2017

# Finding Street Gang Member Profiles on Twitter

Lakshika Balasuriya
*Wright State University*

# FINDING STREET GANG MEMBER PROFILES ON TWITTER

A thesis submitted in partial fulfilment
of the requirements for the degree of
Master of Science

By

LAKSHIKA BALASURIYA
B.Sc., University of Moratuwa, Sri Lanka, 2009

2017
Wright State University

WRIGHT STATE UNIVERSITY
SCHOOL OF GRADUATE STUDIES

NOVEMBER 29, 2017

I HEREBY RECOMMEND THAT THE THESIS PREPARED UNDER MY SUPERVISION BY Lakshika Balasuriya ENTITLED Finding Street Gang Member Profiles on Twitter BE ACCEPTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF Master of Science.

Amit Sheth, Ph.D.
Thesis Co-Director

Derek Doran, Ph.D.
Thesis Co-Director

Mateen M. Rizki, Ph.D.
Chair, Department of Computer Science and Engineering

Committee on
Final Examination

Amit Sheth, Ph.D.

Derek Doran, Ph.D.

Krishnaprasad Thirunarayan, Ph.D.

Barry Milligan, Ph.D.
Interim Dean of the Graduate School

# Abstract

Balasuriya, Lakshika. M.S., Department of Computer Science and Engineering, Wright State University, 2017. Finding Street Gang Member Profiles on Twitter

The crime and violence street gangs introduce into neighborhoods is a growing epidemic in cities around the world. Today, over 1.4 million people, belonging to more than 33,000 gangs, are active in the United States, of which 88% identify themselves as being members of a street gang. With the recent popularity of social media, street gang members have established online presences coinciding with their physical occupation of neighborhoods. Recent studies report that approximately 45% of gang members participate in online offending activities such as threatening, harassing individuals, posting violent videos or attacking someone on the street for something they said online in social media platforms. Thus, their social media posts may be useful to social workers and law enforcement agencies to discover clues about recent crimes or to anticipate ones that may occur in a community. Finding these posts, however, requires a method to discover gang member social media profiles. This is a challenging task since gang members represent a very small population compared to the active social media user base.

This thesis studies the problem of automatically identifying street gang member profiles on Twitter, which is a popular social media platform that is commonly used by street gang members to promote their online gang-related activities. It outlines a process to curate one of the largest sets of verifiable gang member Twitter profiles that have ever been studied. A review of these profiles establishes differences in the language, profile and cover images, YouTube links, and emoji shared on Twitter by gang members compared to the rest of the Twitter population. Beyond the earlier efforts in Twitter profile identification that utilize features derived from the profile and tweet text, this thesis uses additional heterogeneous sets of features from the emoji usage, profile images, and links to YouTube videos reflecting gang-related music culture towards solving the gang member profile identification problem. Features from this review are used to train a series of supervised machine learning classifiers and they are further improved upon by using word embeddings learned over a large corpus of tweets. Experimental results demonstrate that heterogeneous features enabled our classifiers to achieve low false positive rates and promising $F1$-scores.

# Contents

# List of Figures

# List of Tables

# Acknowledgment

This journey has been filled with ups and downs. I would like to take this opportunity to express my gratitude to all those who have helped me throughout this journey.

First and Foremost, I am grateful for my advisors Dr. Amit P. Sheth and Dr. Derek Doran for encouraging me and guiding me. I am thankful for Dr. Sheth for all the opportunities he provided me and for helping me out when I was in need. He supported me, gave me valuable inputs and also let me work on the problems that interest me. I am thankful for Dr. Derek Doran for his guidance and his valuable inputs and his expertise which better shaped my work. I learned a lot from him and I am glad that I had the opportunity to work with him.

I would like to thank Dr. Krishnaprasad Thirunarayan for encouraging me, his thoughtfulness and also for his feedback on my thesis work. I am thankful for Sanjaya Wijeratne, for being my collaborator, mentor and also my loving husband. I am grateful for him for being there with me through thick and thin. He has motivated, guided and helped me throughout this journey. I would also like to thank Dr. Guozhu Dong for his time and guidance and the Department of Computer Science and Engineering at Wright State University for funding my education via a Teaching Assistantship in the past.

I am grateful for my beloved parents and siblings for their unconditional love and support. My precious nieces, Yarra and Chanidi for inspiring me to be better everyday. I am thankful for my loving in-laws, past teachers and friends for encouraging me. I would also like to thank my Sri Lankan friends and their families in Dayton for their friendship and for being my home away from home. Especially, I would like to thank Ajith Ranabahu and Dharshani Nadeeka Herath who have been very kind and caring through the years. I am thankful for Sujan Perera, Sarasi Lalithsena and Kalpa Gunaratna for helping me in various ways.

Last but not least, I would like to thank my past and present colleagues at Kno.e.sis, including the non-academic staff members. I appreciate all the help I have received from them. Everyone of them have helped me one way or the other. Some of my colleagues reviewed our papers, helped us with evaluations and was always there whenever I needed help.

Finally, I would like to acknowledge funding sources that enabled the research in this thesis. This research received 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.

Dedicated to

*my loving parents and husband*

# 1 Introduction

The crime and violence street gangs introduce into neighborhoods is a growing epidemic in cities around the world[1]. Today, over 1.4 million people in the United States are members of a *street gang* [2, 3], which is "*a self-formed association of peers, united by mutual interests, with identifiable leadership and internal organization, who act collectively or as individuals to achieve specific purposes, including the conduct of illegal activity and control of a territory, facility, or enterprise*" [4]. 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 the victims of at least 1.3% of all gang-related[2] homicides are merely innocent bystanders who live in gang occupied neighborhoods [5].

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 [2]. 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 [6]. More recent studies report approximately 45% of gang members participate in online offending activities such as threatening, harassing individuals, posting violent videos or attacking someone on the street for something they said

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

[2] The terms 'gang' and 'street gang' are used interchangeably in this thesis.

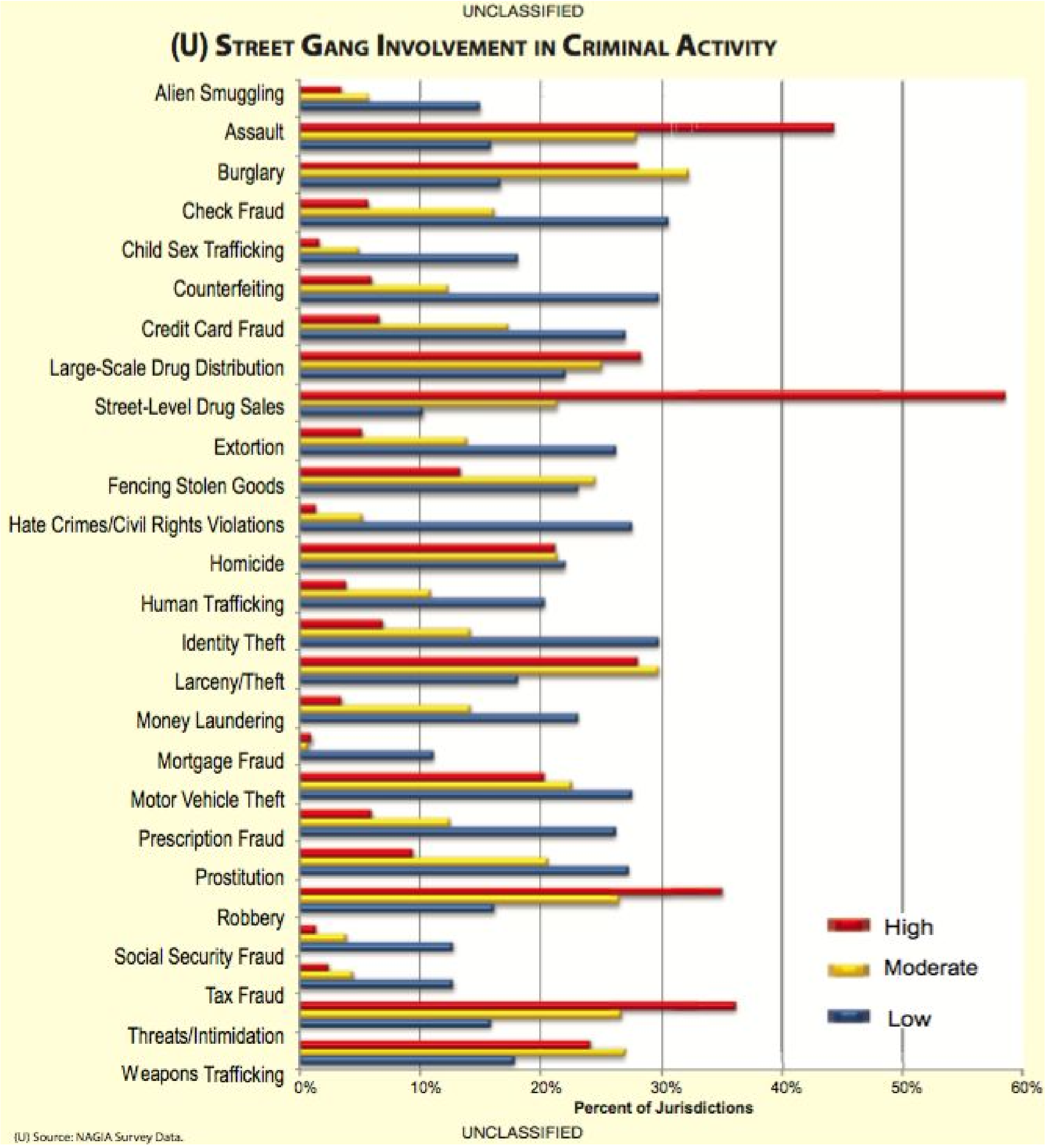
online [7, 8]. This "Cyber-" or "Internet banging" [9] behavior is precipitated by the fact that an increasing number of young members of the society are joining gangs [10], and these young members have become enamored with technology and with the notion of sharing information quickly and publicly through 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 [11].

Past research has shown that social media play an essential role in illicit activities carried out by street gang members [12, 13]. For example, street gang members use social media as a platform to threaten their rival gangs, sell drugs, publicize crimes to gain online reputation and to recruit new gang members [7, 14, 12, 13]. Figure 1.1 depicts a complete list of illicit activities carried out by street gangs as per the 2015 National Gang Report [1]. It further reports that the social media use of street gang members is on the rise. For example, it reports that over 90% of street gang members have used Facebook at least once in 2015 (See Figure 1.2). Among other popular social media websites, YouTube, Instagram and Twitter have also received attention of the gang members. For example, close to 80% of street gang members have used YouTube in 2015 where as Instagram and Twitter have been used by more than 60% of them. Gang members publicly share their activities on these social media websites. However, sites such as Facebook[4] and Instagram[5] do not allow the use of user-generated data for further aggregated analysis without the user's consent, even if the data is publicly available. On the other hand, publicly available data on Twitter can be used for aggregated analysis as long as personally identifiable information related to a user is not revealed in the analysis.

