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

Terra Blevins
Department of
Computer Science
Columbia University
New York, NY, USA
tlb2145@columbia.edu

Robert Kwiatkowski
Department of
Computer Science
Columbia University
New York, NY, USA
rjk2147@columbia.edu

Jamie Macbeth
Department of Electrical and
Computer Systems Engineering
Fairfield University
Fairfield, CT, USA
jmacbeth@fairfield.edu

Kathleen McKeown
Department of
Computer Science
Columbia University
New York, NY, USA
kathy@cs.columbia.edu

Desmond Patton
School of
Social Work
Columbia University
New York, NY, USA
dp2787@columbia.edu

Owen Rambow
Center for Computational
Learning Systems
Columbia University
New York, NY, USA
rambow@ccls.columbia.edu

Abstract

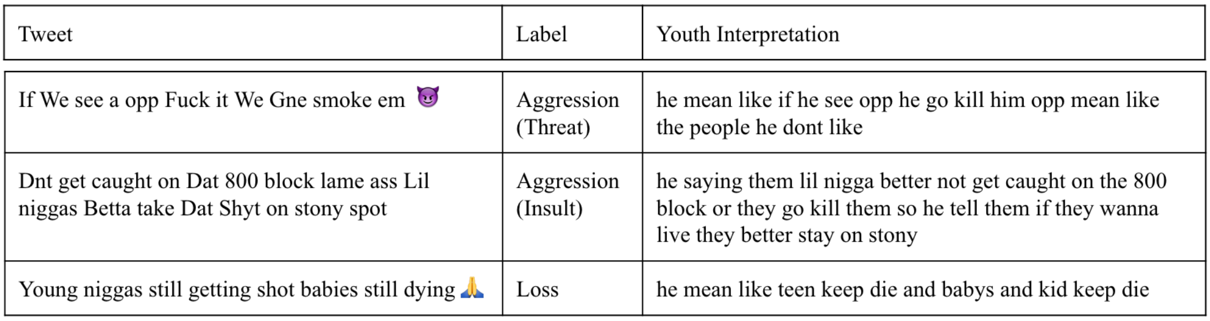
Violence is a serious problems for cities like Chicago and has been exacerbated by the use of social media by gang-involved youths for taunting rival gangs. We present a corpus of tweets from a young and powerful female gang member and her communicators, which we have annotated with discourse intention, using a deep read to understand how and what triggered conversations to escalate into aggression. We use this corpus to develop a part-of-speech tagger and phrase table for the variant of English that is used, as well as a classifier for identifying tweets that express grieving and aggression.

1 Introduction

The USA has the highest rate of firearm related deaths compared to other industrialized countries. Violence is particularly prevalent in cities like Chicago, which has seen a 40% increase in firearm violence in 2015; someone is shot every 2-3 hours in the city. The Chicago Police Department claims that gang violence is exacerbated by taunting between gang members on social media. Recent studies have shown that the new "digital street" is likely to have consequences for one's lived experiences (Moule et al., 2013; Patton et al., 2013; Pyrooz et al., 2015). Gangs that are highly organized have an increased likelihood of engaging in online behaviors that may include harassing others via the web (Moule et al., 2013).

In this paper, we work with a dataset of tweets posted by a young and particularly powerful female Chicago gang member, Gakirah Barnes, and people with whom she communicated. We use the dataset to develop a system that can automatically classify tweets as expressing either *loss*, grieving the death of friends or family who were shot, or *aggression*, threatening to harm others often in retribution for a loss. Tweets that don't fall into either category are classified as *other*. The ultimate goal of our work is to alert community outreach groups when aggressive tweets are identified so that they can intervene to alleviate a potentially violent situation. We are also interested in enabling interventions when youths are traumatized before grief turns to retribution. Our team includes social workers who labeled the data with discourse tags representing the intention behind the tweet, and computer scientists who developed the classification system in close consultation with the social workers.

