HHS Public Access

Author manuscript

Proc IEEE ACM Int Conf Adv Soc Netw Anal Min. Author manuscript; available in PMC 2017 July 13.

Published in final edited form as:

Proc IEEE ACM Int Conf Adv Soc Netw Anal Min. 2016 August ; 206: 685–692. doi:10.1109/ASONAM.2016.7752311.

Finding Street Gang Members on Twitter

Lakshika Balasuriya,
Kno.e.sis Center, Wright State University, Dayton OH, USA

Sanjaya Wijeratne,
Kno.e.sis Center, Wright State University, Dayton OH, USA

Derek Doran, and
Kno.e.sis Center, Wright State University, Dayton OH, USA

Amit Sheth
Kno.e.sis Center, Wright State University, Dayton OH, USA

Abstract

Most street gang members use Twitter to intimidate others, to present outrageous images and statements to the world, and to share recent illegal activities. Their tweets may thus be useful to law enforcement agencies to discover clues about recent crimes or to anticipate ones that may occur. Finding these posts, however, requires a method to discover gang member Twitter profiles. This is a challenging task since gang members represent a very small population of the 320 million Twitter users. This paper studies the problem of automatically finding gang members on Twitter. It outlines a process to curate one of the largest sets of verifiable gang member profiles that have ever been studied. A review of these profiles establishes differences in the language, images, YouTube links, and emojis gang members use compared to the rest of the Twitter population. Features from this review are used to train a series of supervised classifiers. Our classifier achieves a promising F_1 score with a low false positive rate.

Keywords

Street Gangs; Twitter Profile Identification; Gang Activity Understanding; Social Media Analysis

I. Introduction and Motivation

The crime and violence street gangs introduce into neighborhoods is a growing epidemic in cities around the world[1]. Today, over 1.23 million people in the United States are members of a *street gang* [1], which is a coalition of peers, united by mutual interests, with identifiable leadership and internal organization, who act collectively to conduct illegal activity and to control a territory, facility, or enterprise [2]. They promote criminal activities such as drug trafficking, assault, robbery, and threatening or intimidating a neighborhood [3]. Moreover, data from the Centers for Disease Control in the United States suggests that

[1]http://goo.gl/OjWeYf

the victims of at least 1.3% of all gang-related[2] homicides are merely innocent bystanders who live in gang occupied neighborhoods [4].

Street gang members have established online presences coinciding with their physical occupation of neighborhoods. The National Gang Threat Assessment Report confirms that at least tens of thousands of gang members are using social networking websites such as Twitter and video sharing websites such as YouTube in their daily life [1]. They are very active online; the 2007 National Assessment Center's survey of gang members found that 25% of individuals in gangs use the Internet for at least 4 hours a week [5]. Gang members typically use social networking sites and social media to develop online respect for their street gang [6] and to post intimidating, threatening images or videos [7]. This "Cyber-" or "Internet banging" [8] behavior is precipitated by the fact that an increasing number of young members of the society are joining gangs [9], and these young members have become enamored with technology and with the notion of sharing information quickly and publicly through social media[3]. Stronger police surveillance in the physical spaces where gangs congregate further encourages gang members to seek out virtual spaces such as social media to express their affiliation, to sell drugs, and to celebrate their illegal activities [10].

Gang members are able to post publicly on Twitter without fear of consequences because there are few tools law enforcement can use to surveil this medium [11]. Police departments across the United States instead rely on manual processes to search social media for gang member profiles and to study their posts. For example, the New York City police department employs over 300 detectives to combat teen violence triggered by insults, dares, and threats exchanged on social media, and the Toronto police department teaches officers about the use of social media in investigations [12]. Officer training is broadly limited to understanding policies on using Twitter in investigations and best practices for data storage [13]. The safety and security of city neighborhoods can thus be improved if law enforcement were equipped with intelligent tools to study social media for gang activity.

The need for better tools for law enforcement cannot be underscored enough. Recent news reports have shown that many incidents involving gangs start on Twitter, escalate over time, and lead to an offline event that could have been prevented by an early warning. For example, the media reported on a possible connection between the death of a teenage rapper from Illinois and the final set of tweets he posted. One of his last tweets linked to a video of him shouting vulgar words at a rival gang member who, in return, replied *"I'ma kill you"* on social media[4]. In a following tweet, the teenage rapper posted *"im on 069"*, revealing his location, and was shot dead soon after that post. Subsequent investigation revealed that the rivalry leading to his death began and was carried out entirely on social media. Other reporting has revealed how innocent bystanders have also become targets in online fights, leaving everyone in a neighborhood at risk[5].

[2]The terms 'gang' and 'street gang' are used interchangeably in this paper.
[3]http://www.hhs.gov/ash/oah/news/e-updates/eupdate-nov-2013.html
[4]http://www.wired.com/2013/09/gangs-of-social-media/
[5]https://goo.gl/75U3ME

This paper investigates whether gang member profiles can be identified automatically on Twitter, which can enable better surveillance of gang members on social media. Classifying Twitter profiles into particular types of users has been done in other contexts [14], [15], [16], but gang member profiles pose unique challenges. For example, many Twitter profile classifiers search for contextual clues in tweets and profile descriptions [17], but gang member profiles use a rapidly changing lexicon of keywords and phrases that often have only a local, geographic context. This is illustrated in Figure 1, which shows the Twitter profile descriptions of two verified deceased gang members. The profile of @OsoArrogantJoJo provides evidence that he belongs to a rival gang of the Black Disciples by #BDK, a hashtag that is only known to those involved with gang culture in Chicago. @PappyNotPapi's profile mentions #PBG and our investigations revealed that this hashtag is newly founded and stands for the Pooh Bear Gang, a gang that was formerly known as the Insane Cutthroat Gangsters. 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. Instead, this study proposes heterogeneous sets of features derived not only from profile and tweet text but also from the emoji usage, profile images, and links to YouTube videos reflecting their music culture. A large set of gang member profiles, obtained through a careful data collection process, is compared against non-gang member profiles to find contrasting features. Experimental results show that using these sets of features, we can build a classifier that has a low false positive rate and a promising $F1$-score of 0.7755.

