Modeling the Use of Graffiti Style Features to Signal Social Relations within a Multi-Domain Learning Paradigm

Mario Piergallini[1], A. Seza Doğruöz[2], Phani Gadde[1], David Adamson[1], Carolyn P. Rosé[1,3]

[1]Language Technologies Institute
Carnegie Mellon University
5000 Forbes Avenue,
Pittsburgh PA, 15213
{mpiergal,pgadde,
dadamson}@cs.cmu.edu

[2]Tilburg University, TSH,
5000 LE Tilburg, The
Netherlands/
Language Technologies
Institute, Carnegie Mellon
University, 5000 Forbes
Ave.,Pittsburgh PA 15213
a.s.dogruoz@gmail.com

[3]Human-Computer
Interaction Institute
Carnegie Mellon University
5000 Forbes Avenue,
Pittsburgh PA, 15213
cprose@cs.cmu.edu

Abstract

In this paper, we present a series of experiments in which we analyze the usage of graffiti style features for signaling personal gang identification in a large, online street gangs forum, with an accuracy as high as 83% at the gang alliance level and 72% for the specific gang. We then build on that result in predicting how members of different gangs signal the relationship between their gangs within threads where they are interacting with one another, with a predictive accuracy as high as 66% at this thread composition prediction task. Our work demonstrates how graffiti style features signal social identity both in terms of personal group affiliation and between group alliances and oppositions. When we predict thread composition by modeling identity and relationship simultaneously using a multi-domain learning framework paired with a rich feature representation, we achieve significantly higher predictive accuracy than state-of-the-art baselines using one or the other in isolation.

1 Introduction

Analysis of linguistic style in social media has grown in popularity over the past decade. Popular prediction problems within this space include gender classification (Argamon et al., 2003), age classification (Argamon et al., 2007), political affiliation classification (Jiang & Argamon, 2008), and sentiment analysis (Wiebe et al., 2004). From a sociolinguistic perspective, this work can be thought of as fitting within the area of machine learning approaches to the analysis of style (Biber & Conrad, 2009), perhaps as a counterpart to work by variationist sociolinguists in their effort to map out the space of language variation and its accompanying social interpretation (Labov, 2010; Eckert & Rickford, 2001). One aspiration of work in social media analysis is to contribute to this literature, but that requires that our models are interpretable. The contribution of this paper is an investigation into the ways in which stylistic features behave in the language of participants of a large online community for street gang members. We present a series of experiments that reveal new challenges in modeling stylistic variation with machine learning approaches. As we will argue, the challenge is achieving high predictive accuracy without sacrificing interpretability.

Gang language is a type of sociolect that has so far not been the focus of modeling in the area of social media analysis. Nevertheless, we argue that the gangs forum we have selected as our data source provides a strategic source of data for exploring how social context influences stylistic language choices, in part because it is an area where the dual goals of predictive accuracy and interpretability are equally important. In particular, evidence that gang related crime may account for up to 80% of crime in the United States attests to the importance of understanding the social practices of this important segment of society (Johnsons, 2009). Expert testimony attributing meaning to observed, allegedly gang-related social practices is frequently used as evidence of malice in criminal investigations (Greenlee, 2010). Frequently, it is police officers who are given the authority to serve as expert witnesses on this interpretation because of their routine interaction with gang members.

Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, pages 107–115,
Gothenburg, Sweden, April 26-30 2014.

Nevertheless, one must consider their lack of formal training in forensic linguistics (Coulthard & Johnson, 2007) and the extent to which the nature of their interaction with gang members may subject them to a variety of cognitive biases that may threaten the validity of their interpretation (Kahneman, 2011).