The language used in the tweets is quite different from Standard American English and also from language used in Twitter by other populations. Sample tweets are shown in Figure 1, illustrating the many factors that characterize the form of these tweets: the use of dialectal (African American Vernacular English or AAVE) grammar and vocabulary (Rickford, 1999; Green, 2002), gang-related slang, non-standard orthography, emojis, and abbreviated expressions. Individual words do not always mean what



Tweet	Label	Youth Interpretation
If We see a opp Fuck it We Gne smoke em 😈	Aggression (Threat)	he mean like if he see opp he go kill him opp mean like the people he dont like
Dnt get caught on Dat 800 block lame ass Lil niggas Betta take Dat Shyt on stony spot	Aggression (Insult)	he saying them lil nigga better not get caught on the 800 block or they go kill them so he tell them if they wanna live they better stay on stony
Young niggas still getting shot babies still dying 🙏	Loss	he mean like teen keep die and babys and kid keep die

Figure 1: Tweets and a Chicago youth's interpretation of them.

they do in Standard American English. A Chicago youth from the neighborhood helped us interpret the tweets. His interpretations of the sample tweets are shown on the right in Figure 1.

Given this non-standard language, natural language tools that are widely used in the NLP community cannot be used for our task. Out-of-vocabulary words and abbreviated informal expressions mean that part-of-speech taggers are not accurate. Some words carry different meanings than in most other contexts (e.g., *smoke* in the first tweet of Figure 1 means 'kill') and thus even online slang dictionaries such as Wiktionary do not have accurate definitions for words in this context. In fact, 56.9% of the words in our corpus which are not in WordNet (Beckwith et al., 1991) have incorrect definitions in Wiktionary. The intuitions of the computer scientists on the team about the meaning of tweets was often incorrect and thus, interaction with the social workers was critical.

Our approach to classifying tweets features three key contributions:

- A new corpus that is annotated with discourse intention based on a deep read of the corpus, as well as POS tags.[1]
- NLP resources for the sub-language used by Chicago gang members, specifically a POS tagger and a glossary.
- A system to identify the emotion conveyed by tweets, using the Dictionary of Affect in Language.

We developed a part-of-speech (POS) tagger for the gang sublanguage and applied machine translation alignment to produce a phrase table that maps the vocabulary they use to Standard English. Features for the emotion classifier included the POS tags produced by our tagger as well as the Dictionary of Affect in Language's (DAL) quantitative scores representing the affect of words. (Whissell, 2009). In order to access the correct word in the DAL for each Twitter word, we used the glossary we derived to find the standard English terms corresponding to slang. Our supervised classifier is able to recognize *loss* tweets with 62.3% f-measure and *aggression* tweets with 63.6% f-measure, improving over the baseline by 13.7 points (aggression) and 5.8 points (loss).

In the following sections, we describe our annotated corpus, the sublanguage tools we developed and the classifier.

2 Related Work

Wijeratne et al. (2015) engineered a general surveillance platform that uses commonly available sentiment analysis tools as a component, but does not process the language of social media posts based on the specific language and culture of street gangs with an aim towards detecting aggression. Others have analyzed urban gangs' social media presence using spatialized network data (Radil et al., 2010) and automated the analysis of graffiti style features (Piergallini et al., 2014) to predict gang affiliation. Research has also studied the psychological impact of crime on urban populations by analyzing social media, finding that crime exposure over a year can result in negative emotion and anxiety (Valdes et al.,

[1] The dataset is available at http://dx.doi.org/10.7916/D84F1R07 .

2015). Yet others are analyzing news reports to build a database of gun violence incidents (Pavlick and Callison-Burch, 2016).

There has been work on POS tagging for Twitter (Derczynski et al., 2013; Owoputi et al., 2013), including for other languages (Rehbein, 2013). We discuss (Owoputi et al., 2013) in detail in Section 4.1 The most closely related work is (Jørgensen et al., 2016), which studies African American Vernacular English in three genres (movie scripts, lyrics, tweets). They use a very large unlabeled corpus. In contrast, we use a small labeled corpus and investigate domain adaptation using additional data. Given the short amount of time since the publication of (Jørgensen et al., 2016), we have not been able to obtain their data or system to compare to ours, which we intend to do in the future.

Other research has used statistical approaches to automatically characterize dialect variation in Twitter across cities and to show how the geographical distribution of lexical variation changes over time (Eisenstein, 2015). There has been quite a bit of work examining other kinds of phenomena on Twitter; researchers have developed systems to analyze accommodation (Danescu-Niculescu-Mizil et al., 2011), sentiment analysis (e.g., (Agarwal et al., 2011; Rosenthal et al., 2015)) and clues to geolocation (Dredze et al., 2016).