This paper is organized as follows. Section II discusses the related literature and positions how this work differs from other related works. Section III discusses the data collection, manual feature selection and our approach to identify gang member profiles. Section IV gives a detailed explanation for evaluation of the proposed method and the results in detail. Section V concludes the work reported while discussing the future work planned.

II. Related Work

Gang violence is a well studied social science topic dating back to 1927 [18]. However, the notions of "Cyber-" or "Internet banging", which is defined as *"the phenomenon of gang affiliates using social media sites to trade insults or make violent threats that lead to homicide or victimization"* [8], was only recently introduced [19], [11]. Patton *et al.* introduced the concept of "Internet banging" and studied how social media is now being used as a tool for gang self-promotion and as a way for gang members to gain and maintain street credibility [8]. They also discussed the relationship between gang-related crime and hip-hop culture, giving examples on how hip-hop music shared on social media websites targeted at harassing rival gang members often ended up in real-world collisions among those gangs. Decker *et al.* and Patton *et al.* have also reported that street gangs perform Internet banging with social media posts of videos depicting their illegal behaviors, threats to rival gangs, and firearms [20], [21].

The ability to take action on these discoveries is limited by the tools available to discover gang members on social media and to analyze the content they post [19]. Recent attempts to improve our abilities include a proposed architecture for a surveillance system that can learn

the structure, function, and operation of gangs through what they post on social media [11]. However, the architecture requires a set of gang member profiles for input, thus assuming that they have already been discovered. Patton *et al.* [21] devised a method to automatically collect tweets from a group of gang members operating in Detroit, MI. However, their approach required the profile names of the gang members to be known beforehand, and data collection was localized to a single city in the country.

This work builds upon existing methods to automatically discover gang member profiles on Twitter. This type of user profile classification problem has been explored in a diverse set of applications such as political affiliation [14], ethnicity [14], gender [16], predicting brand loyalty [14], and user occupations [17]. However, these approaches may utilize an abundance of positive examples in their training data, and only rely on a single feature type (typically, tweet text). Whereas most profile classifiers focus on a single *type* of feature (e.g. profile text), we consider the use of a variety of feature types, including emoji, YouTube links, and photo features.

III. Discovering Gang Member Profiles

This section discusses the methodology we followed to study and classify the Twitter profiles of gang members automatically. It includes a semi-automatic data collection process to discover a large set of verifiable gang member profiles, an evaluation of the tweets of gang and non-gang member posts to identify promising features, and the deployment of multiple supervised learning algorithms to perform the classification.

A. Data collection

Discovering gang member profiles on Twitter to build training and testing datasets is a challenging task. Past strategies to find these profiles were to search for keywords, phrases, and events that are known to be related to gang activity in a particular city a priori [11], [21]. However, such approaches are unlikely to yield adequate data to train an automatic classifier since gang members from different geographic locations and cultures use local languages, location-specific hashtags, and share information related to activities in a local region [11]. Such region-specific tweets and profiles may be used to train a classifier to find gang members within a small region but not across the Twitterverse. To overcome these limitations, we adopted a semi-automatic workflow, illustrated in Figure 2, to build a dataset of gang member profiles suitable for training a classifier. The steps of the workflow are:

1. **Seed Term Discovery:** Following the success of identifying gang member profiles from Chicago [11], we began our data collection with discovering universal terms used by gang members. We first searched for profiles with hashtags for Chicago gangs noted in [11], namely `#BDK` (Black Disciple Killers) and `#GDK` (Gangster Disciples Killers). Those pro-files were analyzed and manually verified as explained in Step 3. Analysis of these profiles identified a small set of hashtags they all use in their profile descriptions. Searching Twitter profiles using those hashtags, we observed that gang members across the U.S. use them, thus we consider those terms to be location neutral. For example, gang members post `#FreeDaGuys` in their profile to support their fellow members

who are in jail, `#RIPDaGuys` to convey the grieving for fallen gang members, and `#FuckDaOpps` to show their hatred towards police officers. We used these terms as keywords to discover Twitter profiles irrespective of geographical location. We used the Followerwonk Web service API[6] and Twitter REST API[7] to search Twitter profile descriptions by keywords `#FreeDaGuys`, `#FreeMyNigga`, `#RIPDaGuys`, and `#FuckDaOpps`. Since there are different informal ways people spell a word in social media, we also considered variations on the spelling of each keyword; for example, for `#FreeDaGuys`, we searched both `#FreeDaGuys`, and #FreeTheGuys.

2. **Gang Affiliated Rappers' Twitter Profile Discovery:** Finding profiles by a small set of keywords is unlikely to yield sufficient data. Thus, we sought additional gang member profiles with an observation from Patton *et al.* [8] that the influence of hip-hop music and culture on offline gang member activities can also be seen in their social media posts. We thus also consider the influence of hip-hop culture on Twitter by exploring the Twitter network of known gangster rappers who were murdered in 2015 due to gang-related incidents[8]. We searched for these rapper profiles on Twitter and manually checked that the rapper was affiliated to a gang.

3. **Manual verification of Twitter profiles:** We verified each profile discovered manually by examining the profile picture, profile background image, recent tweets, and recent pictures posted by a user. During these checks, we searched for terms, activities, and symbols that we believed could be associated with a gang. For example, profiles whose image or background included guns in a threatening way, stacks of money, showing gang hand signs and gestures, and humans holding or posing with a gun, appeared likely to be from a gang member. Such images were often identified in profiles of users who submitted tweets that contain messages of support or sadness for prisoners or recently fallen gang members, or used a high volume of threatening and intimidating slang language. Only profiles where the images, words, and tweets all suggested gang affiliation were labeled as gang affiliates and added to our dataset. Although this manual verification does have a degree of subjectivity, in practice, the images and words used by gang members on social media are so pronounced that we believe any reasonable analyst would agree that they are gang members. We found that not all the profiles collected belonged to gang members; we observed relatives and followers of gang members posting the same hashtags as in Step 1 to convey similar feelings in their profile descriptions.