Gang-related social identities are known to be displayed through clothing, tattoos, and language practices including speech, writing, and gesture (Valentine, 1995), and even dance (Philips, 2009). Forensic linguists have claimed that these observed social practices have been over-interpreted and inaccurately interpreted where they have been used as evidence in criminal trials and that they may have even resulted in sentences that are not justified by sufficient evidence (Greenlee, 2010). Sociolinguistic analysis of language varieties associated with gangs and other counter-cultural groups attests to the challenges in reliable interpretation of such practices (Bullock, 1996; Lefkowitz, 1989). If we as a community can understand better how stylistic features behave due to the choices speakers make in social contexts, we will be in a better position to achieve high predictive accuracy with models that are nevertheless interpretable. And ultimately, our models may offer insights into usage patterns of these social practices that may then offer a more solid empirical foundation for interpretation and use of language as evidence in criminal trials.

In the remainder of the paper we describe our annotated corpus. We then motivate the technical approach we have taken to modeling linguistic practices within the gangs forum. Next, we present a series of experiments evaluating our approach and conclude with a discussion of remaining challenges.

2 The Gangs Forum Corpus

The forum that provides data for our experiments is an online forum for members of street gangs. The site was founded in November, 2006. It was originally intended to be an educational resource compiling knowledge about the various gang organizations and the street gang lifestyle. Over time, it became a social outlet for gang members. There are still traces of this earlier focus in that there are links at the top of each page to websites dedicated to information about particular gangs. At the time of scraping its contents, it had over a million posts and over twelve thousand active users. Our work focuses on analysis of stylistic choices that are influenced by social context, so it is important to consider some details about the social context of this forum. Specifically, we discuss which gangs are present in the data and how the gangs are organized into alliances and rivalries. Users are annotated with their gang identity at two levels of granularity, and threads are annotated with labels that indicate which gang dominates and how the participating gangs relate to one another.

2.1 User-Level Annotations

At the fine-grained level, we annotated users with the gang that they indicated being affiliated with, including Bloods, Crips, Hoovers, Gangster Disciples, other Folk Nation, Latin Kings, Vice Lords, Black P. Stones, other People Nation, Trinitarios, Norteños, and Sureños. There was also an Other category for the smaller gangs. For a coarser grained annotation of gang affiliation, we also noted the nation, otherwise known as gang alliance, each gang was associated with.

For our experiments, a sociolinguist with significant domain expertise annotated the gang identity of 3384 users. Information used in our annotation included the user's screen name, their profile, which included a slot for gang affiliation, and the content of their posts. We used regular expressions to find gang names or other identifiers occurring within the gang affiliation field and the screen names and annotated the users that matched. If the value extracted for the two fields conflicted, we marked them as claiming multiple gangs. For users whose affiliation could not be identified automatically, we manually checked their profile to see if their avatar (an image that accompanies their posts) or other fields there contained any explicit information. Otherwise, we skimmed their posts for explicit statements of gang affiliation.

Affiliation was unambiguously identified automatically for 56% of the 3384 users from their affiliation field. Another 36% were identified automatically based on their screen name. Manual inspection was only necessary in 9% of the cases. Users that remained ambiguous, were clearly fake or joke accounts, or who claimed multiple gangs were grouped together in an "Other" category, which accounts for 6.2% of the total. Thus, 94% of the users were classified into the 12 specific gangs mentioned above.

At a coarse-grained level, users were also associated with a nation. The nation category was inspired by the well-known gang alliances known as the People Nation and Folks Nation, which are city-wide alliances of gangs in Chicago. We labeled the Crips and Hoovers as a nation since they are closely allied gangs. Historically, the Hoovers began breaking away from the Crips and are rivals with certain subsets of Crips, but allies with the majority of other Crips gangs. The complex inner structure of the Crips alliance will be discussed in Section 5 where we interpret our quantitative results.

There are a large number of gangs that comprise the People and Folks Nations. The major gangs within the People Nation are the Latin Kings, Vice Lords and Black P. Stones. The Folks Nation is dominated by the Gangster Disciples with other Folks Nation gangs being significantly smaller. The People Nation, Blood and Norteños gangs are in a loose, national alliance against the opposing national alliance of the Folks Nation, Crips and Sureños. Remaining gangs were annotated as other, such as the Trinitarios, that don't fit into this national alliance system nor even smaller alliances.