[3] http://www.hhs.gov/ash/oah/news/e-updates/eupdate-nov-2013.html

[4] https://www.facebook.com/legal/FB_Work_Privacy

[5] https://www.instagram.com/about/legal/terms/api/



(U) Source: NAGIA Survey Data.

Figure 1.1: Street gang involvement in various criminal activities in the USA. Image extracted from the 2015 National Gang Report [1].

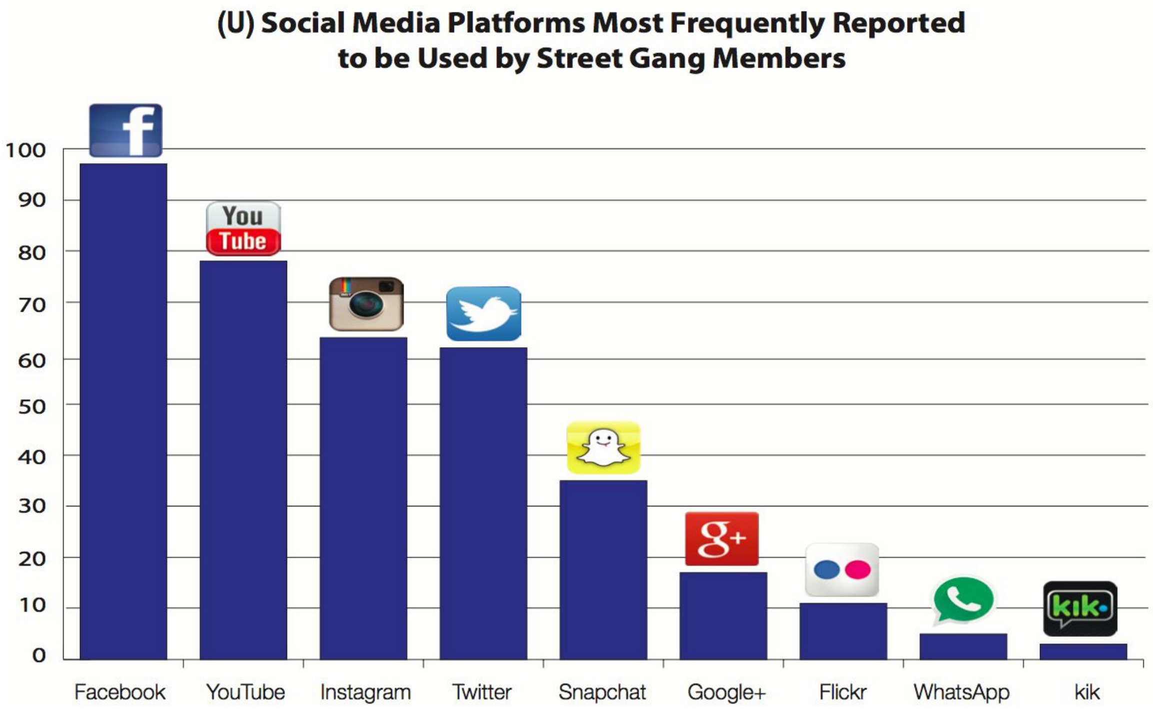


Figure 1.2: Social media use by street gang members in the USA. Image extracted from the 2015 National Gang Report [1].

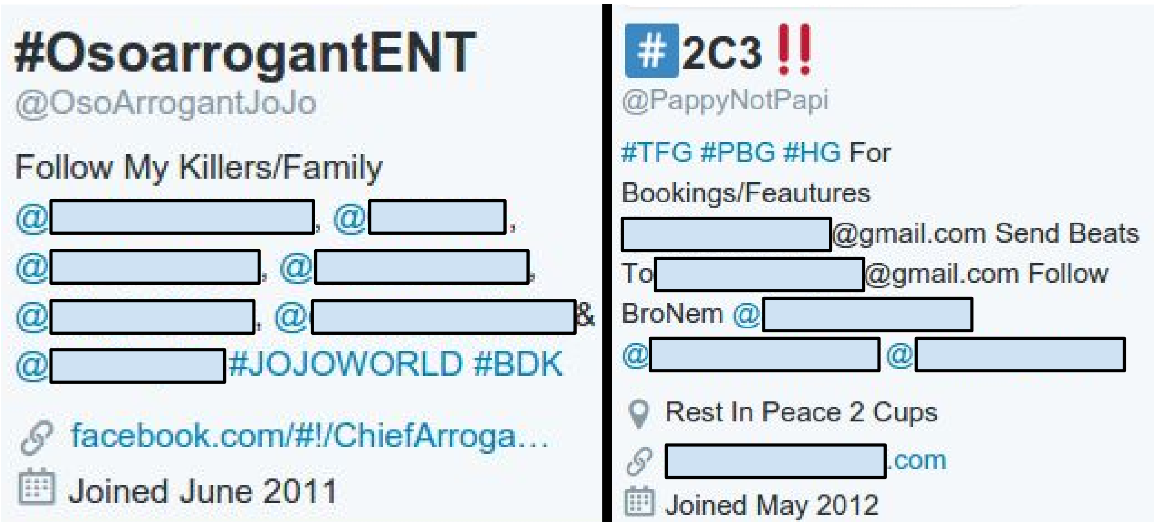
Gang members are able to post publicly on Twitter without fear of consequences because there are few tools law enforcement can use to surveil this medium [15]. Their posts provides live updates on gang activity and can be leveraged by law enforcement and social workers to identify problem areas and send workers in to conflict mediation [16]. 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 [17]. Officer training is broadly limited to understanding policies on using Twitter in investigations and best practices for data storage [18]. From offline clues, the officers monitor just a selected set of social media accounts which are manually discovered and related to a specific investigation. Thus, developing tools to identify gang member profiles on social media is an important step in the direction of using machine intelligence to fight crime. The safety and security of city neighborhoods can thus be improved if law enforcement was equipped with intelligent tools to study social media for gang activity.

The need for better tools for law enforcement and social workers 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 the Englewood,Chicago's teenage rapper Joseph Coleman also known as Lil Jojo 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[6]. In Coleman's subsequent tweets, he posted "im on 069" and revealed his location, and minutes later, was shot dead on the 6900 block of South Princeton Avenue in the Englewood neighborhood of Chicago. Subsequent investigation revealed that the rivalry leading to his death began and was carried out entirely on social media. [19] have studied Twitter communication of one known female gang member in Chicago, Gakirah Barnes, during a two week window in which her friend was killed and then weeks later, she was also killed. They observed how the street culture is reflected in gang related tweets and also found that scripts of reciprocal violence within a local network have real world consequences that resemble street gang behavior [19, 16]. Other reporting has revealed how innocent bystanders have also become targets in online fights, leaving everyone in the neighborhood at risk[7].

This thesis 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

[6] http://www.wired.com/2013/09/gangs-of-social-media/

[7] https://goo.gl/75U3ME



#OsoarrogantENT
@OsoArrogantJoJo

Follow My Killers/Family
@[redacted], @[redacted],
@[redacted], @[redacted],
@[redacted], @[redacted]&
@[redacted]#JOJOWORLD #BDK

facebook.com/#!/ChiefArroga…
Joined June 2011

#️⃣2C3‼️
@PappyNotPapi

#TFG #PBG #HG For
Bookings/Feautures
[redacted]@gmail.com Send Beats
To[redacted]@gmail.com Follow
BroNem @[redacted]
@[redacted] @[redacted]

Rest In Peace 2 Cups
[redacted].com
Joined May 2012

Figure 1.3: 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 thesis. 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.

in other contexts [20, 21, 22], but gang member profiles pose unique challenges. For example, many Twitter profile classifiers search for contextual clues in tweets and profile descriptions [23], 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 1.3, 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 thesis 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 evaluation under various learning algorithms demonstrated a low false positive rate and a promising $F1$-score of 0.7755 for using these sets of features.

Motivated by the recent success of word embeddings-based methods to learn syntactic and semantic structures automatically when provided with large datasets, we then investigate the use of word embeddings to further improve our classifiers. Specifically, we train a Skip-gram model using a large Twitter corpus and generate word embeddings that translate the features into a real vector format amenable for machine learning classification and use them to train another set of supervised classifiers. We show that pre-trained word embeddings improve the machine learning models we developed earlier and help us obtain an $F1$-score of 0.7835 on identifying gang member profiles (a 6.39% improvement in $F1$-score compared to the baseline models which were not trained using word embeddings).

## 1.1 Thesis Organization

The remainder of this thesis is organized as follows. Chapter 2 discusses the related literature. Specifically, it discusses past research related to gang members activity in social media and word embedding techniques and positions how the work presented in this thesis differs from the related work discussed. Chapter 3 discusses the techniques used and steps followed to collect the gang and non-gang member Twitter profiles dataset in detail. Chapter 4 reports a review of different features available in the dataset, highlighting the predictive power of each feature. Chapter 5 discusses the different approaches used to conduct the experiments while Chapter 6 gives a detailed explanation of the evaluation of the proposed method and the results obtained. Chapter 7 concludes the work reported while discussing the potential future work.

## 1.2 Publication of Thesis Work

The work presented in this thesis has been published in the following conferences and workshops.

1. ASONAM 2016 – The creation of the gang member Twitter profile dataset along with building classification models to automatically identify such profiles has been published as a full paper at the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2016). Full citation of the publication is given below.

   L. Balasuriya, S. Wijeratne, D. Doran, and A. Sheth, "Finding Street Gang Members on Twitter," in 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), vol. 8, San Francisco, CA, USA, August 2016, pp. 685–692.

2. SML 2016 – The work conducted on using word embedding models to improve gang member profile identification, including building new classification models has been published as a full workshop paper at the 3rd International Workshop on Semantic Machine Learning (SML 2016), co-located with the 25th International Joint Conference on Artificial Intelligence (IJCAI-16). Full citation of the publication is given below.

   S. Wijeratne, L. Balasuriya, D. Doran, and A. Sheth, "Word Embeddings to Enhance Twitter Gang Member Profile Identification," in IJCAI Workshop on Semantic Machine Learning (SML 2016). New York City, NY: CEUR-WS, 07/2016 2016.

3. ChASM 2016 – The experiments conducted on features that can lead to identification of street gang member Twitter profiles has been published as

an extended abstract at the 4th Computational Approaches to Social Modeling Workshop (ChASM 2016), co-located with the 8th International Conference on Social Informatics (SocInfo 2016). Full citation of the publication is given below.

L. Balasuriya, S. Wijeratne, D. Doran, and A. Sheth, "Signals Revealing Street Gang Members on Twitter," in Workshop on Computational Approaches to Social Modeling (ChASM 2016) co-located with 8th International Conference on Social Informatics (SocInfo 2016), vol. 4, Bellevue, WA, USA, November 2016.