4. **Using Retweets to discover more profiles:** From the set of verified profiles, we explored their retweet and follower networks as a way to expand the dataset. We first considered authors of tweets which were retweeted by a gang member in our seed set. In Twitter, *"retweeting"* is a mechanism by which a user can share

[6] https://moz.com/followerwonk/bio
[7] https://dev.twitter.com/rest/public
[8] http://www.hipwiki.com/List+of+Rappers+Murdered+in+2015

someone else's tweet to their follower audience. Assuming that a user only retweets things that they believe or their audience would be interested in, it may be reasonable to assume that gang members would only be interested in sharing what other gang members have to say, and hence, the authors of gang members' retweets could also be gang members.

5. **Using Followers and Followees to discover more profiles:** We analyzed followers and followees of our seed gang member profiles to find more gang member profiles. A Twitter user can follow other Twitter users so that the individual will be subscribed to their tweets as a follower and they will be able to start a private conversation by sending direct messages to the individual. Motivated by the sociological concept of homophily, which claims that individuals have a tendency to associate and bond with similar others[9], we hypothesized that the followers and followees of Twitter profiles from the seed set may also be gang members. Manual verification of Twitter profiles collected from retweets, followers, and followees of gang members showed that a majority of those profiles are non-gang members who are either family members, hip-hop artists, women or profiles with pornographic content. To ensure that our dataset is not biased towards a specific gang or geographic location, only a limited number of profiles were collected via retweets, followers and followees.

Table I summarizes the number of profiles manually verified as gang members from Twitter profiles collected in step 1, 2, 4 and 5. Altogether we collected 400 gang member's Twitter profiles. This is a large number compared to previous studies of gang member activities on social media that curated a maximum of 91 profiles [11]. Moreover, we believe the profiles collected represent a diverse set of gang members that are not biased toward a particular geographic area or lingo as our data collection process used location-independent terms proven to be used by gang members when they express themselves.

B. Data analysis

We next explore differences between gang and non-gang member Twitter usage to find promising features for classifying profiles. For this purpose, profiles of non-gang members were collected from the Twitter Streaming API[10]. We collected a random sample of tweets and the profiles of the users who authored the tweets in the random sample. We manually verified that all Twitter profiles collected in this approach belong to non-gang members. The profiles selected were then filtered by location to remove non-U.S. profiles by reverse geo-coding the location stated in their profile description by the Google Maps API[11]. Profiles with location descriptions that were unspecified or did not relate to a location in the U.S. were discarded. We collected 2,000 non-gang member profiles in this manner. In addition, we added 865 manually verified non-gang member profiles collected using the location neutral keywords discussed in Section III. Introducing these profiles, which have some characteristics of gang members (such as cursing frequently or cursing at law enforcement)

[9] http://aris.ss.uci.edu/~lin/52.pdf
[10] https://dev.twitter.com/streaming/overview
[11] https://developers.google.com/maps/

but are not, captures local languages used by family/friends of gang members and ordinary people in a neighborhood where gangs operate.

With the Twitter REST API[12], we collected the maximum number of most recent tweets that can be retrieved (3,200) along with profile descriptions and images (profile and cover photos) of every gang and non-gang member profile. The resulting dataset consists of 400 gang member Twitter profiles and 2,865 non-gang member Twitter profiles. The dataset has a total of 821,412 tweets from gang member profiles and 7,238,758 tweets from non-gang member profiles. Prior to analyzing any text content, we removed all of the seed words used to find gang member profiles, all stop words, and performed stemming across all tweets and profile descriptions.

1. *Tweet text:* Figure 3 summarizes the words seen most often in the gang and non-gang members' tweets as clouds. They show a clear difference in language. For example, we note that gang members more frequently use curse words in comparison to ordinary users. Although cursing is frequent in tweets, they represent just 1.15% of all words used [22]. In contrast, we found 5.72% of all words posted by gang member accounts to be classified as a curse word, which is nearly five times more than the average curse word usage on Twitter. The clouds also reflect the fact that gang members often talk about drugs and money with terms such as *smoke, high, hit*, and *money*, while ordinary users hardly speak about finances and drugs. We also noticed that gang members talk about material things with terms such as *got, money, make, real, need* whereas ordinary users tend to vocalize their feelings with terms such as *new, like, love, know, want, look, make, us*. These differences make it clear that the individual words used by gang and non-gang members will be relevant features for gang profile classification.

2. *Twitter Profile Description:* On Twitter, a user can give a self-description as a part of the user's profile. A comparison of the top 10 words in gang members' and non-gang members' Twitter profile descriptions is shown in Figure 4. The first 10 words are the most frequently used words in non-gang members' profiles and the latter 10 words are the most frequently used words in gang members' profiles. Word comparison shows that gang members prefer to use curse words (*nigga, fuck, shit*) in their profile descriptions while non-gang members use words related to their feelings or interests (*love, life, live, music, book*). The terms *rip* and *free* which appear in approximately 12% of all gang member Twitter profiles, suggest that gang members use their profile descriptions as a space to grieve for their fallen or incarcerated gang members. The term *gang* in gang members' profile descriptions suggest that gang members like to self-identify themselves on Twitter. Such lexical features may therefore be of great importance for automatically identifying gang member profiles. We take counts of unigrams from gang and non-gang members' Twitter profile descriptions as classification features.

[12] https://dev.twitter.com/rest/public

3. *Music interests:* It has been recognized that music is a key cultural component in an urban lifestyle and that gang members often want to emulate the scenarios and activities the music conveys [8]. Our analysis confirms that the influence of gangster rap is expressed in gang members' Twitter posts. We found that 51.25% of the gang members collected have a tweet that links to a YouTube video. Following these links, a simple keyword search for the terms `gangsta` and `hip-hop` in the YouTube video description found that 76.58% of the shared links are related to hip-hop music, gangster rap, and the culture that surrounds this music genre. Moreover, this high proportion is not driven by a small number of profiles that prolifically share YouTube links; eight YouTube links are shared on average by a gang member.