3 Our Corpus

3.1 Data Collection, Corpus, and Qualitative Analysis

To create our corpus, we analyzed publicially available Twitter communication from Gakirah Barnes, who became a gang member in Chicago at age 13 and was killed at age 17, as well as tweets from people who communicated with her. Barnes changed her Twitter handle to @TyquanAssassin in memory of her friend Tyquan Tyler, who was killed in 2012. She subsequently swore to avenge Tyler's death and became a known gang leader with 9 killings to her name before she was in turn shot and killed at age 17. We focus on Gakirah because she was highly active on Twitter, posting over 27,000 tweets from December, 2011 until her death on April 11, 2014. Her typical content ranged from discussing friends and intimate relationships to threats and taunts towards rival gangs and grieving the loss of friends killed due to gang or police violence. To start, we used Radian6[2], a social media tracking service, to capture several thousand tweets by, mentions of, and replies to @TyquanAssassin. We then applied a deep read to 718 of these tweets sent during a 34-day period starting on March 15th, 2014, two weeks before another of Gakirah's friends, Raason "Lil B" Shaw, was killed by the Chicago police (March 29th, 2014) and ending one week after Gakirah's death (Thursday, April 17th, 2014). A deep read is a type of textual analysis in which annotators use outside knowledge such as context to interpret textual data. They identify and describe subtle details of the tweet such as moments of escalation. (Patton et al., 2016) We selected this time period because it represents two violent events and the conditions for retaliation are feasible. Figure 1 shows three tweets from this period. We subsequently included 102 tweets from January 14th to January 20th of the same year in the analysis in order to create a test set.

Modeling a social work approach to conducting research, we created an interdisciplinary research team comprised of a social work researcher and computer scientists (Ford, 2014). The social workers developed the annotation categories based on work with two 18 year old African American men, from a Chicago neighborhood with high rates of violence, who we hired as research assistants. They were asked to interpret the 718 tweets from the 34-day period described above. The research assistants were provided an Excel spread sheet with Gakirah Twitter data which listed the author of the tweet, the content (excluding images), the URL to the specific Twitter page and the date and time with which the tweet was posted. They provided their initial reactions about the tweets including: their first impressions, general tone, emotion and explanation of language. They also interpreted emojis that were connected to text when they were able to access the URL for a specific tweet.

Next, the social work team used the Chicago youth interpretations to ensure they had an accurate understanding of the culture, context, and language embedded in the tweets. They then analyzed communications from Gakirah Barnes and other Twitter users in her network. The deep read analysis we developed was based on a coding process that related external events to expressed events. As part of the

[2] http://radian6.com

coding process, the social work team developed a codebook to reflect categories found in the data using a random sample of 50 Twitter communications from Gakirah and others in her Twitter network. The research team then used the codebook to code all 820 tweets. Given the context of the case study, (i.e. gang violence, aggression, and trauma), initial codes identified content in posts that were perceived as threatening or violent. We then focused on posts identified through our coding process and interpretations from youth research assistants as threatening or violent, and asked "why was this communicated?" To achieve this goal, we developed a 6 step process to understand how and what triggered conversations to escalate into aggression, a process we have termed the *Digital Urban Violence Analysis Approach*. These six steps include analyzing: a triggering event; the context about the author; the tweet content; information derived from the conversational network; the linguistic form and tone of the tweet; and finally, the next event or turning point. During this process we acquired a deeper understanding of the context surrounding the variation in Twitter communication. For example, we learned that aggressive and threatening communication was often times preceded by posts that reflected loss or grief. A total of 26 codes were developed through open coding which provided an explanation for why a threatening or violent post was communicated on Twitter. Critical to this process was the coding meetings or "member checking" where the social work team came together with the computer scientists to discuss the validity of codes. Chicago youth called in to discuss how they interpreted posts, the social work team described how they developed codes and identified emerging themes and the computer scientists often asked specifically about the qualitative coding process to better understand why certain text was coded as aggression or threat. A fuller description of the methodology can be found in (Patton et al., 2016).