# 2 Related Work

This chapter discusses research studies that are related to the work presented in this thesis. We first discuss the research related to the social media usage among street gang members. In particular, we discuss several studies that built applications to understand the activities of street members on Twitter. Then we discuss a selected set of studies that aimed at building Twitter profile classification models and emphasize how our approach differs from the other approaches presented prior to our work. Finally, we briefly discuss research on word embedding models and how they can be used for text classification tasks. We also highlight how our work differs from the existing approaches that use word embeddings for text classification.

## 2.1 Social Media Use of Street Gang Members

Gang violence is a well studied social science topic dating back to 1927 [24] while the existence of criminal gangs in America dates back to 1760 [25]. Historical reviews portray American gangs emerging along racial and ethnic lines and developing into organizations designed for illegal business including drug and weapon trafficking, prostitution, human trafficking etc. [25]. 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"* [9], was only recently introduced [26, 15].

Patton *et al.* [26] were the first to introduce the concept of "Internet banging" and they studied how social media is being used as a tool for gang self-promotion and gaining and maintaining street credibility [9]. 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 [7, 14]. Past research also reveals that gang members use social platforms to recruit new members to their gangs. For example, a recent study by Pyrooz *et al.* [12] that interviewed 418 current and former gang members reports that 8% of the participants had stated that their gangs recruited new individuals online. Morselli *et al.* [13] reported that gang members use the Internet and social networking sites as much, if not more, than their non-gang counterparts and gang members have a greater overall propensity for online crime and deviance than former and non-gang respondents.

The ability to take action upon gang members' activity on social media is limited by the tools available to discover gang members on social media sites and to analyze the content they post [26]. Recent attempts to improve the abilities of analyzing social media posts by gang members 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 [15]. The proposed surveillance system, which extends the Twitris social media platform [27], had four design goals aimed at understanding gang member posts, namely, (i) monitor negative community effects of gang activities, (ii) discover opinion leaders who influence the thoughts and actions of other gang members, (iii) evaluate the sentiment of posts targeting communities, locations, and groups (including rival gangs), and (iv) monitor community and gang responses to community support programs. The designers of the surveillance system argued that it should be able to analyze the spatio-temporal-thematic (where,when, and what), people-content-networking (who and how), and emotion-sentiment (perceptions and intent) dimensions of social media posts in order to support the proposed design goals. However, the said architecture requires a set of gang member profiles for input, thus assuming that they have already been discovered.

Patton *et al.* [14] devised a method to automatically collect tweets from a group of gang members operating in Detroit, MI. They manually identified Twitter profiles belong to known street gang members who operate in the Detroit area and then used keywords related to crime and violence to further filter out tweets posted by them. Similar to Wijeratne *et al.* [15], this approach requires the Twitter profile names of the gang members to be known beforehand, and data collection was localized to a single city in the country. In another study that examined how gang members use social media, Decary-Hetu *et al.* [28] used 28 keywords which are related to U.S. and Canada-based gangs to collect tweets and Facebook posts that discussed their gang-related activities. They reported that there has been an increase in social media use by gang members and the amount of information being shared online on gang activities. Decary-Hetu *et al.*'s [28] data collection approach was also based on pre-identified gang names that are specific to a set of to U.S. and Canadian cities where a large number of gangs operate. Radil *et al.* [29] studied the rivalry network of Los Angeles-based street gangs using social network analysis techniques. Piergallini *et al.* [30] studied the graffiti style features used by street gang members in online Web forums to develop methods to distinguish their gang affiliation. Radil *et al.* [29]'s approach is only limited to street gangs operating in Los Angeles while Piergallini *et al.* [30]'s is limited to twelve gangs which include Bloods, Crips, Hoovers, Gangster Disciples, other Folk Nation, Latin Kings, Vice Lords, Black P. Stones, other People Nation, Trinitarios, Norteños, and Sureños.

The work reported in this thesis differs from the related research discussed above in two ways. Firstly, we propose a location-agnostic method to collect Twitter profiles of the street gang members. Instead of using gang names as keywords to search for Twitter profiles or manually identifying street gang members' Twitter handles, we use hashtags that are commonly used by street gang members across U.S. to (i) support their fellow members who are in jail (e.g., `#FreeDaGuys`), (ii) convey the grieving for fallen gang members (e.g., `#RIPDaGuys`), and (iii) show their hatred towards police officers (e.g., `#FuckDaOpps`). By doing so, we were able to create a gang members dataset that is not specific to a particular city or neighborhood. Secondly, this thesis uses additional heterogeneous sets of features than to what is proposed in earlier studies in the classification models. For example, we extract features from the emoji usage, profile images, and links to YouTube videos reflecting gang-related music culture in addition to features extracted from tweets and Twitter profile descriptions. Experimental results demonstrate that heterogeneous features enabled our classifiers to achieve low false positive rates and promising $F1$-scores.

## 2.2 Twitter User Profile Classification

Twitter user profile classification is a well-studied problem where a class label is assigned to a Twitter profile from a set of pre-defined labels. Concrete examples of Twitter profile classification include user political affiliation classification [20], ethnicity classification [20], gender identification [22], brand loyalty prediction [20], and user occupation classification [23]. Majority of these applications rely only on textual features extracted from content posted on Twitter or user profiles. Pennacchiotti *et al.* [20] proposed a machine learning framework to classify Twitter profiles by using the Twitter user profile description, user's tweeting behavior, linguistic content of tweets and user's follower/followee network as features. Pennacchiotti *et al.* showed that their framework can be used to identify user attributes such as a user's ethnicity, political affiliation or brand loyalty. Liu *et* *al.* [22] tried to incorporate user's self-reported first name into a gender classifier and showed that, when combined with other textual features obtained from tweets, first name can improve the gender classification of Twitter users. Purohit *et al.* [23] developed a method to generate user summaries or 'User Tag Lines' for Twitter users based on the content posted on their Twitter profiles. They utilized Twitter profile description-based features along with features extracted from tweets (e.g., entities present in tweets and word phrases) to generate user summaries, which could then be used in a user profile classification task.

The work reported in this thesis builds upon the existing methods to automatically classify Twitter profiles. Unlike the above approaches that utilize an abundance of positive examples in their training data, and only rely on one or two feature types[8](typically, tweet text and profile description), we consider the use of a variety of feature types, including emoji, YouTube links, and image features. We show that integrating multiple types of features could significantly improve the classification accuracy of gang member Twitter profile classification problem.

## 2.3 Word Embedding Models

In addition to using a diverse set of feature types, this thesis also explores the possibility of further improving Twitter profile classification results by mapping the above identified features types into a considerably smaller feature space through the use of word embeddings. A word embedding model is a neural network that learns rich representations of words in a text corpus. It takes data from a large, $n$-dimensional 'word space' (where $n$ is the number of unique words in a corpus) and learns a transformation of the data into a lower $k$-dimensional space of real-valued numbers. This transformation is developed in a way that similarities between the $k$-dimensional

[8] The terms 'feature type' and 'content type' are used interchangeably in this thesis. These terms refer to the different types of content used for feature extraction.

vector representation of two words reflects semantic relationships among the words themselves. These semantics are not captured by typical bag-of-words or $n$-gram models for classification tasks on text data [31, 32].

Word embeddings have led to state-of-the-art results in many natural language processing tasks [33]. In fact, word embedding learning is an important step for many statistical language modeling tasks in text processing systems. Bengio *et al.* were the first ones to introduce the idea of learning a distributed representation for words over a text corpus [34]. They learned representations for each word in the word corpus using a neural network model that modeled the joint probability function of word sequences in terms of the feature vectors of the words in the sequence. Mikolov *et al.* showed that word embeddings learned over a text corpus can be used to perform simple algebraic operations on them, which leads to findings such as word embedding vector of the word “King” $-$ the word embedding vectors of “Man” $+$ “Woman” would results in a word embedding vector that is closest to the word embedding vector of the word “Queen” [31]. Recent successes in using word embeddings to improve text classification for short text [35, 36], encouraged us to explore how they can be used to improve gang and non-gang member Twitter profile classification.

Word embeddings can be performed under different neural network architectures; two popular ones are the Continuous Bag-of-Words (CBOW) and Continuous Skip-gram (Skip-gram) models [37]. The CBOW model learns a neural network such that given a set of context words surrounding a target word, it predict a target word. The Skip-gram model differs by predicting context words given a target word and by capturing the ordering of word occurrences. Recent improvements to Skip-gram model make it better able to handle less frequent words, especially when negative sampling is used [32].

Previous research has shown word embedding-based methods can improve classification of short text [35, 36]. Thus, we investigate using word embeddings to further improve the process of identifying gang member profiles on Twitter. We believe our corpus of gang and non-gang member tweets, with nearly 64.6 million word tokens, could act as a rich resource to train word embeddings for distinguishing gang and non-gang member Twitter users. Our method differs from other word embedding-based text classification systems such as [35, 36] due to the fact that we use a set of heterogeneous features including emojis in tweets and image tags extracted from profile and cover images available in Twitter in our classification task [38]. Experimental results demonstrate that heterogeneous features enabled our classifiers to achieve low false positive rates and promising $F1$-scores.

# 3 Data Curation

This section discusses the methodology we followed to create the gang and non-gang member datasets we used in our study. It includes a semi-automatic data collection process to discover one of the largest sets of verifiable gang member Twitter profiles that have ever been studied.

## 3.1 Gang Member 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 [15, 14]. For example, Wijeratne *et al.* [15] studied Chicago-based street gangs based on a Twitter profile dataset collected using local street gang names. Patton *et al.* [14] studied Detroit-based street gangs by manually identifying the gang members' Twitter profiles. 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 [15]. 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 to build a dataset of gang member profiles suitable for training a classifier. The steps of the workflow are: (i) seed term discovery, (ii) gang affiliated rappers' Twitter profile discovery, (iii) manual verification of Twitter profiles, (iv) using retweets to discover gang member Twitter profiles, and (v) using followers and followees to discover gang member Twitter profiles. The workflow is illustrated in Figure 3.1 and each step of the workflow is discussed in detail below.