   Recognizing the frequency with which gang members post YouTube links on gangster rap and hip-hop, we consider the YouTube videos posted in a user's tweets as features for the classifier. In particular, for each YouTube video tweeted, we used the YouTube API[13] to retrieve the video's description and its comments. Further analysis of YouTube data showed a difference between terms in gang members' YouTube data and non-gang members' YouTube data. For example, the top 5 terms (after stemming and stop word removal) used in YouTube videos shared by gang members are *shit, like, nigga, fuck, lil* while *like, love, peopl, song, get* are the top 5 terms in non-gang member video data. To represent a user profile based on their music interests, we generated a bag of words from the video descriptions and comments from all shared videos.

4. *Emoji:* Motivated by recent work involving the use of emojis by gang members [23], we also studied if and how gang and non-gang members use emoji symbols in their tweets. Our analysis found that gang members have a penchant for using just a small set of emoji symbols that convey their anger and violent behavior through their tweets. Figure 5 illustrates the emoji distribution for the top 20 most frequent emojis used by gang member profiles in our dataset. The fuel pump emoji was the most frequently used emoji by the gang members, which is often used in the context of selling or consuming marijuana. The pistol emoji is the second most frequent in our dataset, which is often used with the guardsman emoji or the police cop emoji in an 'emoji chain'. Figure 6 presents some prototypical 'chaining' of emojis used by gang members. The chains may reflect their anger at law enforcement officers, as a cop emoji is often followed up with the emoji of a weapon, bomb, or explosion. We found that 32.25% of gang members in our dataset have chained together the police and the pistol emoji, compared to just 1.14% of non-gang members. Moreover, only 1.71% of non-gang members have used the hundred points emoji and pistol emoji together in tweets while 53% of gang members have used them. A variety of the angry face emoji such as devil face emoji and imp emoji were also common in gang member tweets. The frequency of each emoji symbol used across the set of user's tweets are thus considered as features for our classifier.

[13] https://developers.google.com/youtube/

5. *Profile image:* In our profile verification process, we observed that most gang member profiles portray a context representative of gang culture. Some examples of these profile pictures are shown in Figure 7, where the user holds or points weapons, is seen in a group fashion which displays a gangster culture, or is showing off graffiti, hand signs, tattoos and bulk cash. Descriptions of these images may thus empower our classifier. Thus, we translated profile images into features with the Clarifai web service[14]. Clarifai offers a free API to query a deep learning system that tags images with a set of scored keywords that reflect what is seen in the image. We tagged the profile image and cover image for each profile using 20 tags identified by Clarifai. Figure 8 offers the 20 most often used tags applied to gang and non-gang member profiles. Since we take all the tags returned for an image, we see common words such as *people and adult* coming up in the top 20 tag set. However, gang member profile images were assigned unique tags such as *trigger, bullet, worship* while non-gang images were uniquely tagged with *beach, seashore, dawn, wildlife, sand, pet*. The set of tags returned by Clarifai were thus considered as features for the classifier.

C. Learning algorithms

The unigrams of tweets, profile text, and linked YouTube video descriptions and comments, along with the distribution of emoji symbols and the profile image tags were used to train four different classification models: a Naive Bayes net, a Logistic Regression, a Random Forest, and a Support Vector Machine (SVM). These four models were chosen because they are known to perform well over text features, which is the dominant type of feature considered. The performance of the models are empirically compared to determine the most suitable classification technique for this problem. Data for the models are represented as a vector of term frequencies where the terms were collected from one or more feature sets described above.

IV. Evaluation

We next evaluate the performance of classifiers that use the above features to discover gang member profiles on Twitter. For this purpose, we use the training set discussed in Section III with 400 gang member profiles (the ‘positive’/‘gang’ class) and 2,865 non-gang member profiles (the ‘negative’/‘non-gang’ class). We trained and evaluated the performance of the classifiers mentioned in Section III-C under a 10-fold cross validation scheme. For each of the four learning algorithms, we consider variations involving only tweet text, emoji, profile, image, or music interest (YouTube comments and video description) features, and a final variant that considers all types of features together. The classifiers that use a single feature type were intended to help us study the quality of their predictive power by itself. When building these single-feature classifiers, we filtered the training dataset based on the availability of the single feature type in the training data. For example, we only used the twitter profiles that had at least a single emoji in their tweets to train classifiers that consider

[14] http://www.clarifai.com/

emoji features. We found 3,085 such profiles out of the 3,265 profiles in the training set. When all feature types were considered, we developed two different models:

1. *Model(1)*: This model is trained with all profiles in the training set.
2. *Model(2)*: This model is trained with profiles that contain *every* feature type.

Because a Twitter profile may not have every feature type, *Model(1)* represents a practical scenario where not every Twitter profile contains every type of feature. In this model, the non-occurrence of a feature is represented by 'zeroing out' the feature value during model training. *Model(2)* represents the ideal scenario where all profiles contain *every* feature type. For this model, we used 1,358 training instances (42% of all training instances), out of which 172 were gang members (43% of all gang members) and 1,186 were non-gang members (41% of all non-gang members). We used version 0.17.1 of scikit-learn[15] machine learning library to implement the classifiers.

For each 10-fold cross validation experiment, we report three evaluation metrics for the 'gang' and 'non-gang' classes, namely, the Precision = $tp/(tp + fp)$, Recall = $tp/(tp + fn)$, and $F1\text{-score} = 2 * (Precision * Recall)/(Precision + Recall)$, where tp is the number of true positives, fp is the number of false positives, tn is the number of true negatives, and fn is the number of false negatives. We report these metrics for the positive 'gang' and negative 'non-gang' classes separately because of class imbalance in our dataset.