Based on the coding meetings, we then engaged in a second round of coding, or selective coding, which was used to further examine why a category existed and to collapse the 26 codes further. We noticed that the majority of our codes fit into three broad categories: 1) aggression, 2) grief and 3) other. The collapsed aggression code contained examples of insults, threats, bragging, hypervigilance and challenges with authority. The collapsed grief code included examples of distress, sadness, loneliness and death. The "other" codes contained examples of general conversations between users, discussions about women, and tweets that represented happiness. The January data (the test set) was coded by two annotators; inter-annotator agreement on the test set is $\kappa = 0.62$, which is moderate agreement.

3.2 Data Used in Computational Experiments

The dataset used for our NLP experiments contains the 820 tweets from Gakirah and people with whom she communicated as just described. This data is partitioned into a training set of 616 tweets, a development set of 102 tweets, and a test set of 102 tweets. The training and development set come from March and April of 2014, and the test set consists of tweets from January of the same year.

We manually annotated this data set for part of speech (POS) tags. One annotator tagged the dev and train sets and another annotated the dev and test sets. Inter-annotator agreement was $\kappa = 0.80$ on the dev set. There is a large amount of domain specific language which our annotators frequently were unfamiliar with as well as many tweets with a variety of words used in a manner different from Standard English. One such example is the use of the word *ass* which at times can be used as an adverb, adjective, or noun, whereas in Standard English *ass* is almost always used exclusively as a noun. An example can be found in the second tweet in Figure 1, *lame ass Lil niggas*.[3] The first annotator interpreted *ass* as an intensifying modifier to the adjective *lame* and tagged it is an adverb, while the second annotator read it as the second in a string of three adjectives modifying *niggas*. Additionally, the noun phrase *stony spot* in the same tweet is read by the first annotator as a common noun phrase (an adjective modifying a noun) whereas the second annotator interpreted it as the name of a location and as such tagged both as proper nouns. These discrepancies and others like them lead to a difficult task, involving reconciling problems that do not exist in newswire data; for example, confusion between common and proper nouns account for 20% of the inter-annotator disagreement. Experiments to train a system to automatically produce

[3] Note that this is not the "Ass Camouflage Construction" (ACC) discussed by Collins et al. (2008) and others, in which a phrase of type *your ass* acts as a pronoun. Instead, this is an instance of the following unnamed construction: "An [AAVE] construction distinct from the ACC, one not common to standard colloquial English, involves the combination of adjectives or nouns with the nouns *ass* and *behind* to form complex adjectives only usable pre-nominally" (Collins et al., 2008, p.32).

POS annotation are described in Section 4.1.

For the classification experiments, we used the collapsed categories of *aggression* and *loss*. Tweets that do not have an aggression or loss label are grouped into a miscellaneous *other* group. We experimented with using the full data containing all three labels and a subset of the data containing only aggression and loss annotations; we describe these experiments in Section 5.

4 NLP Analysis for the Language of Twitter Posts

4.1 Part-of-Speech Tagging

Part of speech (POS) tagging is used as a source of features in many NLP classification tasks. Our data, being fairly different from most Standard English corpora, necessitated the creation of a tagger specific to this domain: the Stanford POS tagger trained on newswire achieves an accuracy of only 34.8% on our dev set (Table 1), and even the CMU Tweet-specific POS tagger (Owoputi et al., 2013) achieves only 81.5% (as compared to 91.5% on the CMU test set). We therefore hand-annotated our corpus with POS tags (see Section 3.2). We used the CMU tokenization scheme and tagset with minor changes. We tokenized the raw data using the CMU "twokenizer" for tweets, and then we performed a second tokenization step that splits all unicode emojis into individual emoji symbols separated by spaces. As our corpus had more acronyms and other miscellaneous words which CMU tags uniformly as "G" for garbage, we use the context of the word in the sentence to give it an appropriate tag (such as "N"). Furthermore, our data also includes many emojis as well as emoticons. CMU's tagset only had an "E" tag for emoticons but not for emojis; we tag all emojis with "E" as well. As such, our final tagset included all 25 tags of CMU tagset, with the exception of the "G" tag, resulting in 24 tags in the tagset. These differences caused an unfair decrease in accuracy for the CMU tagger which we want to use as a baseline in fair comparison; therefore, for evaluation of the CMU tagger we created a separate evaluation corpus on which to test CMU wherein all emojis were replaced with the emoticon ":)" and all "G" tags were not counted. Our own tagger was trained and tested on the unmodified data with all emojis preserved. A similar transformation was also necessary in converting the output from the Stanford Tagger due to the PTB tagset differing from the CMU tagset. The transformation is fairly straightforward as PTB tags have more detail than CMU tags. Additionally, because there are some CMU tags specific to Twitter language such as the "#" tag for hashtags or the "L" tag for words with contractions such as *I'm*, all such tags were not included in the accuracy rating for Stanford.