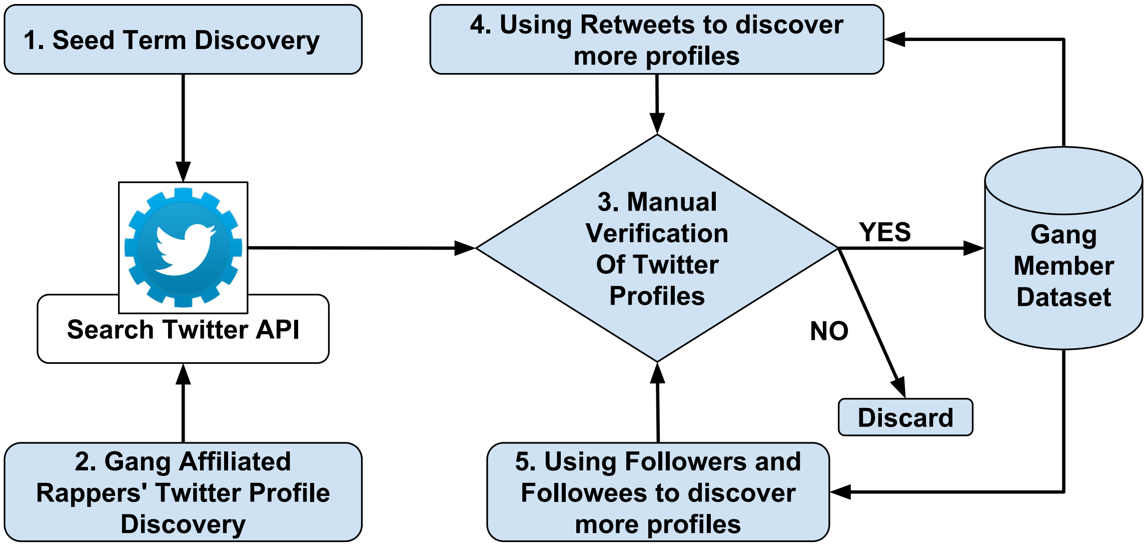

1. Seed Term Discovery
4. Using Retweets to discover more profiles
3. Manual Verification Of Twitter Profiles
YES
NO
Gang Member Dataset
Search Twitter API
Discard
2. Gang Affiliated Rappers' Twitter Profile Discovery
5. Using Followers and Followees to discover more profiles

Figure 3.1: Gang member dataset creation.

### 3.1.1 Seed Term Discovery

Following the success of identifying gang member profiles from Chicago [15], 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 [15], namely #BDK (Black Disciple Killers) and #GDK (Gangster Disciples Killers). Those profiles were analyzed and manually verified as explained in subsection 3.1.3.

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[9] and Twitter REST API[10] 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.

### 3.1.2 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.* [9] 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[11]. We searched for these rapper profiles on Twitter and manually checked that the rapper was affiliated to a gang.

### 3.1.3 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 the user. During these checks, we searched for terms, activities, and symbols that we believed could be associated with a gang including self-identification of gang affiliation in their

[9] https://moz.com/followerwonk/bio

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

[11] http://www.hipwiki.com/List+of+Rappers+Murdered+in+2015

Twitter profiles. 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.

### 3.1.4 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 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.

### 3.1.5 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[12], we hypothesized that the followers and followees of Twitter profiles from the seed set may also be gang members. However, 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 3.1 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 [15]. 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.

[12] http://aris.ss.uci.edu/~lin/52.pdf

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

Table 3.1: Number of gang member profiles captured.

## 3.2 Non-Gang Member Data collection

For this study, profiles of non-gang members were collected from the Twitter Streaming API[13]. We first collected a random sample of tweets and retrieved 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[14]. 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 3.1.3. Introducing these profiles, which have some characteristics of gang members (such as cursing frequently or cursing at law enforcement) but are not, captures local languages used by family/friends of gang members and ordinary people in a neighborhood where gangs operate.

[13] `https://dev.twitter.com/streaming/overview`

[14] `https://developers.google.com/maps/`

## 3.3 Dataset

Using the Twitter REST API[15], 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.

[15] `https://dev.twitter.com/rest/public`

# 4 Data Analysis For Feature Extraction

Feature engineering is an important part of any study that uses supervised-machine learning. Specifically, studies have shown that carefully identified features can improve the performance of Twitter-based supervised learning tasks [39, 40]. Thus, we next explore the differences between gang and non-gang members' Twitter usage patterns to find promising features for classifying their Twitter profiles. Based on previous studies and our observations during the manual verification of gang member profiles, we explored 5 different feature types that are listed below to see whether they can be used to discriminate gang member profiles in Twitter. They are:

1. Tweet Text – This includes the textual content present in a tweet. We extract unigrams from the tweet text and treat each unigram as a feature.

2. Twitter Profile Description – This includes user-provided description of a Twitter profile. We extract unigrams from the text appear in the Twitter profile description and treat each unigram as a feature.

3. Music Interests – We process each YouTube video shared along with tweets and extract unigram features from the video title, description and comments posted on the YouTube video.

4. Emoji – We extract emoji from tweet text and treat each emoji as a feature.

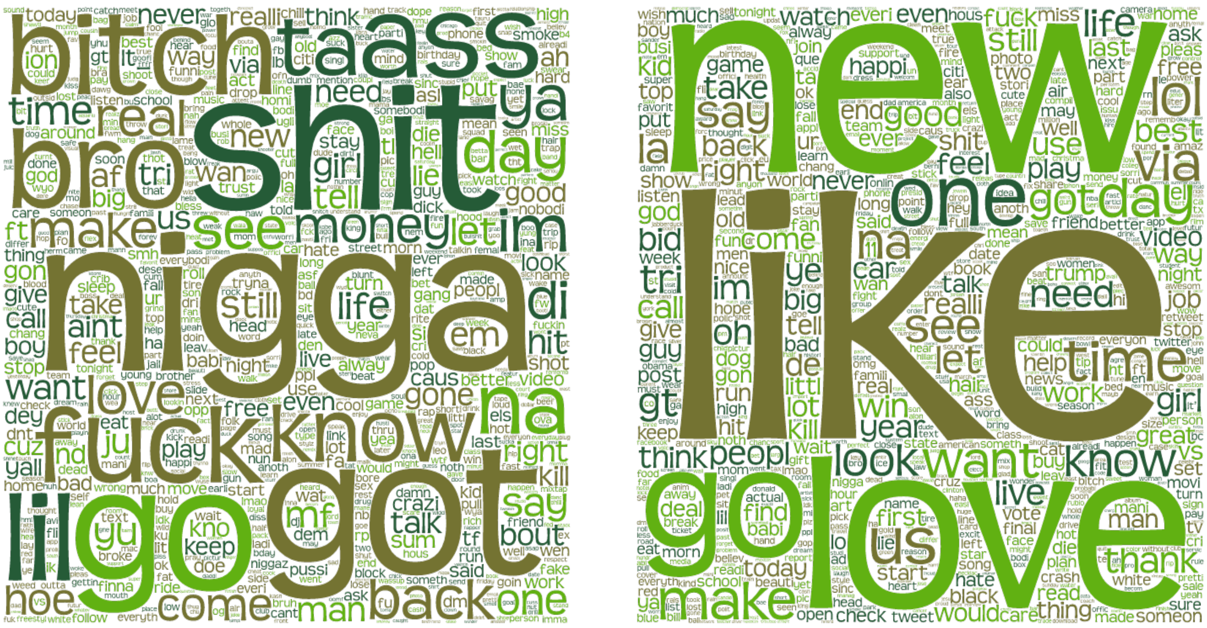
5. Profile Image – We extract image tags using a third-party service for each profile and cover image posted on Twitter and treat the image tags as features.

This chapter provides a detail analysis of each of the above feature types and how well each of them contributed to the task of identifying Twitter gang member profiles.

## 4.1 Tweet text

Tweet text is commonly used to extract features in many Twitter-based studies that analyze the content posted on Twitter [39]. Common features extracted from Tweet text include n-grams, which are the contiguous sequences of n words that appear in a tweet text fragment, and Part-of-Speech (PoS) tags, which are the categories of words that exhibit similar properties or functions based on how words are used in the language. In our experiment, we use unigrams extracted from tweet text as features. We avoid using PoS tags as features in our experiments as we noticed that gang members' tweets contain words that are not available in lexicons that were used to train state-of-the-art Twitter PoS taggers (also known as out-of-vocabulary words or OOV), leading PoS taggers to output PoS tag patterns that are not meaningful.

Figure 4.1 summarizes the words seen most often in the gang and non-gang members' tweets as word 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 [41]. In contrast, we found 5.72% of all words posted by gang member accounts to be classified as curse words, which is nearly five times more than the average curse word usage on Twitter. The word 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.



(a) Gang members.

(b) Non-gang members.

Figure 4.1: Comparison of words used in tweets.

## 4.2 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 4.2. 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

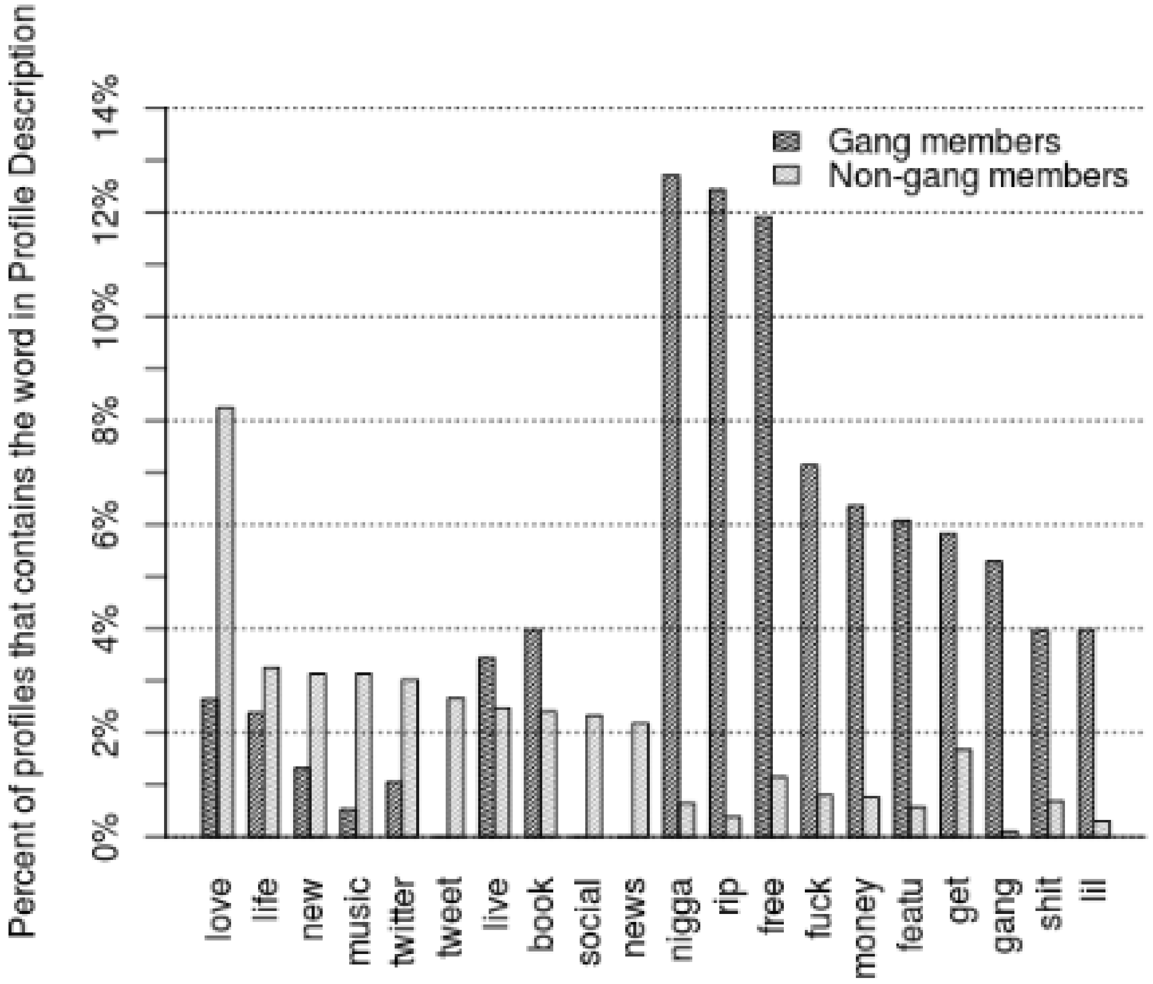


Figure 4.2: Word usage in profile descriptions: gang vs non-gang.