A. Experimental results

Table II presents the average precision, recall, and $F1$-score over the 10 folds for the single-feature and combined feature classifiers. The table includes, in braces ('{ }'), the number of gang and non-gang profiles that contain a particular feature type, and hence the number of profiles used for the 10-fold cross validation. It is reasonable to expect that *any* Twitter profile is not that of a gang member, predicting a Twitter user as a non-gang member is much easier than predicting a Twitter user as a gang member. Moreover false positive classifications of the 'gang' class may be detrimental to law enforcement investigations, which may go awry as they surveil an innocent person based on the classifier's suggestion. We thus believe that a small false positive rate of the 'gang' class to be an especially important evaluation metric. We say that a classifier is 'ideal' if it demonstrates high precision, recall, and $F1$-score for the 'gang' class while performing well on the 'non-gang' class as well.

The best performing classifier that considers single features is a Random Forest model over tweet features (T), with a reasonable $F1$-score of 0.7229 for the 'gang' class. It also features the highest $F1$-score for the 'non-gang' class (0.9671). Its strong performance is intuitive given the striking differences in language as shown in Figure 3 and discussed in Section III-B1. We also noted that music features offer promising results, with an $F1$-score of 0.6505 with a Naive Bayes classifier, as well as emoji features with an $F1$-score of 0.6067 also achieved by a Naive Bayes classifier. However, the use of profile data and image tags by themselves yield relatively poor $F1$-scores no matter which classifier considered. There may

[15] http://scikit-learn.org/stable/index.html

be two reasons for this despite the differences we observed in Section III-B. First, these two feature types did not generate a large number of specific features for learning. For example, descriptions are limited to just 160 characters per profile, leading to a limited number of unigrams (in our dataset, 10 on average) that can be used to train the classifiers. Second, the profile images were tagged by a third party Web service which is not specifically designed to identify gang hand signs, drugs and guns, which are often shared by gang members. This led to a small set of image tags in their profiles that were fairly generic, i.e., the image tags in Figure 8 such as 'people', 'man', and 'adult'.

Combining these diverse sets of features into a single classifier yields even better results. Our results for *Model(1)* show that the Random Forest achieves the highest F1-scores for both 'gang' (0.7364) and 'non-gang' (0.9690) classes *and* yields the best precision of 0.8792, which corresponds to a low false positive rate when labeling a profile as a gang member. Despite the fact that it has lower positive recall compared to the second best performing classifier (a Random Forest trained over only tweet text features (T)), for this problem setting, we should be willing to increase the chance that a gang member will go unclassified if it means reducing the chance of applying a 'gang' label to a non-gang member. When we tested *Model(2)*, a Random Forrest classifier achieved an F1-score of 0.7755 (improvement of 7.28% with respect to the best performing single feature type classifier (T)) for 'gang' class with a precision of 0.8961 (improvement of 6.26% with respect to (T)) and a recall of 0.6994 (improvement of 9.26% with respect to (T)). *Model(2)* thus outperforms *Model(1)*, and we expect its performance to improve with the availability of more training data with all feature types.

B. Evaluation Over Unseen Profiles

We also tested the trained classifiers using a set of Twitter profiles from a separate data collection process that may emulate the classifier's operation in a real-time setting. For this experiment, we captured real-time tweets from Los Angeles, CA[16] and from ten South Side, Chicago neighborhoods that are known for gang-related activities [11] using the Twitter streaming API. We consider these areas with known gang presence on social media to ensure that some positive profiles would appear in our test set. We ultimately collected 24,162 Twitter profiles: 15,662 from Los Angeles, and 8,500 from Chicago. We populated data for each profile by using the 3,200 most recent tweets (the maximum that can be collected from Twitter's API) for each profile. Since the 24,162 profiles are far too many to label manually, we qualitatively study those profiles the classifier placed into the 'gang' class.

We used the training dataset to train our best performing random forest classifier (which use all feature types) and tested it on the test dataset. We then analyzed the Twitter profiles that our classifier labeled as belonging to the 'gang' class. Each of those profiles had several features which overlap with gang members such as displaying hand signs and weapons in their profile images or in videos posted by them, gang names or gang-related hashtags in their profile descriptions, frequent use of curse words, and the use of terms such as *"my homie"* to refer to self-identified gang members. Representative tweets extracted from those

[16] http://isithackday.com/geoplanet-explorer/index.php?woeid=2442047

profiles are depicted in Figure 9. The most frequent words found in tweets from those profiles were *shit, nigga, got, bitch, go, fuck etc.* and their user profiles had terms such as *free, artist, shit, fuck, freedagang, and ripthefallen*. They had frequently used emojis such as *face with tears of joy, hundred points symbol, fire, skull, money bag, and pistol*. For some profiles, it was less obvious that the classifier correctly identified a gang member. Such profiles used the same emojis and curse words commonly found in gang members profiles, but their profile picture and tweet content was not indicative of a gang affiliation. In conclusion, we find that in a real-time-like setting, the classifier to be able to extract profiles with features that strongly suggest gang affiliation. Of course, these profiles demand further investigation and extensive evidence from other sources in order to draw a concrete conclusion, especially in the context of a law enforcement investigation. We refrain from reporting any profile names or specific details about the profiles labeled as a 'gang' member to comply with the applicable IRB governing this human subject research.

V. Conclusion and Future Work

This paper presented an approach to address the problem of automatically identifying gang member profiles on Twitter. Despite the challenges in developing such automated systems, mainly due to difficulties in finding online gang member profiles for developing training datasets, we proposed an approach that uses features extracted from textual descriptions, emojis, images and videos shared on Twitter (textual features extracted from images, and videos). Exploratory analysis of these types of features revealed interesting, and sometimes striking differences in the ways gang and non-gang members use Twitter. Classifiers trained over features that highlight these differences, were evaluated under 10-fold cross validation. Our best classifier achieved a promising F1-score of 0.7755 over the 'gang' profiles when all types of features were considered.