2.2 Thread-Level Annotations

In addition to person-level annotations of gang and nation, we also annotated 949 threads with dominant gang as well as thread composition, by which we mean whether the users who participated on the thread were only from allied gangs, included opposing gangs, or contained a mix of gangs that were neither opposing nor allied. These 949 threads were ones where a majority of the users who posted were in the set of 3384 users annotated with a gang identity.

For the dominant gang annotation at the gang level, we consider only participants on the thread for whom there was an annotated gang affiliation. If members of a single gang produced the majority of the posts in the thread, then that was annotated as the dominant gang of the thread. If no gang had a majority in the thread, it was instead labeled as Mixed. For dominant gang at the nation level, the same procedure was used, but instead of looking for which gang accounted for more of the members, we looked for which gang alliance accounted for the majority of users.

For the thread composition annotation, we treated the Bloods, People Nation, and Norteños as allied with each other as the "Red set". We treated Crips, Hoovers, Folks Nation, and Sureños as allies with each other as the "Blue set". The Red and Blue sets oppose one another. The Latin Kings and Trinitarios also oppose one another. Thread composition was labeled as Allied, Mixed or Opposing depending on the gangs that appeared in the thread. As with the dominant gang annotation, only annotated users were considered. If all of the posts were by users of the same gang or allied gangs, the thread was labeled as Allied. If there were any posts from rival gangs, it was labeled as Opposing. Otherwise, it was labeled as Mixed. If the users were all labeled with Other as their gang it was also labeled as Mixed.

3 Modeling Language Practices at the Feature Level

In this section, we first describe the rich feature representation we developed for this work. Finally, we discuss the motivation for employing a multi-domain learning framework in our machine-learning experiments.

3.1 Feature Space Design: Graffiti Style Features

While computational work modeling gang-related language practices is scant, we can learn lessons from computational work on other types of sociolects that may motivate a reasonable approach. Gender prediction, for example, is a problem where there have been numerous publications in the past decade (Corney et al., 2002; Argamon et al., 2003; Schler et al., 2005; Schler, 2006; Yan & Yan, 2006; Zhang et al., 2009). Because of the complex and subtle way gender influences language choices, it is a strategic example to motivate our work.

Gender-based language variation arises from multiple sources. Among these, it has been noted that within a single corpus comprised of samples of male and female language that the two genders do not speak or write about the same topics. This is problematic because word-based features such as unigrams and bigrams, which are very frequently used, are highly likely to pick up on differences in topic (Schler, 2006) and possibly perspective. Thus, in cases where linguistic style variation is specifically of interest, these features do not offer good generalizability (Gianfortoni et al., 2011). Similarly, in our work, members of different

gangs are located in different areas associated with different concerns and levels of socioeconomic status. Thus, in working to model the stylistic choices of gang forum members, it is important to consider how to avoid overfitting to content-level distinctions.

Typical kinds of features that have been used in gender prediction apart from unigram features include part-of-speech (POS) ngrams (Argamon et al., 2003), word-structure features that cluster words according to endings that indicate part of speech (Zhang et al., 2009), features that indicate the distribution of word lengths within a corpus (Corney et al., 2002), usage of punctuation, and features related to usage of jargon (Schler et al., 2005). In Internet-based communication, additional features have been investigated such as usage of internet specific features including "internet speak" (e.g., lol, wtf, etc.), emoticons, and URLs (Yan & Yan, 2006).

Transformation	**Origin or meaning**
b^, c^, h^, p^	"Bloods up" Positive towards Bloods, Crips, Hoovers, Pirus, respectively
b → bk, c → ck	Blood killer, Crip killer
h → hk, p → pk	Hoover killer, Piru killer
ck → cc, kc	Avoid use of 'ck' since it represents Crip killer
o → x, o → ø	Represents crosshairs, crossing out the '0's in a name like Rollin' **60s** Crips
b → 6	Represents the six-pointed star. Symbol of Folk Nation and the affiliated Crips.
e → 3	Various. One is the trinity in Trinitario.
s → 5	Represents the five-pointed star. Symbol of People Nation and the affiliated Bloods.