Tagger	Oct27 Test	Dev Set	Test Set
Stanford	52.2%	34.8%	26.0%
CMU	**91.5**%	81.5%	78.0%
Our Tagger	90.3%	**89.8**%	**81.5**%

Table 1: POS Tagger Accuracy Results

For features, we use word unigram and bigram features, the predicted tags from the previous two words ("Tags In Window"), character n-grams for the target word, and miscellaneous binary character features such as whether or not there was punctuation, capitalization, etc. in the word in question. Furthermore, our tagger also leverages Brown Clusters created by CMU for the task of POS tagging tweets.

We train our tagger on the entirety of the Oct27 CMU dataset containing around 1800 tweets as well as our manually annotated gang tweets training data (616 tweets – see Section 3.2). In order to leverage the similarity to standard tweets we made use of domain adaptation (Daumé III, 2007). We also tried an even simpler method: by adding an additional feature corresponding to the domain of the sentence in the training data as well as the domain of the sentence to be tagged, the classifier is able to effectively give a weighting to the value added by each of the domains when tagging the other. This simple method of domain adaptation performed slightly better than the Daumé method for this tagging task. In Table 2 we can see that the CMU data without domain adapatation adds 0.8%, and our simple domain adaptation adds another 0.9%.

The results on the CMU test set, our dev set, and our test set are shown in Table 1. The differences between our tagger and the CMU tagger on the dev and test sets are statistically significant ($p < 0.0001$, McNemar's test). There is a large difference between the dev and test set accuracy among all taggers,

Tagger	Dev Set
Our Tagger	**89.8**%
- Misc Char Features	89.7%
- Word Bigrams	89.5%
- Word Unigrams	89.1%
- Domain Adaptation	88.9%
- CMU data	88.1%
- Tags In Window	88.7%
- CMU Brown Clusters	88.0%
- Char n-grams	86.9%

Table 2: POS tagger feature ablation study: we show accuracy results when each listed feature is removed

due presumably to the difference in annotators. Accuracy for all three taggers is higher on the CMU test set than on the Gang dev and test sets as well. This is likely in part due to the very specific nature of the language in our tweets.

Table 2 shows an ablation study in which we remove one feature at a time. A lower result means that this feature contributes more. Surprisingly, the single most important feature is the character n-grams, followed by the CMU Brown Clusters. Because of the similarity of much of the vocabulary used, the Brown Clusters produce a reasonable increase in accuracy similar to the increase reported for CMU's tweet tagger. The CMU Brown Clusters had a hit rate of 93% for words in our corpus, excluding URLs, hashtags, user handles and emojis. The high hit rate, coupled with the fact that these clusters were derived from Twitter data, likely contributed to the value.

4.2 Extracting a Glossary

Another challenging NLP task involved the creation of a glossary for the gang tweets. Our method involved using the machine translation software Moses (Koehn et al., 2007). We glossed about 400 of the tweets from our corpus into Standard American English, and used MGIZA++ to extract an alignment. (We did not succeed in creating a phrase table, presumably because the corpus was too small.) From the alignments that Moses generated, we created a simplified phrasebook, mapping one gang tweet word to one or more English words. This approach was most effective in translating the many acronyms and abbreviations that exist in gang tweets.

5 Predicting Aggression in Twitter Posts

We experiment with three supervised classification systems to predict which tweets are aggressive or demonstrate loss. Two of our systems are Support Vector Machines (SVM) (Cortese and Vapnik, 1995); these experiments include ternary classification on the full dataset (TCF) and binary classification on the aggression-loss subset (BCS). Our TCF experiments include two binary classifiers, which classify tweets as aggression versus other and loss versus other; all tweets not classified as aggression or loss are labeled other. We also implemented an additional model, in an attempt to improve performance on the full dataset. This system is a cascading classifier (CC), which uses two SVM models. One model is trained to identify one class containing all aggression and loss tweets and a second class containing all other tweets using a binary classifier on the full dataset. This enables automatic generation of an aggression/loss subset. The tweets selected by the first SVM are then passed to a second model. This second model is the same model as the BCS for Loss and Aggression.