Future work will strengthen our training dataset by including more gang member Twitter profiles by searching for more location-independent keywords. We also plan to develop our own image classification system specifically designed to classify images found on gang member profiles. We would also like to experiment with building dictionaries that contain gang names to understand whether *"having a gang name in the profile description"* as a feature can improve our results. Finally, we would also like to study how can we further improve our classifier models using word embeddings [24] and social networks of known gang members.

Acknowledgments

We are thankful to Uday Kiran Yeda for helping us with data collection. We acknowledge partial support from the National Science Foundation (NSF) award: CNS-1513721: "Context-Aware Harassment Detection on Social Media", National Institutes of Health (NIH) award: MH105384-01A1: "Modeling Social Behavior for Healthcare Utilization in Depression" and Grant No. 2014-PS-PSN-00006 awarded by the Bureau of Justice Assistance. The Bureau of Justice Assistance is a component of the U.S. Department of Justice's Office of Justice Programs, which also includes the Bureau of Justice Statistics, the National Institute of Justice, the Office of Juvenile Justice and Delinquency Prevention, the Office for Victims of Crime, and the SMART Office. Points of view or opinions in this document are those of the authors and do not necessarily represent the official position or policies of the U.S. Department of Justice, NSF or NIH.

References

1. 2011 National Gang Threat Assessment Issued Emerging Trends. 2011.
2. Miller, WB. Crime by youth gangs and groups in the United States. US Department of Justice, Office of Justice Programs, Office of Juvenile Justice and Delinquency Prevention; Washington, DC: 1992.
3. National Gang Report. National Gang Intelligence Center; 2013.
4. Gang homicides-five us cities, 2003–2008. Morbidity and mortality weekly report. 2012; 61(3):46–51. [PubMed: 22278158]
5. Survey of Gang Members Online Habits and Participation. Survey results reported at the i-SAFE Annual Internet Safety Education Review Meeting Carlsbad; California. National Assessment Center; 2007.
6. King JE, Walpole CE, Lamon K. Surf and turf wars online-growing implications of internet gang violence. Journal of Adolescent Health. 2007; 41(6, Supplement):S66–S68. [PubMed: 18047950]
7. Pyrooz DC, Decker SH, RKM. Criminal and routine activities in online settings: Gangs, offenders, and the internet. Justice Quarterly. 2015; 32(3):471–499.
8. Patton DU, Eschmann RD, Butler DA. Internet banging: New trends in social media, gang violence, masculinity and hip hop. Computers in Human Behavior. 2013; 29(5):A54– A59.
9. Howell JC. Gang prevention: An overview of research and programs. juvenile justice bulletin. Office of Juvenile Justice and Delinquency Prevention. 2010
10. Ito M, Baumer S, Bittanti M, Cody R, Herr-Stephenson B, Horst HA, Lange PG, Mahendran D, Martínez KZ, Pascoe C, et al. Hanging out, messing around, and geeking out. Digital media. 2010
11. Wijeratne, S., Doran, D., Sheth, A., Dustin, JL. Analyzing the social media footprint of street gangs. IEEE International Conference on Intelligence and Security Informatics (ISI); May 2015; 2015. p. 91-96.
12. Forum, PER. Tech Rep. United States Office of Community Oriented Policing Services and United States Department of Justice; 2013. Social media and tactical considerations for law enforcement.
13. Brunty, J., Miller, L., Helenek, K. Social media investigation for law enforcement. Routledge; 2014.
14. Pennacchiotti M, Popescu A-M. A machine learning approach to twitter user classification. 2011
15. Tinati, R., Carr, L., Hall, W., Bentwood, J. Identifying communicator roles in twitter. Proceedings of the 21st International Conference on World Wide Web, ser. WWW '12 Companion; New York, NY, USA: ACM; 2012. p. 1161-1168.
16. Liu W, Ruths D. Whats in a name? using first names as features for gender inference in twitter. 2013
17. Purohit, H., Dow, A., Alonso, O., Duan, L., Haas, K. User taglines: Alternative presentations of expertise and interest in social media. 2012 International Conference on Social Informatics (Social Informat-ics); Washington, D.C., USA. December 14–16; 2012. p. 236-243.
18. Thrasher, FM. The gang: A study of 1,313 gangs in Chicago. University of Chicago Press; 1963.
19. Patton DU, Hong JS, Ranney M, Patel S, Kelley C, Eschmann R, Washington T. Social media as a vector for youth violence: A review of the literature. Computers in Human Behavior. 2014
20. Decker S, Pyrooz D. Leaving the gang: Logging off and moving on. council on foreign relations. 2011
21. Patton, DU. Gang violence, crime, and substance use on twitter: A snapshot of gang communications in detroit. Society for Social Work and Research 19th Annual Conference: The Social and Behavioral Importance of Increased Longevity; jan 2015;
22. Wang, W., Chen, L., Thirunarayan, K., Sheth, AP. Cursing in english on twitter. Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing, ser. CSCW '14; New York, NY, USA: ACM; 2014. p. 415-425.
23. Patton DU, Lane J, Leonard P, Macbeth J, Smith-Lee JR. Gang violence on the digital street: Case study of a south side chicago gang members twitter communication. New Media & Society. 2016

24. Wijeratne, S., Balasuriya, L., Doran, D., Sheth, A. Word embeddings to enhance twitter gang member profile identification. IJCAI Workshop on Semantic Machine Learning (SML 2016); New York City, NY: CEUR-WS; 2016.