Table 1: Orthographical substitutions from gang graffiti symbolism

In order to place ourselves in the best position to build an interpretable model, our space of graffiti style features was designed based on a combination of qualitative observations of the gangs forum data and reading about gang communication using web accessible resources such as informational web pages linked to the forum and other resources related to gang communication (Adams & Winter, 1997; Garot, 2007). Specifically, in our corpus we observed gang members using what we refer to as graffiti style features to mark their identity. Gang graffiti employs shorthand references to convey affiliation or threats (Adams & Winter, 1997). For example, the addition of a <k> after a letter representing a rival gang stands for "killer." So, writing <ck> would represent "crip killer." A summary of these substitutions can be seen in Table 1. Unfortunately, only about 25% of the users among the 12,000 active users employ these features in their posts, which limits their ability to achieve a high accuracy, but nevertheless offers the opportunity to model a frequent social practice observed in the corpus.

The graffiti style features were extracted using a rule-based algorithm that compares words against a standard dictionary as well as using some phonotactic constraints on the position of certain letters. The dictionary was constructed using all of the unique words found in the AQUAINT corpus (Graff, 2002). If a word in a post did not match any word from the AQUAINT corpus, we tested it against each of the possible transformations in Table 1. Transformations were applied to words using finite state transducers. If some combination transformations from that table applied to the observed word could produce some term from the AQUAINT corpus, then we counted that observed word as containing the features associated with the applied transformations.

The transformations were applied in the order of least likely to occur in normal text to the most likely. Since 'bk' only occurs in a handful of obscure words, for example, almost any occurrence of it can be assumed to be a substitution and the 'k' can safely be removed before the next step. By contrast, 'cc' and 'ck' occur in many common words so they must be saved for last to ensure that the final dictionary checks have any simultaneous substitutions already removed.

When computing values for the graffiti style features for a text, the value for each feature was computed as the number of words (tokens) that contained the feature divided by the total number of words (tokens) in the document. We used a set of 13 of these features, chosen on the basis of how frequently they occurred and how strongly they distinguished gangs from one another (for example, substituting '$' for 's' was a transformation that was common across gangs in

our qualitative analysis, and thus did not seem beneficial to include).

Transformation	Freq.	False Positive rate	False Negative rate
b^, c^, h^, p^	15103	0%	0%
b → bk	26923	1%	0%
c → ck	16144	25%	8%
h → hk	10053	1%	0%
p → pk	5669	3%	0%
ck → cc, kc	72086	2%	0%
o → x, o → ø	13646	15%	5%
b → 6	2470	16%	0%
e → 3	8628	28%	1%
s → 5	13754	6%	0%

Table 2: Evaluation of extraction of graffiti style features over the million post corpus

The feature-extraction approach was developed iteratively. After extracting the features over the corpus of 12,000 active users, we created lists of words where the features were detected, sorted by frequency. We then manually examined the words to determine where we observed errors occurring and then made some minor adjustments to the extractors. Table 2 displays a quantitative evaluation of the accuracy of the graffiti style feature extraction.

Performance of the style features was estimated for each style-feature rule. For each rule, we compute a false positive and false negative rate. For false positive rate, we begin by retrieving the list of words marked by the feature extraction rule containing the associated style marking. From the full set of words that matched a style feature rule, we selected the 200 most frequently occurring word types. We manually checked that complete set of word tokens and counted the number of misfires. The false positive rate was then calculated for each feature by dividing the number of tokens that were misfires over the total number of tokens in the set. In all cases, we ensured that at least 55% of the total word tokens were covered, so additional words may have been examined.

In the case of false negatives, we started with the set of word types that did not match any word in the dictionary and also did not trigger the style feature rule. Again we sorted word types in this list by frequency and selected the top 200 most frequent. We then manually checked for missed instances where the associated style feature was used but not detected. The false negative rate was then the total number of word tokens within this word type set divided by the total number of word tokens in the complete set of word types.