We compute features for these classifiers from our Twitter data, including unigrams, predicted POS tags, and emotion scores. For unigram features, Twitter handles are mapped to a common token, and URLs are handled similarly. Emojis behave as regular words for all features.

5.1 Emotion Features

Our approach to identifying aggression and loss in tweets depends on identifying the emotion expressed. We use the Dictionary of Affect in Language (DAL) in order to obtain the emotional content of individual words. The DAL is a lexicon that maps over 8000 English words to a three dimensional score. The three dimensions of this score are pleasantness; activation, which is a measure of a word's intensity; and imagery, which is a measure of the ease with which a word can be visualized. Our system extends the DAL with WordNet in order to identify the emotional content of Standard English words in our data that do not occur in the DAL following Rosenthal and McKeown (2013). For each word that is not found in the DAL and is found in WordNet, the synonyms from the first (most common) synset are searched against the DAL. We assume that the emotion of a synonym will be similar to that of the original word. Thus, if there is a match between the synonyms and the DAL, the emotion score of the synonym is used for the original word.

💯	HUNDRED POINTS SYMBOL
🙏	PERSON WITH FOLDED HANDS
😂	FACE WITH TEARS OF JOY
😈	SMILING FACE WITH HORNS
💸	MONEY WITH WINGS

Figure 2: The five most common emojis in our dataset and their unabbreviated descriptions.

A more difficult task is to apply the DAL to the nonstandard English and Twitter-specific elements of the tweets. We assume each token that is not found in the DAL or WordNet is not a Standard English word. We considered various lexicons for "translating" these tokens to standard English. One such lexicon is the phrasebook automatically generated through machine translation (Section 4.2). We also attempted to translate the tweets by using a larger knowledge base, Wiktionary. With Wiktionary, we considered the definition of a word to be its translation. Wiktionary contains an entry for about half (47.7%) of the nonstandard words in the tweets; however, due to the obscure nature of most of these words, only 45.1% of these definitions are correct. In comparison, the MT-generated phrasebook manages to identify a comparable number (43.6%) of the nonstandard words to Wikipedia. Additionally, the phrasebook is much more accurate on the words that it manages to translate than Wiktionary - 83.2% of the translations from the phrasebook are accurate. We thus use the MT-phrasebook we derived from the training data instead of Wiktionary as a translation lexicon in our final system.

We use a similar technique to obtain an emotion score for the emojis found in many of the tweets. Emojis are Unicode symbols that depict faces, animals, objects, and many other entities (Figure 2). They have recently become very popular in online communication, replacing the older "emoticon" (a facial expression depicted by punctuation, for example :)). Emojis are often used to contribute to or clarify the emotion of the words they accompany. Additionally, a significant number (12.6%) of non-stopword tokens in our data are emojis. Since emojis play a significant role in the overall emotional content of a tweet, it is imperative that we include the emotional content of these emojis when scoring the tweets for their overall emotion. We attempt to solve this problem by using an additional lexicon for emojis, which maps these symbols to a representative English word or phrase. Our Emoji Lexicon uses abbreviated versions of the Unicode "names," or informative glosses, that describe the symbol in words. Thus, similar to the process we use to translate nonstandard words and slang, we utilize this lexicon to obtain a English "translation" of each emoji we come across.

We obtained an emotion score for each tweet using these techniques. We preprocess the data, removing the stopwords and Twitter specific features that do not add emotional content, such as URLs and Twitter handles. For each nonstandard token, we search a translation lexicon (either the MT-generated phrasebook or the Emoji Lexicon) to obtain a Standard English translation. Once a translation is obtained for a nonstandard element, it is applied to the DAL system described above to obtain an emotion score. For words whose emotion scores are found directly in the DAL or through WordNet, the translation process is skipped. Once the three-dimensional emotion score of each individual word is identified, the scores are combined to represent the overall emotion of the tweet. A number of different methods of combining the emotion scores were tested; however, the best results were obtained by using, for each