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.

## 4.3 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 [9]. 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 in our dataset.

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[16] 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.

## 4.4 Emoji

Emoji has become a widely used language construct to express emotion in social media. Studies have shown that people associate different meanings to emoji when they use the same emoji in different message contexts [42, 43]. Due to the recent work by Patton *et al.* that discusses the use of emoji by gang members [19], we were motivated to study 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

[16] `https://developers.google.com/youtube/`

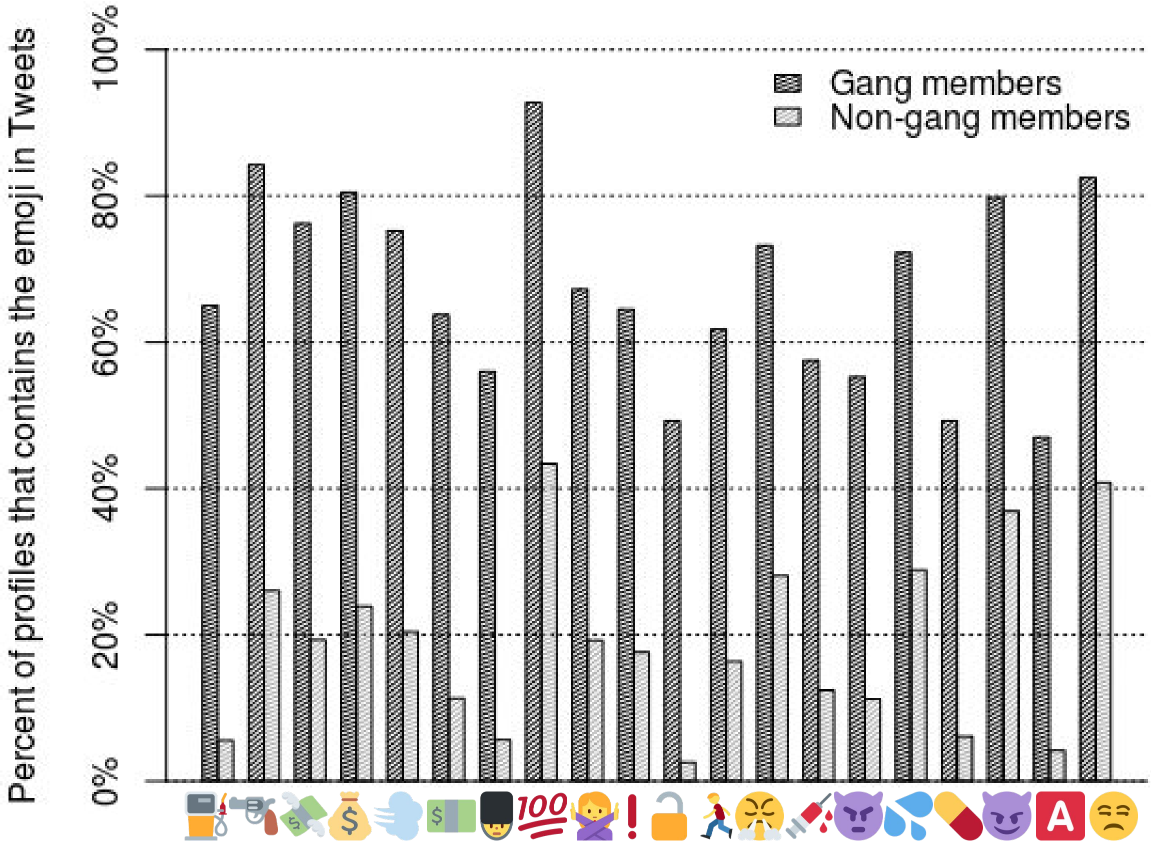
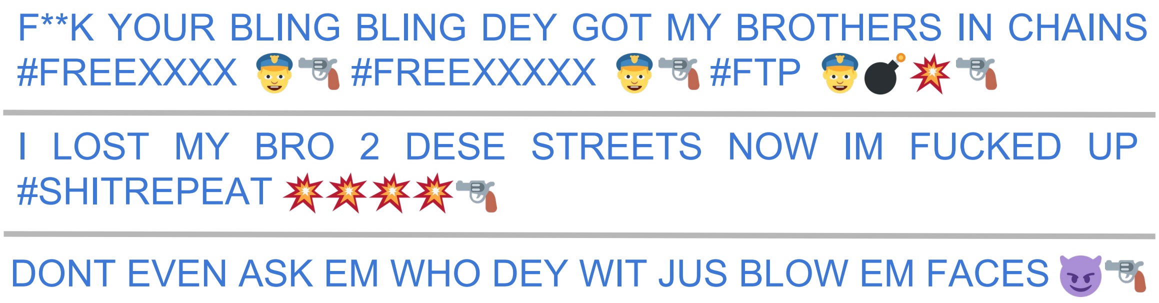


Figure 4.3: Emoji usage distribution: gang vs non-gang.

tweets. We also noticed that gang members use emoji in non-traditional ways when discussing drug-related incidents in their tweets. This aligns with the context-based emoji meanings reported in the emoji-related literature [42, 43].

Figure 4.3 illustrates the emoji distribution for the top 20 most frequent emojis used in 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 4.4 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 by the emoji of a gun 🔫, bomb 💣, or



F**K YOUR BLING BLING DEY GOT MY BROTHERS IN CHAINS #FREEXXXX 👮🔫 #FREEXXXXX 👮🔫 #FTP 👮💣💥🔫

I LOST MY BRO 2 DESE STREETS NOW IM FUCKED UP #SHITREPEAT 💥💥💥💥🔫

DONT EVEN ASK EM WHO DEY WIT JUS BLOW EM FACES 😈🔫

Figure 4.4: Examples for gang members' tweets with emojis.

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.

## 4.5 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 4.6, 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 using Clarifai web service[17]. 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 4.5 offers the 20

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

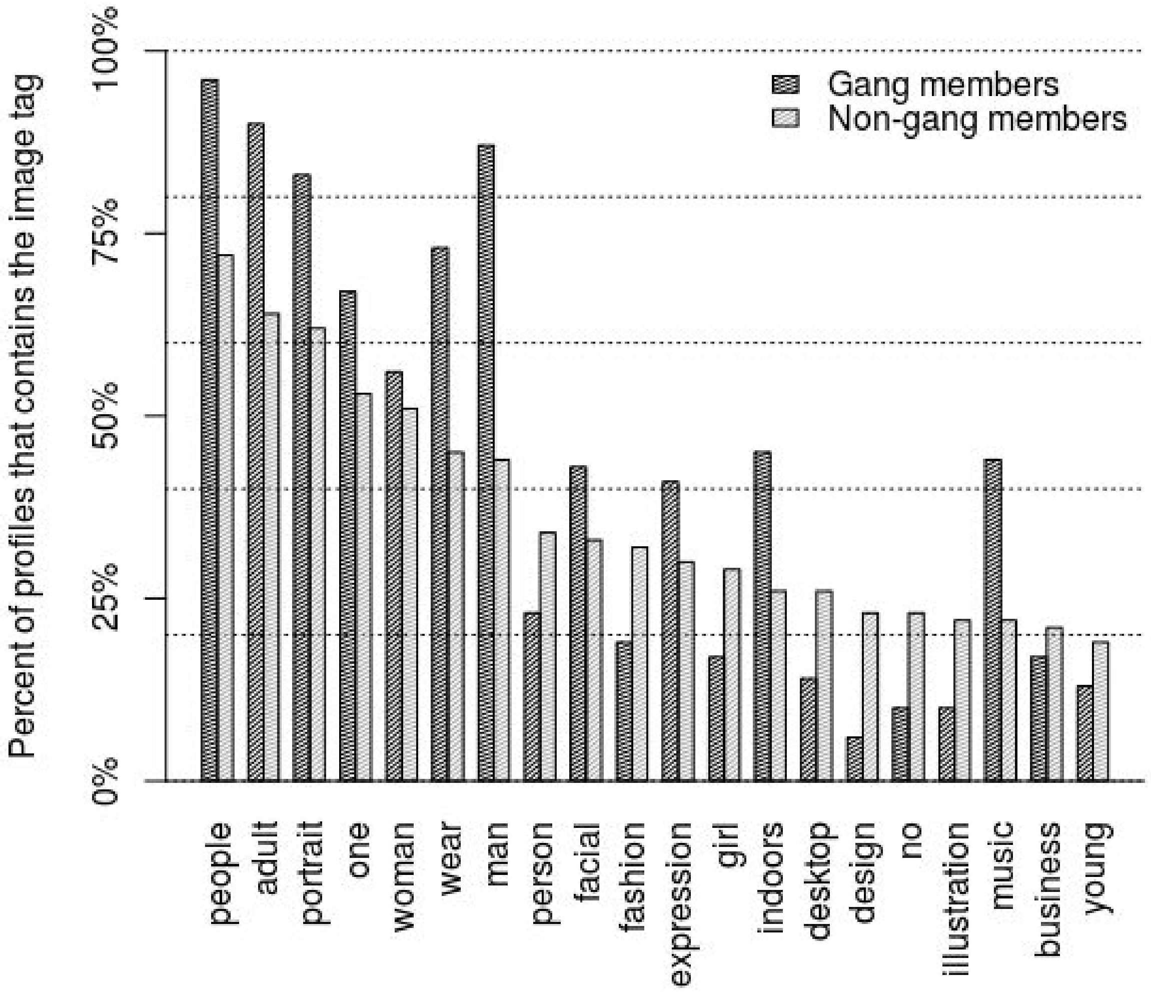


Figure 4.5: Image tags distribution: gang vs non-gang.

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.