Another type of feature we used referenced the nicknames gangs used for themselves and other gangs, which we refer to as Names features. The intuition behind this is simple: someone who is a member of the Crips gang will talk about the Crips more often. The measure is simply how often a reference to a gang occurs per document. Some of these nicknames we included were gang-specific insults, with the idea that if someone uses insults for Crips often, they are likely not a Crip. The last type of reference is words that refer to gang alliances like the People Nation and Folks Nation. Members of those Chicago-based gangs frequently refer to their gang as the "Almighty [gang name] Nation".

Gang	Positive/Neutral Mentions	Insults
Crips	crip, loc	crab, ckrip, ck
Bloods	blood, damu, piru, ubn	slob, bklood, pkiru, bk, pk
Hoovers	hoover, groover, crim, hgc, hcg	snoover, hkoover, hk
Gangster Disciples	GD, GDN, Gangster Disciple	gk, dk, nigka
Folks Nations	folk, folknation, almighty, nation	
People Nation	people, peoplenation, almighty, nation	
Latin Kings	alkqn, king, queen	
Black P. Stones	stone, abpsn, moe, black p.	
Vice Lords	vice, lord, vl, avln, foe, 4ch	

Table 3: Patterns used for gang name features. For all gangs listed in the table, there are slang terms used as positive mentions of the gang. For some gangs there are also typical insult names.

We used regular expressions to capture occurrences of these words and variations on them such as the use of the orthographic substitutions mentioned previously, plurals, feminine forms, etc. Additionally, in the Blood and Hoover features, they sometimes use numbers to replace the 'o's representing the street that their gang is located on. So the Bloods from 34th Street, say, might write "Bl34d".

3.2 Computational Paradigm: Multi-domain learning

The key to training an interpretable model in our work is to pair a rich feature representation with a model that enables accounting for the structure of the social context explicitly. Recent work in the area of multi-domain learning offers such an opportunity (Arnold, 2009; Daumé III, 2007; Finkel & Manning, 2009). In our work, we treat the dominant gang of a thread as a domain for the purpose of detecting thread composition. This decision is based on the observation that while it is a common practice across gangs to express their attitudes towards allied and opposing gangs using stylistic features like the Graffiti style features, the particular features that serve the purpose of showing affiliation or opposition differ by gang. Thus, it is not the features themselves that carry significance, but rather a combination of who is saying it and how it is being said.

As a paradigm for multi-domain learning, we use Daume's Frustratingly Easy Domain Adaptation approach (Daumé III, 2007) as implemented in LightSIDE (Mayfield & Rosé, 2013). In this work, Daumé III proposes a very simple "easy adapt" approach, which was originally proposed in the context of adapting to a specific target domain, but easily generalizes to multi-domain learning. The key idea is to create domain-specific versions of the original input features depending on which domain a data point belongs to. The original features represent a domain-general feature space. This allows any standard learner to appropriately optimize the weights of domain-specific and domain-general features simultaneously. In our work, this allows us to model how different gangs signal within-group identification and across-group animosity or alliance using different features. The resulting model will enable us to identify how gangs differ in their usage of style features to display social identity and social relations.

It has been noted in prior work that style is often expressed in a topic-specific or even domain-specific way (Gianfortoni et al., 2011). What exacerbates these problems in text processing approaches is that texts are typically represented with features that are at the wrong level of granularity for what is being modeled. Specifically, for practical reasons, the most common types of features used in text classification tasks are still unigrams, bigrams, and part-of-speech bigrams, which are highly prone to over-fitting. When text is represented with features that operate at too fine-grained of a level, features that truly model the target style are not present within the model. Thus, the trained models are not able to capture the style itself and instead capture features that correlate with that style within the data (Gianfortoni et al., 2011).

This is particularly problematic in cases where the data is not independent and identically distributed (IID), and especially where instances that belong to different subpopulations within the non-IID data have different class value distributions. In those cases, the model will tend to give weight to features that indicate the subpopulation rather than features that model the style. Because of this insight from prior work, we contrast our stylistic features with unigram features and our multi-domain approach with a single-domain approach wherever appropriate in our experiments presented in Section 4.