Experiment	Label	Precision	Recall	F-measure
TCF	**Aggression**	0.525	0.600	0.560
	Baseline (unigrams)	0.462	0.514	0.486
TCF	**Loss**	0.500	0.625	0.556
	Baseline (unigrams)	0.500	0.688	0.578
TCF	Average of **Aggression** and **Loss**	0.513	0.613	0.558
TCF	**Aggression** or **Loss**	0.588	0.800	0.678
CC	**Aggression**	0.471	0.923	0.623
CC	**Loss**	0.483	0.933	0.636
CC	Average of **Aggression** and **Loss**	0.477	0.928	0.630
BCS	**Aggression**	0.868	0.943	0.904
	Baseline (unigrams)	0.906	0.829	0.866
BCS	**Loss**	0.750	0.938	0.833
	Baseline (unigrams)	0.813	0.813	0.813

Table 3: Experimental Results on the test set. TCF is a Ternary Classification on the Full dataset (the three classes being **Aggression**, **Loss**, and neither). We provide separate results for our two classes of interest, as well as the macro-average for the two classes. We also give results for a binary task in which we collapse **Aggression** and **Loss** into one class ("**TCF Aggression** or **Loss**"). CC is the Cascading Classifier whose first step is an identification of **Aggression** or **Loss** (the system in line labeled "**TCF Aggression** or **Loss**"), and whose second step is a binary classification on the positively identified data points from the first step using the BCS system. We again provide separate results for our two classes of interest and the macro-average. BCS is Binary Classification on the aggression-loss Subset of the training data.

dimension, the minimum and maximum scores across all words in the tweet.

5.2 Results

We experimented with different approaches to classifying our data according to the aggression, loss, and other categories. In addition to SVMs, we experimented with a number of ML approaches, but we found SVMs to work best for this task. The results are shown in Table 3, as well as the f-scores of baseline unigram models for each experiment.[4] The results are better on the aggression-loss subset (BCS) than on the full dataset (TCF). However, the aggression-loss subset does not represent real-world data, as all tweets that are not labeled loss or aggression were removed prior to the experiment. The Cascading Classifier (CC) was thus implemented in an attempt to achieve better results on a realistic dataset. The CC performs better than the unigram baselines and the TCF models on both the aggression and loss categories, with both improvements statistically significant (using randomization) at $p = 0.023$ for aggression and $p = 0.039$ for loss. For our task, the high recall of our system is beneficial; it ensures that all the tweets that could potentially escalate to violence are recommended to the user, so that they can decide whether or not to intervene.

We also report results on the development set and demonstrate the contributions made by the POS tags and emotion scores as features over the baseline with respect to the dev set (Table 4). Note that the POS tags were separated into two features: a unigram POS language model, and a bigram model. We only show results for those single features in combination with unigrams that improve over the baseline. The last line for each experiment/label pair represents the final feature set that was used by each experiment on the test set.

All of our features have an impact on our classifiers. Emotion scores are useful for classification on the aggression/loss subset of the data and for classification of the aggression label of the full dataset. POS tags are useful for almost all experiments with the exception of classification of the aggression label on

[4] The "other" category had a precision of 0.706, a recall of 0.462, and a 0.558% f-measure with the TCF classifier.

Experiment	Label	Features	F-measure
TCF	**Aggression**	unigrams (baseline)	0.609
		unigrams, bigrams	0.674
		unigrams, POS-unigrams	0.674
		unigrams, emotion score	0.659
		unigrams, bigrams, POS-unigrams, emotion score	0.741
TCF	**Loss**	unigrams (baseline)	0.756
		unigrams, POS-bigrams	0.818
TCF	**Aggression + Loss**	unigrams (baseline)	0.727
		unigrams, bigrams	0.738
		unigrams, POS-bigrams	0.812
		unigrams, bigrams, POS-bigrams	0.821
BCS	**Aggression**	unigrams (baseline)	0.866
		unigrams, bigrams	0.884
		unigrams, emotion score	0.914
		unigrams, bigrams, emotion score	0.926
BCS	**Loss**	unigrams (baseline)	0.708
		unigrams, POS-unigrams	0.766
		unigrams, emotion score	0.723
		unigrams, POS-unigrams, emotion score	0.800

Table 4: Results on the development set and a breakdown of impact of the feature sets. The first line given for each experiment and label is the unigram baseline, and the last line is the full feature set.

the subset and of the loss label on the full dataset. Since the subset models were used as part of the CC, the features for these models are also important to our cascading classifiers.