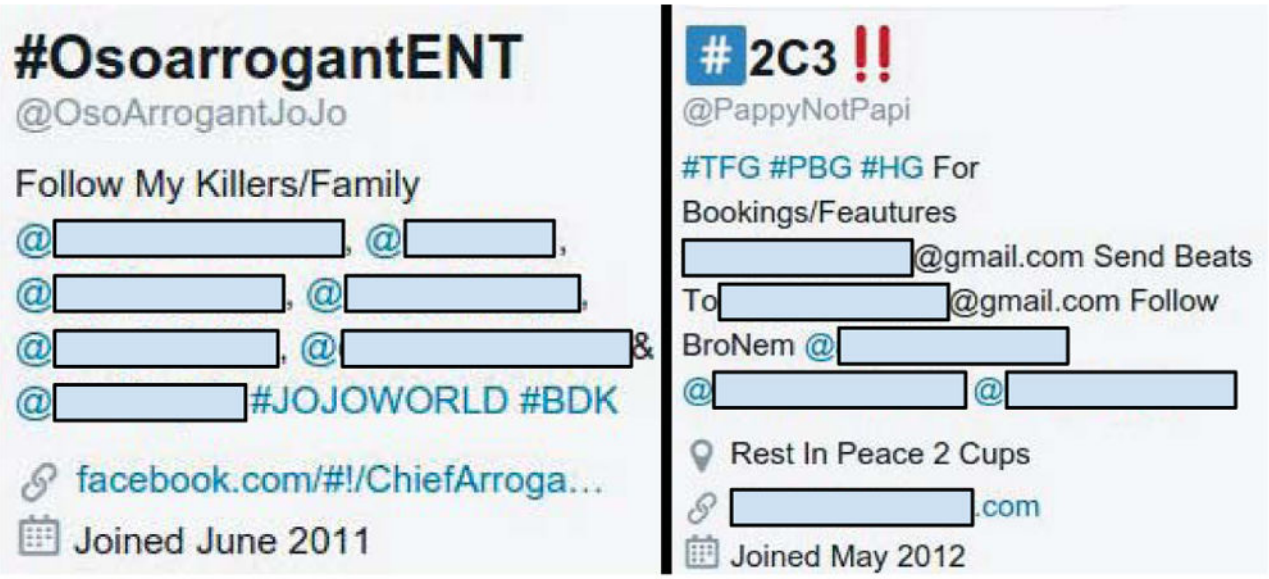

Fig. 1.

Twitter profile descriptions of known gang members. Pursuant to an IRB governing human subject research, we are prohibited from revealing personally identifiable information in this paper. We only report Twitter handles that have already been revealed in widely reported publications and were not collected by the research team for this work.

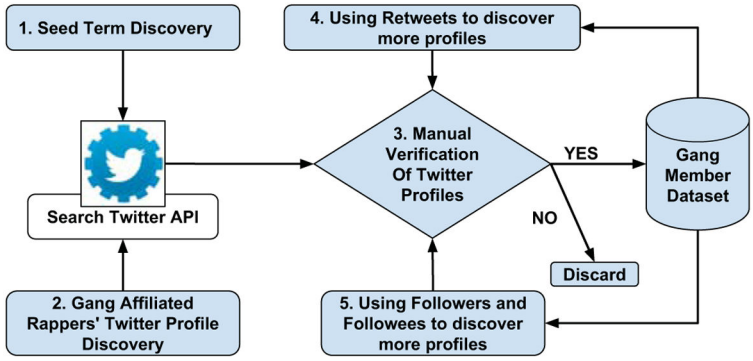

Fig. 2.
Gang member dataset creation.

(a) Gang members. (b) Non-gang members.

Fig. 3.
Comparison of words used in tweets.

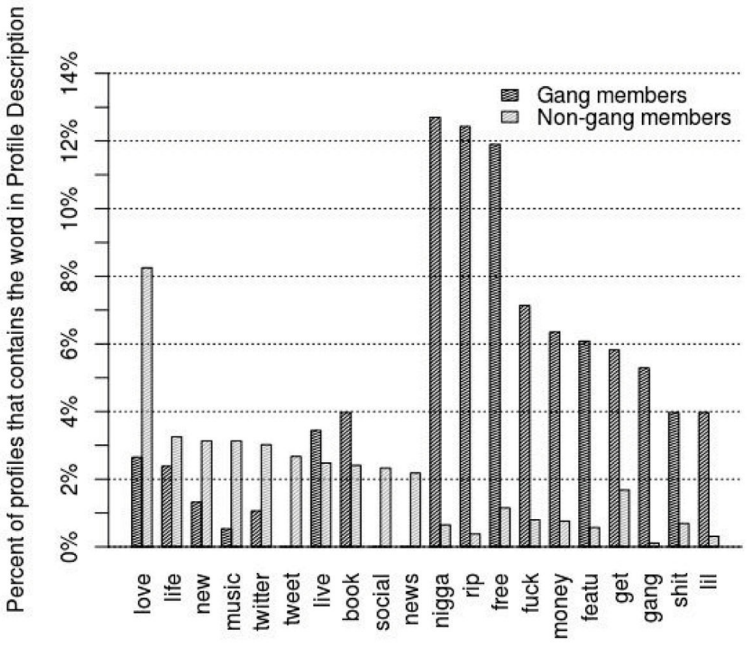

Fig. 4.
Word usage in profile descriptions: gang vs non-gang.

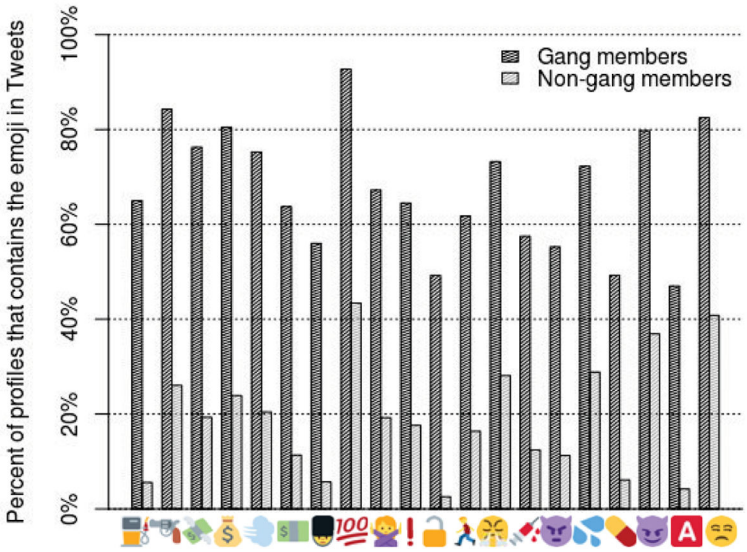

Fig. 5.
Emoji usage distribution: gang vs non-gang.