4 Prediction Experiments

In this section we present a series of prediction experiments using the annotations described in Section 2. We begin by evaluating our ability to identify gang affiliation for individual users. Because we will use dominant gang as a domain feature in our multi-domain learning approach to detect thread composition, we also present an evaluation of our ability to automatically predict dominant gang for a thread. Finally, we evaluate our ability to predict thread composition. All of our experiments use L1 regularized Logistic regression.

4.1 Predicting Gang Affiliation per User

The first set of prediction experiments we ran was to identify gang affiliation. For this experiment, the full set of posts contributed by a user was concatenated together and used as a document from which to extract text features. We conducted this experiment using a 10-fold cross-validation over the full set of users annotated for gang affiliation. Results contrasting alternative feature spaces at the gang level and nation level are displayed in Table 4. We begin with a unigram feature space as the baseline. We contrast this with the Graffiti style features described above in Section 3.1. Because all of the Graffiti features are encoded in words as pairs of characters, we contrast the carefully extracted Graffiti style features with character

bigrams. Next we test the nickname features also described in Section 3.1. Finally, we test combinations of these features.

	Gang	**Nation**
Unigrams	70%	81%
Character Bigrams	64%	76%
Graffiti Features	44%	68%
Name Features	63%	78%
Name + Graffiti	67%	81%
Unigrams + Name	70%	82%
Unigrams + Character Bigrams	71%	82%
Unigrams + Graffiti	71%	82%
Unigrams + Name + Graffiti	<u>72%</u>	<u>83%</u>
Unigrams + Name + Character Bigrams	<u>72%</u>	79%

Table 4: Results (percent accuracy) for gang affiliation prediction at the gang and nation level.

We note that the unigram space is a challenging feature space to beat, possibly because only about 25% of the users employ the style features we identified with any regularity. The character bigram space actually significantly outperforms the Graffiti features, in part because it captures aspects of both the Graffiti features, the name features, and also some other gang specific jargon. When we combine the stylistic features with unigrams, we start to see an advantage over unigrams alone. The best combination is Unigrams, Graffiti style features, and Name features, at 72% accuracy (.65 Kappa) at the gang level and 83% accuracy (.69 Kappa) at the nation level. Overall the accuracy is reasonable and offers us the opportunity to expand our analysis of social practices on the gangs forum to a much larger sample in our future work than we present in this first foray.

4.2 Predicting Dominant Gang per Thread

In Section 4.3 we present our multi-domain learning approach to predicting thread composition. In that work, we use dominant gang on a thread as a domain. In those experiments, we contrast results with hand-annotated dominant gang and automatically-predicted dominant gang. In order to compute an automatically-identified dominant gang for the 949 threads used in that experiment, we build a model for gang affiliation prediction using data from the 2689 users who did not participate on any of those threads as training data so there is no overlap in users between train and test.

The feature space for that classifier included unigrams, character bigrams, and the gang name features since this feature space tied for best performing at the gang level in Section 4.1 and presents a slightly lighter weight solution than Unigrams, graffiti style features, and gang name features. We applied that trained classifier to the users who participated on the 949 threads. From the automatically-predicted gang affiliations, we computed a dominant gang using the gang and nation level for each thread using the same rules that we applied to the annotated user identities for the annotated dominant gang labels described in Section 2.2. We then evaluated our performance by comparing the automatically-identified dominant gang with the more carefully annotated one. Our automatically identified dominant gang labels were 73.3% accurate (.63 Kappa) at the gang level and 76.6% accurate (.72 Kappa) at the nation level. This experiment is mainly important as preparation for the experiment presented in Section 4.3.

4.3 Predicting Thread Composition

Our final and arguably most important prediction experiments were for prediction of thread composition. This is where we begin to investigate how stylistic choices reflect the relationships between participants in a discussion. We conducted this experiment twice, specifically, once with the annotated dominant gang labels (Table 5) and once with the automatically predicted ones (Table 6). In both cases, we evaluate gang and nation as alternative domain variables. In both sets of experiments, the multi-domain versions significantly outperform the baseline across a variety of feature spaces, and the stylistic features provide benefit above the unigram baseline. In both tables the domain and nation variables are hand-annotated. * indicates the results are significantly better than the no domain unigram baseline. <u>Underline</u> indicates best result per column. And **bold** indicates overall best result.