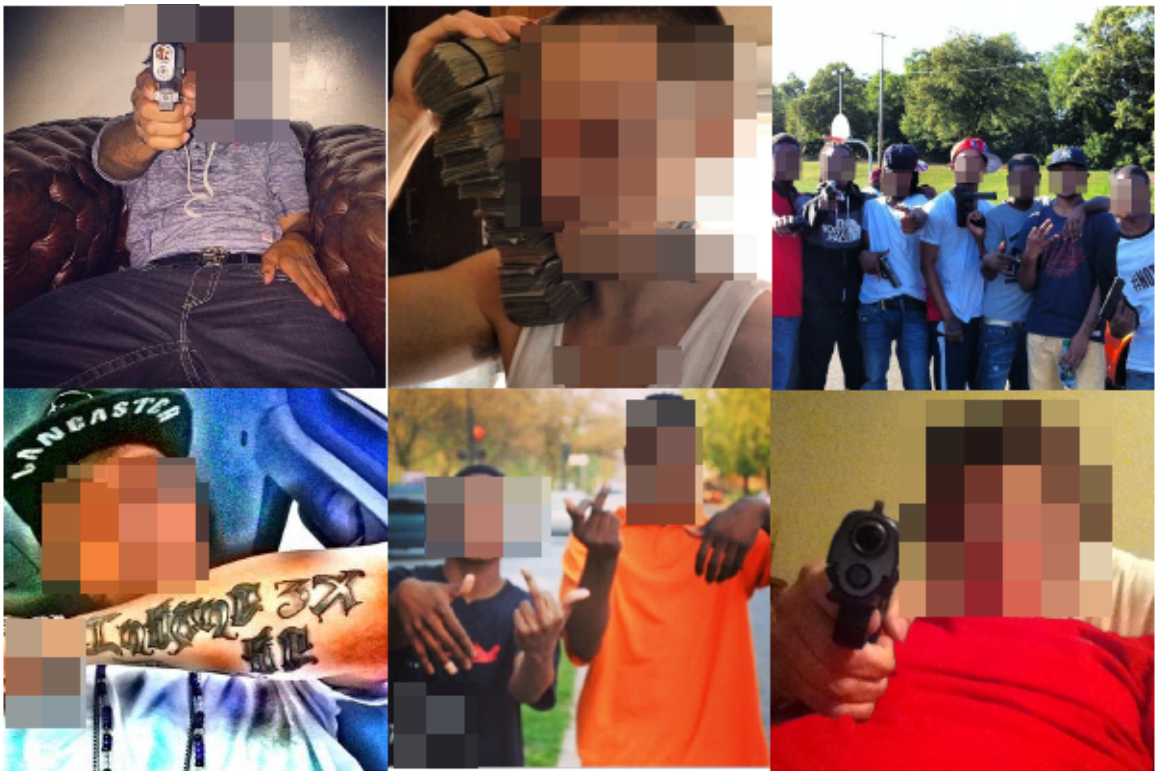


Figure 4.6: Few examples for gang member profile images.

# 5 Approach

This chapter discusses the approach we used to classify gang member profiles on Twitter using a heterogeneous set of features discussed earlier. It also discusses the word embeddings-based methods used to represent the features.

## 5.1 Using Heterogeneous Features

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 algorithms. They are:

1. Naive Bayes Classifier (NB) – This is a conditional probabilistic learning-based classifier which is based on the assumption that the value of a feature is independent of the value of any other feature for a given the class variable.

2. Logistic Regression Classifier (LR) – This is a classification algorithm which takes a categorical dependent variable and requires the outcome to take membership in one of a limited number of categories.

3. Random Forest Classifier (RF) – This is an ensemble of decision trees which is based on the intuition that a large set of weak learners (decision trees in random forest) can be used together to create a strong learner (random forest). Random forest creates a set of decision trees where each decision tree is created from a random sample with replacement of the training set and a random subset of the features.

4. Support Vector Machine (SVM) – This is a popular supervised machine learning algorithm that tries to find the best hyperplane which can separate the classes in the training data. The hyperplane that represents the largest separation, or margin, between the two classes (maximum-margin hyperplane) can be selected as the best hyperplane. To perform non-linear classification SVM is using a technique known as 'kernel' to map inputs into a high-dimensional feature space.

These four algorithms 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.

## 5.2 Representing Text Using Word Embeddings

We also explored using word embeddings to represent our features. Word embedding models are neural language models that tries to learn rich representations for words in a text corpus in a way that the representations it learn better capture the syntactic and semantic similarities of the words in the corpus. They try to learn embeddings in high dimensional spaces (words mapped in to vectors) thus capturing semantic similarities among words which were not possible to capture using other well performing models such as bag-of-words or n-gram models. Recent studies have shown that word embeddings learned with skip-gram based models using negative sampling better capture the context of a word, thus learning to rich word embeddings [32]. Due to their recent success in variety of text processing tasks along with the introduction of easy to use off the shelf tools like Word2Vec to learn word embeddings, they have become very popular and continue to improve the state-of-the-art in text classification

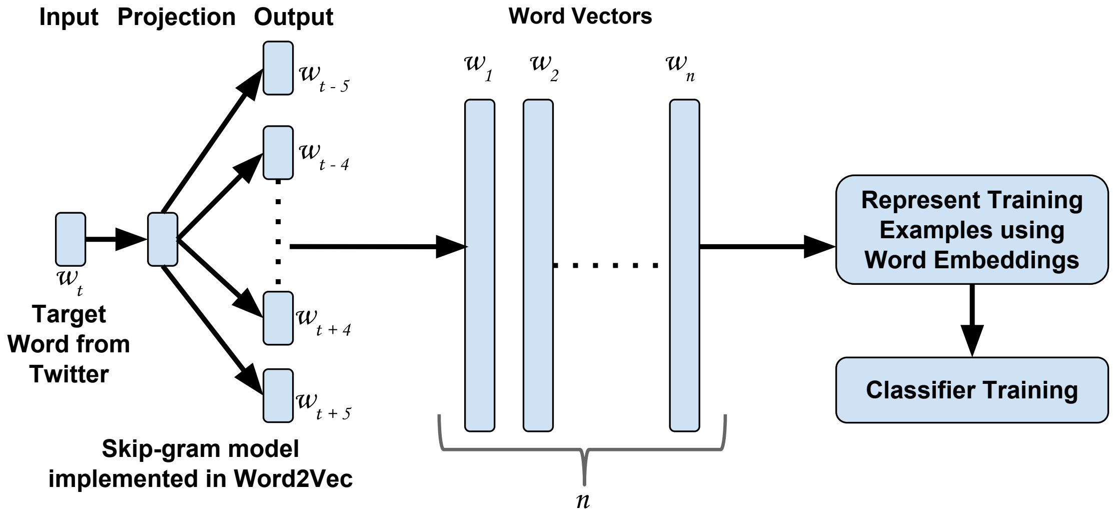


Figure 5.1: Classifier training with word embeddings.

tasks [44]. Previous research have shown that word embeddings work best when it is given with large amounts of training data [45]. Therefore we choose to use word embeddings to improve our classifiers using the 3,265 gang and non-gang member profile dataset we collected as training data to learn word embeddings.

Figure 5.1 shows the steps involved in learning the word embeddings and using them to build classifiers. First we converted non textual features such as emojis, profile and cover images into textual features. Then the seed words used for data collection were removed. We further pre-processed the dataset by removing stop words and stemming all profile descriptions and tweet text. We used the Word2Vec tool along with our pre-processed dataset to train a skip-gram model with negative sampling. Skip-gram model tries to predict a target word given it's context words, which are typically the words surrounded by the target word. It is formally defined in [32].

When training the skip-gram model, we set the negative sampling to 10 sample words, which seems to work well with medium size datasets. We set the context word window to be 5, so that it will consider 5 words to left and right of the target

| Number of Words in | Gang Members | Non-gang Members | **Total** |
|---|---|---|---|
| Tweets | 3,825,092 | 45,213,027 | **49,038,119** |
| Profiles | 3,348 | 21,182 | **24,530** |
| Emoji | 732,712 | 3,685,669 | **4,418,381** |
| Videos | 554,857 | 10,459,235 | **11,014,092** |
| Images | 10,162 | 73,252 | **83,414** |
| **Total** | **5,126,171** | **59,452,365** | **64,578,536** |

Table 5.1: Statistics of the dataset used for training of word embeddings.

word. This setting is suitable for sentences where average sentence length is less than 11 words, which is the case in tweets. We ignore the words that occur less than 5 times in our training corpus. Table 5.1 provides statistics on the number of words found in each type of feature used to train the word embedding model. We obtain word vectors of size 300 using Word2Vec tool. In Figure 5.1, the the total number of word vectors are denoted by $n$ and $i$th word vector is denoted by $w_i$. Once the word vectors are trained, they are used to represent features which is then fed to the learning algorithm used in the classifier.

To represent a Twitter profile, we retrieve word vectors for all the words that appear in a particular profile including the words appear in tweets, profile description, words extracted from emoji, cover and profile images converted to textual formats, and words extracted from YouTube video comments and descriptions for all YouTube videos shared in to the user's timeline. Those word vectors are combined to compute the final feature vector for the Twitter profile. To combine the word vectors, we consider five different methods. Letting the size of a word vector be $k = 300$, for a Twitter profile $p$ with $n$ unique words and the vector of the $i^{th}$ unique word in $p$ denoted by $w_{ip}$, we compute the feature vector for the Twitter profile $V_p$ by:

1. **Sum of word embeddings** $V_{p_{sum}}$ – Sum of the word embedding vectors

obtained for all words in a Twitter profile:

$$V_{p_{sum}} = \sum_{i=0}^{n} w_{ip}$$

2. **Mean of word embeddings** $V_{p_{avg}}$ – Mean of the word embedding vectors of all words found in a Twitter profile:

$$V_{p_{avg}} = 1/n \sum_{i=0}^{n} w_{ip}$$

3. **Sum of word embeddings weighted by term frequency** $V_{p_{sum(count)}}$ – Each word embedding vector multiplied by the word's frequency for the Twitter profile:

$$V_{p_{sum(count)}} = \sum_{i=0}^{n} w_{ip}.c_{ip}$$

where $c_{ip}$ is the term frequency for the $i^{th}$ word in profile $p$.

4. **Sum of word embeddings weighted by** *tf-idf* $V_{p_{sum(tf-idf)}}$ – Each word vector multiplied by the word's *tf-idf* for the Twitter profile:

$$V_{p_{sum(tf-idf)}} = \sum_{i=0}^{n} w_{ip}.t_{ip}$$

where $t_{ip}$ is the *tf-idf* value for the $i^{th}$ word in profile $p$.

5. **Mean of word embeddings weighted by term frequency** $V_{p_{avg(sum(count))}}$ – Mean of the word embedding vectors weighted by term frequency:

$$V_{p_{avg(sum(count))}} = 1/n \sum_{i=0}^{n} w_{ip}.c_{ip}$$

# 6 Evaluation

This chapter presents the evaluation of our approach to automatically find gang member profiles on Twitter. We first discuss the experimental setup used and then we report the evaluation results for our approach using heterogeneous content types. Finally, we present the results for using word embedding along with heterogeneous content types.

