i) find suspicious tweets and individuals based on hash tags analysis (ii) classify the user profile based on Twitter features (iii) identify influencers in the FOAF networks of the senders (iiii) analyze these influencers’ background and history to find hints of past or current criminal activity.

1. **Twitter User Account Analysis**
2. **2 Detection of Events Based on Twitter Communication**
3. **Predicting Crime Location Using Linear Regression**
4. **Detecting Crimes Based on Nodes Analysis**
5. **Prediction Based on Twitter Data Analysis**
6. **Prediction Based on Sentiment Analysis**

**Technology**

1. Garden hose streaming API or Followerwonk Web service API
2. Word Embedding Models
3. VADER for sentiment Analysis (Valence Aware Dictionary and sentiment Reasoner) <https://github.com/cjhutto/vaderSentiment>
4. Textblob
5. NetworkX
6. <https://towardsdatascience.com/the-real-world-as-seen-on-twitter-sentiment-analysis-part-one-5ac2d06b63fb>

**List of identified gangs by FBI:**

Mooda Crowd\_  
LilAntFrmOMB  
Javarri\_  
trouble\_thf  
Free Lil Gudda  
4546\_thf  
Jdog\_fromDaHit  
Moe\_Gotti2x  
ZaysavageEBK  
Kamari\_Keeper  
big600\_manny  
Big600Manny\_  
Big600Booka  
Oso\_3\_Much  
ahmed\_bangbang  
RaheemGang  
KukuGang\_064  
lil800DASHOOTER  
lildurk  
Sixhuncho  
@rockapopSumob41  
@dumpstreet\_Tazz  
@LevOnFrm6200  
@antoniodorsey41  
@OsoArrogantZayy  
@AstronautPrice

**Problem Outlook**

This doc serves as a baseline of words and graphics that we want to identify in a message. To keep it simple, say a message uses 3 of the words in the baseline then we want to give it a yellow priority. But also each word may have a different "weight" meaning kill or bullet may have a weight of 10 while candy has a weight of 3. So the message "I'm going to kill you with so many bullets" will get a higher "score" than "I'm selling candy at 4pm". Thus messages with a higher score will need to be prioritized for analysis.

One model can help identify the context of a message by telling the difference between "Did you see the number of hits I got in that game" and "be careful you don't get any hits to the back". Since right now, it has to be done manually for all 25,000 messages that are gathered each day.

**Files Description**

1. Chi-Data: Contains 2 csv files ideco.csv and idecoNov30.csv ideco is big file.
2. Ideco\_ni\_decoded.csv/ IdecoNov30\_ni\_decoded.csv: There are 2 columns ni and raw. So, ni is processed and tweets are captured but they are unclean and in json format.
3. Signal\_keywords: File contains ‘Alert Term Templates’ and keywords used by gangs. Extracted from file Voice4Impact Baseline Words\_Graphics.docx

**New Learning**

1. **Network Analysis**

<https://www.datacamp.com/community/tutorials/social-network-analysis-python>

Social network analysis (SNA) using Python and NetworkX, a Python library for the study of the structure, dynamics, and functions of complex networks.

Each network consists of:

* **Nodes**: The individuals whose network we are building. Actors in the above example.
* **Edges**: The connection between the nodes. It represents a relationship between the nodes of the network. In our example, the relationship was that the actors have worked together.

1. **Vader Sentiment Analysis**

**VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text**

<http://comp.social.gatech.edu/papers/icwsm14.vader.hutto.pdf>

Summary:  
(1) Uses existing lexicons and added emojis and abbreviations  
(2)  Ground truth obtained from crowd sourcing for each token (word/emoji/abbreviation). The score is from -4 to 4. With -4 representing a extremely negative token and 4, extremely positive token.  
added rules:  
(a) Punctuation (! is more extreme than . )  
(b) Capitalising indicates intensity.  
(c) checks Degree modifies (e.g words such as extremely)  
(d)  checks contrasting conjuction (such as 'but')  
(e) uses tri-grams

I also looked at their vader\_lexicon.txt to see how we can add our keywords:  
Here is the structure of the txt file  
Token | mean | std | Ground truth |  
(1) Token is the word/ emoji/ abbreviation  
(2) mean of the ground truth scores  
(3) standard deviation of the ground truth has to be (std <= 2.5).  
(4) 10 scores (between -4 and 4) from crowd sourcing. This is taken as the ground truth.

**Eigenvector centrality**

Eigenvector centrality is used to measure the level of influence of a node within a network. Each node within the network will be given a score or value: the higher the score the greater the level of influence within the network. This score is relative to the number of connections a node will have to other nodes. Connections to high-scoring eigenvector centrality nodes contribute more to the score of the node than equal connections to low-scoring nodes.

To put this into context, a node with a high degree score (i.e. many connections) may only have a relatively low eigenvector centrality score because many of those connections are with similarly low-scored nodes. Also, a node may have a high betweenness score (indicating it connects disparate parts of a network) but a low Eigenvector Centrality score because it’s still some distance from the centres of power in the network.

You would use eigenvector centrality to identify who or what has a wide-reaching influence within a given network.

### The Pagerank

The pagerank is a variant of the Eigenvector centrality score, but because it uses backlinks/in-degrees it is used in directed networks. That said, there are no “hard or fast” rules stating you can only use pagerank when analysing directed networks, eigenvector centrality will work, but you will need to justify your answer. It does however make no sense in using pagerank scores in undirected networks.

Like eigenvector centrality, the pagerank can be considered as the “importance score” of a web page or social network node. This importance score will always be a non-negative real number and all the scores will add to 1.This score is based on the links made to that page/node from other pages/nodes. The links to a given page/nodes are called the backlinks/in-degrees for that page/node. The web/social network thus becomes a democracy where pages/nodes vote for the importance of other pages by linking to them.

<https://cambridge-intelligence.com/eigencentrality-pagerank/>

Two approaches for the measure of the influence of a node in a network.

Eigen Centrality

Like degree centrality, EigenCentrality measures a node’s influence by counting the number of links it has to other nodes within the network. However, EigenCentrality goes a step further by also taking into account how well connected a node is, and how many links their connections have, and so on through the network.

#### What does EigenCentrality tell me?

A high EigenCentrality score indicates a strong influence over other nodes in the network. It is useful because it indicates not just direct influence, but also implies influence over nodes more than one ‘hop’ away. A node may have a high degree score (i.e. many connections) but a relatively low EigenCentrality score if many of those connections are with similarly low-scored nodes.

Also, a node may have a high betweenness score (indicating it connects disparate parts of a network) but a low EigenCentrality score because it is still some distance from the centers of power in the network.

PageRank: The Google Algorithm

PageRank is a variant of EigenCentrality, designed and made famous by Google founders Larry Page and Sergei Brin. Designed for ranking webpages, PageRank uses links between pages as a measure of importance. Each webpage is treated as a node in a network, and is assigned a score based upon its number of in-coming links (its ‘indegree’). These links are also weighted depending on the relative score of its originating node.The result is that nodes with many in-coming links are influential, and nodes to which they are connected share some of that influence.

#### What does PageRank tell me?

Like EigenCentrality, PageRank can help uncover influential or important nodes whose reach extends beyond just their direct connections.The main difference to EigenCentrality, is that PageRank takes link direction and weight into account.