The best performing models in both cases used a multi-domain model paired with a stylistic feature space rather than a unigram space. Both models performed significantly better than any of the unigram models, even the multi-domain versions with annotated domains. Where gang was used as the domain variable and Graffiti style features were the features used for prediction, we found that the high weight features associated with Allied threads were

either positive about gang identity for a variety of gangs other than their own (like B^ in a Crips dominated thread) or protective (like CC in a Bloods dominated thread).

	No Domain	**Dominant Gang**	**Dominant Nation**
Unigrams	53%	58%*	60%*
Character Bigrams	49%	55%	56%
Graffiti Features	53%	54%	61%*
Name Features	54%	63%*	**66%***
Name + Graffiti	54%	61%*	65%*
Unigrams + Name	52%	58%*	61%*
Unigrams + Graffiti	53%	57%	57%
Unigrams + Name + Graffiti	54%	61%*	65%*

Table 5: Results (percent accuracy) for thread composition prediction, contrasting a single domain approach with two multi-domain approaches, one with dominant gang as the domain variables, and the other with dominant nation as the domain variable. In this case, the domain variables are annotated.

	No Domain	**Dominant Gang**	**Dominant Nation**
Unigrams	53%	57%	57%
Character Bigrams	49%	53%	55%
Graffiti Features	53%	**65%***	58%*
Name Features	54%	61%*	59%*
Name + Graffiti	54%	60%*	59%*
Unigrams + Name	52%	56%	56%
Unigrams + Graffiti	53%	58%*	57%
Unigrams + Name + Graffiti	54%	60%*	59%*

Table 6: Results (percent accuracy) for thread composition prediction, contrasting a single domain approach with two multi-domain approaches with predicted domain variables, one with dominant gang as the domain variables, and the other with dominant nation as the domain variable.

Crips-related features were the most frequent within this set, perhaps because of the complex social structure within the Crips alliance, as discussed above. We saw neither features associated with negative attitudes of the gang towards others nor other gangs towards them in these Allied threads, but in opposing threads, we see both, for example, PK in Crips threads or BK in Bloods threads. Where unigrams are used as the feature space, the high weight features are almost exclusively in the general space rather than the domain space, and are generally associated with attitude directly rather than gang identity. For example, "lol," and "wtf."

5 Conclusions

We have presented a series of experiments in which we have analyzed the usage of stylistic features for signaling personal gang identification and between gang relations in a large, online street gangs forum. This first foray into modeling the language practices of gang members is one step towards providing an empirical foundation for interpretation of these practices. In embarking upon such an endeavor, however, we must use caution. In machine-learning approaches to modeling stylistic variation, a preference is often given to accounting for variance over interpretability, with the result that interpretability of models is sacrificed in order to achieve a higher prediction accuracy. Simple feature encodings such as unigrams are frequently chosen in a (possibly misguided) attempt to avoid bias. As we have discussed above, however, rather than cognizant introduction of bias informed by prior linguistic work, unknown bias is frequently introduced because of variables we have not accounted for and confounding factors we are not aware of, especially in social data that is rarely IID. Our results suggest that a strategic combination of rich feature encodings and structured modeling approach leads to high accuracy and interpretability. In our future work, we will use our models to investigate language practices in the forum at large rather than the subset of users and threads used in this paper[1].

[1] An appendix with additional analysis and the specifics of the feature extraction rules can be found at http://www.cs.cmu.edu/~cprose/Graffiti.html. This work was funded in part by ARL 000665610000034354.

References

Adams, K. & Winter, A. (1997). Gang graffiti as a discourse genre, *Journal of Sociolinguistics* 1/3. Pp 337-360.

Argamon, S., Koppel, M., Fine, J., & Shimoni, A. (2003). Gender, genre, and writing style in formal written texts, *Text*, 23(3), pp 321-346.