## 6.1 Evaluation - Using Heterogeneous Features

We first evaluate the performance of classifiers that use the heterogeneous features to discover gang member profiles on Twitter. For this purpose, we use the training set discussed in Section 3 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 5.1 under a 10-fold cross validation scheme. 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 $F1$-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.

For each of the four learning algorithms (i.e., NB, LR, RF, and SVM), we consider variations involving only tweet text, emoji, profile, image, or music interest (YouTube

| **Features** | **Total Number of Profiles** |
|---|---|
| Tweets (T) | 3,265 {400 : 2,865} |
| Emojis (E) | 3,085 {396 : 2,689} |
| Profile data (P) | 2,996 {378 : 2,618} |
| Image tags (I) | 2,910 {357 : 2,553} |
| Music interest (Y) | 1,630 {196 : 1,434} |
| Model(1) {T+E+P+I+Y} | 3,265 {400 : 2,865} |
| Model(2) {T+E+P+I+Y} | 1,358 {172 : 1,186} |

Table 6.1: Number of profiles available for each feature type.

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 Twitter profiles that had at least one emoji in their tweets to train classifiers which are entirely based on emoji features. We found 3,085 such profiles out of the 3,265 profiles in the training set. Table 6.1 reports, 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. 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[18] machine learning library to implement the classifiers.

### 6.1.1 Experimental results

Table 6.2 presents the average precision, recall, and $F1$-score over the 10 folds for the single-feature and combined feature classifiers. 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 4.1 and discussed in Section 4.1 of Chapter 4. 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

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

| **Features** | **Classifier** | **Results** | | | | | |
|---|---|---|---|---|---|---|---|
| | | **Gang** | | | **Non-Gang** | | |
| | | **Precision** | **Recall** | **$F1$-score** | **Precision** | **Recall** | **$F1$-score** |
| Tweets (T) | 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) | 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) | 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) | 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) | 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} | 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} | Naive Bayes | 0.4405 | 0.9386 | 0.5926 | 0.9889 | 0.8254 | 0.8991 |
| | 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 |

Table 6.2: Classification results based on 10-fold cross validation.

$F1$-scores no matter which classifier considered. There may be two reasons for this despite the differences we observed in Chapter 4. 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 4.5 such as 'people', 'man', and 'adult'.

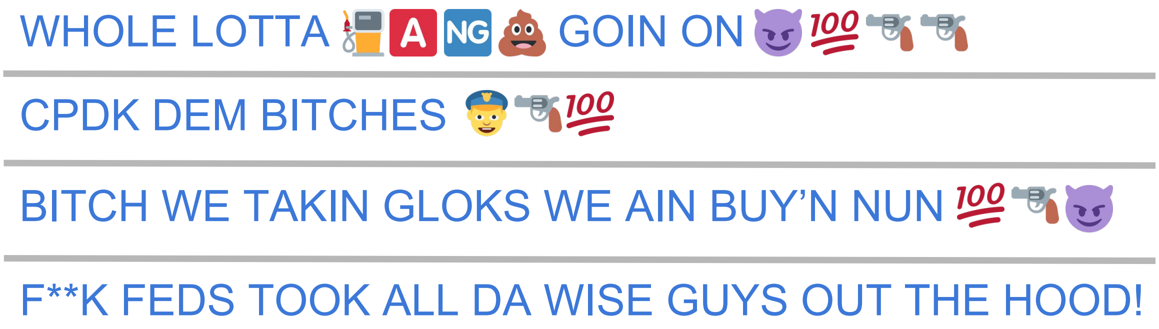
Combining these diverse sets of features into a single classifier yields even better results. Our results for *Model(1)* show that the Random Forest achieves the highest $F1$-scores for both 'gang' (0.7364) and 'non-gang' (0.9690) classes *and* yields the best precision of 0.8792, which corresponds to 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.

### 6.1.2 Evaluation Over Unseen Profiles

To evaluate our classifiers on completely unseen Twitter profiles, we first created a Twitter dataset of random Twitter profiles collected from two U.S. cities that are known for gang-related activities. We captured real-time tweets from Los Angeles, CA[19] and from ten South Side, Chicago neighborhoods [15] 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 then tested the trained classifiers using the above unseen dataset. First, we used our best performing random forest classifier (which use all feature types) and tested it on the unseen 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 profiles are depicted in Figure 6.1. 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

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



WHOLE LOTTA ⛽🅰🆖💩 GOIN ON 😈💯🔫🔫

CPDK DEM BITCHES 👮🔫💯

BITCH WE TAKIN GLOKS WE AIN BUY'N NUN 💯🔫😈

F**K FEDS TOOK ALL DA WISE GUYS OUT THE HOOD!

Figure 6.1: Sample tweets from identified gang members.

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.

## 6.2 Evaluation - Representing Text Using Word Embeddings

We built classifiers using three different learning algorithms, namely Logistic Regression (LR), Random Forest (RF), and Support Vector Machines (SVM). We used version 0.17.1 of scikit-learn[20] machine learning library for Python to implement the classifiers. An open source Python library, Gensim [46] was used to generate the word embeddings. We compare our results with the two best performing models reported in our previous experiment.

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

| Model | Classifier | Gang | | | Non-Gang | | |
|---|---|---|---|---|---|---|---|
| | | **Precision** | **Recall** | **$F1$-score** | **Precision** | **Recall** | **$F1$-score** |
| Baseline *Model(1)* | Random Forest | 0.8792 | 0.6374 | 0.7364 | 0.9507 | **0.9881** | 0.9690 |
| Baseline *Model(2)* | Random Forest | **0.8961** | 0.6994 | 0.7755 | 0.9575 | 0.9873 | 0.9720 |
| $V_{p_{sum}}$ | Logistic Regression | 0.6007 | 0.7045 | 0.6459 | 0.9576 | 0.9346 | 0.9458 |
| | Random Forest | 0.7412 | 0.7085 | 0.7213 | 0.9596 | 0.9659 | 0.9626 |
| | SVM | 0.5929 | **0.7728** | 0.6559 | **0.9661** | 0.9116 | 0.9369 |
| $V_{p_{avg}}$ | Logistic Regression | 0.8394 | 0.5789 | 0.6824 | 0.9442 | 0.9850 | 0.9641 |
| | Random Forest | 0.7627 | 0.7439 | 0.7501 | 0.9650 | 0.9675 | 0.9662 |
| | SVM | 0.8405 | 0.7217 | 0.7740 | 0.9624 | 0.9807 | 0.9715 |
| $V_{p_{sum(count)}}$ | Logistic Regression | 0.6768 | 0.6699 | 0.6681 | 0.9537 | 0.9540 | 0.9537 |
| | Random Forest | 0.7484 | 0.7346 | 0.7386 | 0.9631 | 0.9648 | 0.9639 |
| | SVM | 0.5656 | 0.7180 | 0.6267 | 0.9594 | 0.9212 | 0.9395 |
| $V_{p_{sum(tf-idf)}}$ | Logistic Regression | 0.7901 | 0.7078 | 0.7438 | 0.9595 | 0.9742 | 0.9667 |
| | Random Forest | 0.7979 | 0.7074 | 0.7470 | 0.9598 | 0.9746 | 0.9671 |
| | SVM | 0.7352 | 0.6810 | 0.6952 | 0.9557 | 0.9628 | 0.9589 |
| $V_{p_{avg(sum(count))}}$ | Logistic Regression | 0.8490 | 0.7327 | **0.7835** | 0.9634 | 0.9815 | **0.9723** |
| | Random Forest | 0.7657 | 0.7443 | 0.7519 | 0.9650 | 0.9678 | 0.9663 |
| | SVM | 0.7921 | 0.7194 | 0.7500 | 0.9615 | 0.9735 | 0.9674 |

Table 6.3: Classification results based on 10-fold cross validation.

Table 6.3 presents 10-fold cross validation experiment results for baseline models (first and second rows) and our word embeddings-based models (from third row to seventh row). As mentioned earlier both baseline models use a random forest classifier trained on term frequencies of unigram features extracted from all feature types, and the two baseline models only differs on the training data filtering method based on the availability of features in the training dataset as described in [47]. The baseline *Model(1)* uses all profiles in the dataset and has a $F1$-score of 0.7364 for ‘gang’ class and 0.9690 for ‘non-gang’ class. The baseline *Model(2)* which only uses profiles that contain each and every feature type has a $F1$-score of 0.7755 for ‘gang’ class and $F1$-score of 0.9720 for ‘non-gang’ class.

Vector sum ($V_{p_{sum}}$) is one of the basic operations we can perform on word embedding vectors. The random forest classifier performs the best among vector sum-based classifiers where logistic regression and SVM classifiers also perform comparatively well. Using vector mean ($V_{p_{avg}}$) improves all classifier results and SVM classifier trained on mean of word embeddings achieves very close results to the baseline *Model(2)*. Multiplying vector sum with corresponding word counts for each word in word embeddings ($V_{p_{sum(count)}}$) degrades the classifier accuracy for correctly identifying the positive class. When we multiply words by their corresponding $tf$-*idf* values before taking the vector sum, we again observe an increase in classifier accuracy ($V_{p_{sum(tf-idf)}}$). But we achieve the best performance by averaging the vector sum weighted by term frequency ($V_{p_{avg(sum(count))}}$). Here we multiply the mean of the word embeddings by count of each word, which beats all other word embeddings-based models and the two baselines. In this setting, logistic regression classifier trained on word embeddings performs the best with a $F1$-score of 0.7835. This is a 6.39% improvement in performance when compared to the baseline *Model(1)* and a 1.03% improvement in performance when compared to baseline *Model(2)*. Overall, out of the five vector operations that we used to train machine learning classifiers, four gave us classifier models that beat baseline *Model(1)*. Two vector based operations gave us classifier models that either achieved very similar results to baseline *Model(2)* or beat it. This evaluation demonstrates the promise of using pre-trained word embeddings to boost the accuracy of supervised learning algorithms for Twitter gang member profile classification.

# 7 Conclusion and Future Work

This thesis presented an approach to address the problem of automatically identifying gang member profiles on Twitter. Developing such automated systems is challenging, mainly due to difficulties in finding online gang member profiles for developing training datasets. We outlined a process to curate one of the largest sets of verifiable gang member Twitter profiles that have ever been studied. 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 classifiers achieved promising F1-score over the gang profiles. *Model(1)* uses all profiles in the dataset and has a F1-score of 0.7364 and *Model(2)* which only uses profiles that contain each and every feature type has a F1-score of 0.7755. We then explored using word embeddings to represent features in our classifiers. Our experiments demonstrated that word embeddings achieved superior performance a $F1$–score of 0.7835. This is a 6.39% improvement in performance when compared to the *Model(1)* and a 1.03% improvement in performance when compared to *Model(2)*.