Fig. 6.
Examples for gang members' tweets with emojis.

Fig. 7.
Sample gang member profile images.

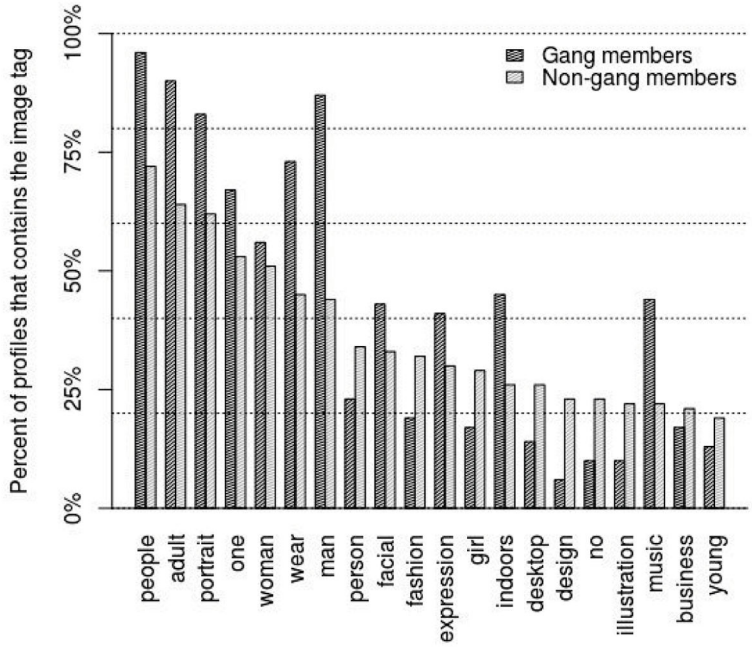

Fig. 8.
Image tags distribution: gang vs non-gang.

Fig. 9.
Sample tweets from identified gang members.

TABLE I

Number of gang member profiles captured.

Method	Number of Profiles
Seed term discovery	280
Gang Affiliated Rappers	22
Retweets, Followers & Followees	98
Total	**400**

TABLE II

Classification results based on 10-fold cross validation.

Features	Total Number of Profiles {#Gang : #Non-Gang}	Classifier	Results					
			Gang			Non-Gang		
			Precision	Recall	*F*1-score	Precision	Recall	*F*1-score
Tweets (T)	3,265 {400 : 2,865}	Naive Bayes	0.4354	**0.9558**	0.5970	**0.9929**	0.8278	0.9028
		Logistic Regression	0.6760	0.6623	0.6666	0.9529	0.9544	0.9536
		Random Forest	0.8433	0.6401	0.7229	0.9517	0.9832	0.9671
		SVM	0.6301	0.6545	0.6388	0.9514	0.9442	0.9477
Emojis (E)	3,085 {396 : 2,689}	Naive Bayes	0.4934	0.7989	0.6067	0.9676	0.8785	0.9207
		Logistic Regression	0.6867	0.3995	0.4969	0.9164	0.9733	0.9438
		Random Forest	0.7279	0.5079	0.5931	0.9292	0.9721	0.9500
		SVM	0.4527	0.5642	0.4955	0.9329	0.8953	0.9133
Profile data (P)	2,996 {378 : 2,618}	Naive Bayes	0.6000	0.243	0.464	0.8765	**1.0000**	0.9341
		Logistic Regression	0.8015	0.2160	0.3362	0.8974	0.9924	0.9424
		Random Forest	0.5719	0.1441	0.2239	0.8886	0.9859	0.9346
		SVM	0.7501	0.2225	0.3394	0.8978	0.9897	0.9414
Image tags (I)	2,910 {357 : 2,553}	Naive Bayes	0.2692	0.6973	0.3851	0.9458	0.7357	0.8271
		Logistic Regression	0.4832	0.1853	0.2624	0.8950	0.9722	0.9318
		Random Forest	0.4131	0.1512	0.2147	0.8911	0.9731	0.9300
		SVM	0.3889	0.1454	0.205	0.8898	0.9679	0.9270
Music interest (Y)	1,630 {196 : 1,434}	Naive Bayes	0.5865	0.7424	0.6505	0.9632	0.9297	0.9460
		Logistic Regression	0.7101	0.5447	0.6110	0.9395	0.9679	0.9534
		Random Forest	0.8403	0.3953	0.5277	0.9232	0.9895	0.9550
		SVM	0.6232	0.6067	0.6072	0.9463	0.9476	0.9467
Model(1) {T + E + P + I + Y}	3,265 {400 : 2,865}	Naive Bayes	0.3718	0.9387	0.5312	0.9889	0.7791	0.8715
		Logistic Regression	0.7250	0.6880	0.7038	0.9564	0.9637	0.9599
		Random Forest	0.8792	0.6374	0.7364	0.9507	0.9881	0.9690
		SVM	0.6442	0.6791	0.6583	0.9546	0.9469	0.9506
Model(2) {T + E + P + I + Y}	1,358 {172 : 1,186}	Naive Bayes	0.4405	0.9386	0.5926	0.9889	0.8254	0.8991

Proc IEEE ACM Int Conf Adv Soc Netw Anal Min. Author manuscript; available in PMC 2017 July 13.

Features	Total Number of Profiles {#Gang : #Non-Gang}	Classifier	Results					
			Gang			Non-Gang		
			Precision	Recall	*F1*-score	Precision	Recall	*F1*-score
		Logistic Regression	0.7588	0.7396	0.7433	0.9639	0.9662	0.9649
		Random Forest	**0.8961**	0.6994	**0.7755**	0.9575	0.9873	**0.9720**
		SVM	0.7185	0.7394	0.7213	0.9638	0.9586	0.9610