Argamon, S., Koppel, M., Pennebaker, J., & Schler, J. (2007). Mining the blogosphere: age, gender, and the varieties of self-expression. *First Monday* 12(9).

Arnold, A. (2009). Exploiting Domain And Task Regularities For Robust Named Entity Recognition. PhD thesis, Carnegie Mellon University, 2009.

Biber, D. & Conrad, S. (2009). *Register, Genre, and Style*, Cambridge University Press

Bullock, B. (1996). Derivation and Linguistic Inquiry: Les Javnais, *The French Review* 70(2), pp 180-191.

Corney, M., de Vel, O., Anderson, A., Mohay, G. (2002). Gender-preferential text mining of e-mail discourse, in the Proceedings of the 18th Annual Computer Security Applications Conference.

Coulthard, M. & Johnson, A. (2007). *An Introduction to Forensic Linguistics: Language as Evidence*, Routledge

Daumé III, H. (2007). Frustratingly Easy Domain Adaptation. In Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, pages 256-263.

Eckert, P. & Rickford, J. (2001). *Style and Sociolinguistic Variation*, Cambridge: University of Cambridge Press.

Finkel, J. & Manning, C. (2009). Hierarchical Bayesian Domain Adaptation. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics.*

Garot, R. (2007). "Where You From!": Gang Identity as Performance, Journal of Contemporary Ethnography, 36, pp 50-84.

Gianfortoni, P., Adamson, D. & Rosé, C. P. (2011). Modeling Stylistic Variation in Social Media with Stretchy Patterns, in *Proceedings of First Workshop on Algorithms and Resources for Modeling of Dialects and Language Varieties*, Edinburgh, Scottland, UK, pp 49-59.

Graff, D. (2002). The AQUAINT Corpus of English News Text, Linguistic Data Consortium, Philadelphia

Greenlee, M. (2010). Youth and Gangs, in M. Coulthard and A. Johnson (Eds.). *The Routledge Handbook of Forensic Linguistics*, Routledge.

Jiang, M. & Argamon, S. (2008). Political leaning categorization by exploring subjectivities in political blogs. In *Proceedings of the 4th International Conference on Data Mining*, pages 647-653.

Johnsons, K. (2009). FBI: Burgeoning gangs behind up to 80% of U.S. Crime, in USA Today, January 29, 2009.

Kahneman, D. (2011). *Thinking Fast and Slow*, Farrar, Straus, and Giroux

Krippendorff, K. (2013). *Content Analysis: An Introduction to Its Methodology* (Chapter 13), SAGE Publications

Labov, W. (2010). *Principles of Linguistic Change: Internal Factors (Volume 1)*, Wiley-Blackwell.

Lefkowitz, N. (1989). Talking Backwards in French, *The French Review* 63(2), pp 312-322.

Mayfield, E. & Rosé, C. P. (2013). LightSIDE: Open Source Machine Learning for Text Accessible to Non-Experts, in *The Handbook of Automated Essay Grading*, Routledge Academic Press. http://lightsidelabs.com/research/

Philips, S. (2009). Crip Walk, Villian Dance, Pueblo Stroll: The Embodiment of Writing in African American Gang Dance, Anthropological Quarterly 82(1), pp69-97.

Schler, J., Koppel, M., Argamon, S., Pennebaker, J. (2005). Effects of Age and Gender on Blogging, Proceedings of AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs.

Schler, J. (2006). Effects of Age and Gender on Blogging. *Artificial Intelligence*, *86*, 82-84.

Wiebe, J., Bruce, R., Martin, M., Wilson, T., & Ball, M. (2004). Learning Subjective Language, *Computational Linguistics*, 30(3).

Yan, X., & Yan, L. (2006). Gender classification of weblog authors. *AAAI Spring Symposium Series Computational Approaches to Analyzing Weblogs* (p. 228–230).

Zhang, Y., Dang, Y., Chen, H. (2009). Gender Difference Analysis of Political Web Forums : An Experiment on International Islamic Women's Forum, Proceedings of the 2009 IEEE international conference on Intelligence and security informatics, pp 61-64.