The work discussed in this thesis can be extended in several ways. One obvious way to improve the classification models is to strengthen our training dataset by including more gang member Twitter profiles by searching for more location-independent keywords. We believe more labeled data can lead to better word embedding models which will eventually improve the accuracy of the final classification models. Another way to improve the classification models is by introducing custom image tagging models that are specifically designed to identify commonly seen objects in gang members' profile images. The image tagging service we used was not trained on images specific to gang member tweets such as gang hand signs or pointed guns. Thus, we noticed that the image-based features obtained from the Clarify image tagging service tend to tag images with generic keywords such as 'people' or 'hands'. Building our own image classification system specifically designed to classify images found on gang member profiles could improve the image-based classification models. Past research has also shown that carefully incorporating domain-specific knowledge into machine learning problems can improve the performance them [48]. Thus, crowd-sourced knowledge-bases such as HipWiki[21] that can be utilized to automatically extract gang names and gang-related slang terms can be used to further improve word embedding models. Another way to improve the classification accuracy is to experiment whether *"having a gang name in the profile description"* as a feature can improve our results rather than treating gang names as unigram features.

[21] http://www.hipwiki.com/Hip+Hop+Wiki

# Bibliography

[1] *National Gang Report, National Gang Intelligence Center*, 2015.

[2] *2011 National Gang Threat Assessment Issued Emerging Trends*, 2011.

[3] *National Gang Report, National Gang Intelligence Center*, 2013.

[4] W. B. Miller, *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.

[5] "Gang homicides-five us cities, 2003-2008." *Morbidity and mortality weekly report*, vol. 61, no. 3, pp. 46–51, 2012.

[6] *Survey of Gang Members' Online Habits and Participation (2007) Survey results reported at the i-SAFE Annual Internet Safety Education Review Meeting Carlsbad, California.* National Assessment Center.

[7] S. Decker and D. Pyrooz, "Leaving the gang: Logging off and moving on. council on foreign relations," 2011.

[8] D. C. Pyrooz, S. H. Decker, and R. K. M. Jr., "Criminal and routine activities in online settings: Gangs, offenders, and the internet," *Justice Quarterly*, vol. 32, no. 3, pp. 471–499, 2015.

[9] D. U. Patton, R. D. Eschmann, and D. A. Butler, "Internet banging: New trends in social media, gang violence, masculinity and hip hop," *Computers in Human Behavior*, vol. 29, no. 5, pp. A54 – A59, 2013.

[10] J. C. Howell, "Gang prevention: An overview of research and programs. juvenile justice bulletin." *Office of Juvenile Justice and Delinquency Prevention*, 2010.

[11] M. Ito, S. Baumer, M. Bittanti, R. Cody, B. Herr-Stephenson, H. A. Horst, P. G. Lange, D. Mahendran, K. Z. Martínez, C. Pascoe *et al.*, "Hanging out, messing around, and geeking out," *Digital media*, 2010.

[12] D. C. Pyrooz, S. H. Decker, and R. K. Moule Jr, "Criminal and routine activities in online settings: Gangs, offenders, and the internet," *Justice Quarterly*, no. ahead-of-print, pp. 1–29, 2013.

[13] C. Morselli and D. Décary-Hétu, "Crime facilitation purposes of social networking sites: A review and analysis of the 'cyberbanging' phenomenon," *Small Wars & Insurgencies*, vol. 24, no. 1, pp. 152–170, 2013.

[14] D. U. Patton, "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.

[15] S. Wijeratne, D. Doran, A. Sheth, and J. L. Dustin, "Analyzing the social media footprint of street gangs," in *IEEE International Conference on Intelligence and Security Informatics (ISI), 2015*, May 2015, pp. 91–96.

[16] D. U. Patton, R. D. Eschmann, C. Elsaesser, and E. Bocanegra, "Sticks, stones and facebook accounts: What violence outreach workers know about social media and urban-based gang violence in chicago," *Computers in human behavior*, vol. 65, pp. 591–600, 2016.

[17] P. E. R. Forum, "Social media and tactical considerations for law enforcement," United States Office of Community Oriented Policing Services and United States Department of Justice, Tech. Rep., 2013.

[18] J. Brunty, L. Miller, and K. Helenek, *Social media investigation for law enforcement.* Routledge, 2014.

[19] D. U. Patton, J. Lane, P. Leonard, J. Macbeth, and J. R. Smith-Lee, "Gang violence on the digital street: Case study of a south side chicago gang member's twitter communication," *New Media & Society*, 2016.

[20] M. Pennacchiotti and A.-M. Popescu, "A machine learning approach to twitter user classification," 2011.

[21] R. Tinati, L. Carr, W. Hall, and J. Bentwood, "Identifying communicator roles in twitter," in *Proceedings of the 21st International Conference on World Wide Web*, ser. WWW '12 Companion. New York, NY, USA: ACM, 2012, pp. 1161–1168.

[22] W. Liu and D. Ruths, "What's in a name? using first names as features for gender inference in twitter," 2013.

[23] H. Purohit, A. Dow, O. Alonso, L. Duan, and K. Haas, "User taglines: Alternative presentations of expertise and interest in social media," in *2012 International Conference on Social Informatics (Social Informatics), Washington, D.C., USA, December 14-16*, 2012, pp. 236–243.

[24] F. M. Thrasher, *The gang: A study of 1,313 gangs in Chicago.* University of Chicago Press, 1963.

[25] R. L. Bonn, *Criminology*, ser. McGraw-Hill Series in Criminology and Criminal Justice. McGraw-Hill Publishing, 1984.

[26] D. U. Patton, J. S. Hong, M. Ranney, S. Patel, C. Kelley, R. Eschmann, and T. Washington, "Social media as a vector for youth violence: A review of the literature," *Computers in Human Behavior*, 2014.

[27] A. Sheth, H. Purohit, G. A. Smith, J. Brunn, A. Jadhav, P. Kapanipathi, C. Lu, and W. Wang, *Twitris: A System for Collective Social Intelligence*, 2nd ed. New York: Springer-Verlag New York, 05/2018 2018, pp. 1–23.

[28] D. Décary-Hétu and C. Morselli, "Gang presence in social network sites," *International Journal of Cyber Criminology*, vol. 5, no. 2, p. 876, 2011.

[29] S. M. Radil, C. Flint, and G. E. Tita, "Spatializing social networks: Using social network analysis to investigate geographies of gang rivalry, territoriality, and violence in los angeles," *Annals of the Association of American Geographers*, vol. 100, no. 2, pp. 307–326, 2010. [Online]. Available: https://doi.org/10.1080/00045600903550428

[30] M. Piergallini, A. S. Doğruöz, P. Gadde, D. Adamson, and C. Rose, "Modeling the use of graffiti style features to signal social relations within a multi-domain learning paradigm," in *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, 2014, pp. 107–115.

[31] T. Mikolov, W. Yih, and G. Zweig, "Linguistic regularities in continuous space word representations," in *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.* Atlanta, Georgia: Association for Computational Linguistics, June 2013, pp. 746–751. [Online]. Available: http://www.aclweb.org/anthology/N13-1090

[32] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in Neural Information Processing Systems 26.* Curran Associates, Inc., 2013, pp. 3111–3119. [Online]. Available: http://papers.nips.cc/paper/

5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf

[33] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[34] Y. Bengio, R. Ducharme, P. Vincent, and C. Janvin, "A neural probabilistic language model," *J. Mach. Learn. Res.*, vol. 3, pp. 1137–1155, Mar. 2003. [Online]. Available: http://dl.acm.org/citation.cfm?id=944919.944966

[35] P. Wang, B. Xu, J. Xu, G. Tian, C.-L. Liu, and H. Hao, "Semantic expansion using word embedding clustering and convolutional neural network for improving short text classification," *Neurocomputing*, vol. 174, Part B, pp. 806 – 814, 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0925231215014502

[36] J. Lilleberg, Y. Zhu, and Y. Zhang, "Support vector machines and word2vec for text classification with semantic features," in *Cognitive Informatics Cognitive Computing (ICCI*CC), 2015 IEEE 14th International Conference on*, July 2015, pp. 136–140.

[37] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *CoRR*, vol. abs/1301.3781, 2013. [Online]. Available: http://arxiv.org/abs/1301.3781

[38] S. Wijeratne, L. Balasuriya, D. Doran, and A. Sheth, "Word embeddings to enhance twitter gang member profile identification," in *IJCAI Workshop on Semantic Machine Learning (SML 2016)*. New York City, NY: CEUR-WS, 07/2016 2016.

[39] S. Wijeratne, A. Sheth, S. Bhatt, L. Balasuriya, H. S. Al-Olimat, M. Gaur, A. H. Yazdavar, and K. Thirunarayan, *Feature Engineering for Twitter-based*

*Applications*. Chapman and Hall. Data Mining and Knowledge Discovery Series, December 2017.

[40] L. Balasuriya, S. Wijeratne, D. Doran, and A. Sheth, "Signals revealing street gang members on twitter," in *Workshop on Computational Approaches to Social Modeling (ChASM 2016) co-located with 8th International Conference on Social Informatics (SocInfo 2016)*, vol. 4, Bellevue, WA, USA, 11 2016.

[41] W. Wang, L. Chen, K. Thirunarayan, and A. P. Sheth, "Cursing in english on twitter," in *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing*, ser. CSCW '14. New York, NY, USA: ACM, 2014, pp. 415–425.

[42] S. Wijeratne, L. Balasuriya, A. Sheth, and D. Doran, "Emojinet: Building a machine readable sense inventory for emoji," in *8th International Conference on Social Informatics (SocInfo)*, Bellevue, WA, USA, November 2016, pp. 527–541.

[43] S. Wijeratne, L. Balasuriya, A. P. Sheth, and D. Doran, "Emojinet: An open service and api for emoji sense discovery," in *11th International AAAI Conference on Web and Social Media (ICWSM)*, Montreal, Canada, May 2017, pp. 437–446.

[44] Q. V. Le and T. Mikolov, "Distributed representations of sentences and documents," *CoRR*, vol. abs/1405.4053, 2014. [Online]. Available: http://arxiv.org/abs/1405.4053

[45] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1798–1828, Aug. 2013. [Online]. Available: http://dx.doi.org/10.1109/TPAMI.2013.50

[46] R. Řehůřek and P. Sojka, "Software Framework for Topic Modelling with Large Corpora," in *Proceedings of the LREC 2010 Workshop on New Challenges for*

*NLP Frameworks*. Valletta, Malta: ELRA, May 2010, pp. 45–50. [Online]. Available: http://is.muni.cz/publication/884893/en

[47] L. Balasuriya, S. Wijeratne, D. Doran, and A. Sheth, "Finding street gang members on twitter," in *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, vol. 8, San Francisco, CA, USA, August 2016, pp. 685–692.

[48] A. P. Sheth, S. Perera, S. Wijeratne, and K. Thirunarayan, "Knowledge will propel machine understanding of content: extrapolating from current examples," in *Proceedings of the International Conference on Web Intelligence, Leipzig, Germany, August 23-26, 2017*, 2017, pp. 1–9. [Online]. Available: http://doi.acm.org/10.1145/3106426.3109448