6 Conclusion

We have presented a new corpus of tweets written by young African Americans associated with gangs in Chicago. The tweets present a challenge to natural language processing since they exhibit many features that differentiate them from Standard American English and from a representative collection of English language tweets, and since they carry complex meaning in context. We have discussed a methodology which involves a close reading of tweets in conjunction with informants, and which leads to an annotation scheme for the tweets which interprets them in the social and communicative context. We have shown that we can use POS tagging at a reasonable level if we annotate a small corpus with POS tags. We have then used this POS tagger in conjunction with a glossary to develop a system that can tag tweets as expressing two categories from the annotation scheme, namely loss and aggression, with F-measures above 60% on our test set for both categories.

The work we describe in this paper is only a first step towards our goal of creating a tool that can alert social workers to the need to intervene, with the ultimate goal of reducing gang-related violence. In future work, we will extend our corpus to include more authors, more time periods, and greater geographical variation. We also intend to further investigate how close the relationship between expressions of aggression on Twitter and real world aggression is.

7 Acknowledgments

The research described here was supported in part by the National Science Foundation (NSF) under IIS-1422863. Any opinions, findings and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the NSF.

We thank the Social Media Listening Center at Clemson University for access to the Radian6 social media tracking platform. We also thank the anonymous reviewers for their thoughtful comments.

References

Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, and Rebecca Passonneau. 2011. Sentiment analysis of Twitter data. In *Proceedings of the Workshop on Language in Social Media (LSM 2011)*, pages 30–38, Portland, Oregon, June. Association for Computational Linguistics.

Richard Beckwith, Christiane Fellbaum, Derek Gross, and George Miller. 1991. WordNet: A lexical database organized on psycholinguistic principles. In Uri Zernik, editor, *Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon*, pages 211–232. Erlbaum.

Chris Collins, Simanique Moody, and Paul M. Postal. 2008. An AAE camouflage construction. *Language*, 84(1):29–68.

Corinna Cortese and Vladimir Vapnik. 1995. Support vector networks. *Machine Learning*, 20:273–297.

Cristian Danescu-Niculescu-Mizil, Michael Gamon, and Susan Dumais. 2011. Mark my words! Linguistic style accommodation in social media. In *Proceedings of WWW*, pages 745–754.

Hal Daumé III. 2007. Frustratingly easy domain adaptation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 256–263, Prague, Czech Republic, June. Association for Computational Linguistics.

Leon Derczynski, Alan Ritter, Sam Clark, and Kalina Bontcheva. 2013. Twitter part-of-speech tagging for all: Overcoming sparse and noisy data. In *RANLP*, pages 198–206.

Mark Dredze, Miles Osborne, and Prabhanjan Kambadur. 2016. Geolocation for Twitter: Timing matters. In *North American Chapter of the Association for Computational Linguistics (NAACL)*.

Jacob Eisenstein. 2015. Written dialect variation in online social media. In Charles Boberg, John Nerbonne, and Dom Watt, editors, *Handbook of Dialectology*. Wiley.

Heather Ford. 2014. Big data and small: Collaborations between ethnographers and data scientists. *Big Data & Society*, 1(2).

Lisa Green. 2002. *African American English: A Linguistic Introduction*. Cambridge University Press.

Anna Jørgensen, Dirk Hovy, and Anders Søgaard. 2016. Learning a POS tagger for AAVE-like language. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1115–1120, San Diego, California, June. Association for Computational Linguistics.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07, pages 177–180, Stroudsburg, PA, USA. Association for Computational Linguistics.

Richard K Moule, David C Pyrooz, and Scott H Decker. 2013. From 'What the f#@% is a Facebook?'to 'Who doesn't use Facebook?': The role of criminal lifestyles in the adoption and use of the Internet. *Social Science Research*, 42(6):1411–1421.

Olutobi Owoputi, Brendan O'Connor, Chris Dyer, Kevin Gimpel, Nathan Schneider, and Noah A. Smith. 2013. Improved part-of-speech tagging for online conversational text with word clusters. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 380–390, Atlanta, Georgia, June. Association for Computational Linguistics.

Desmond Upton Patton, Robert D Eschmann, and Dirk A Butler. 2013. Internet banging: New trends in social media, gang violence, masculinity and hip hop. *Computers in Human Behavior*, 29(5):A54 – A59.

Desmond Upton Patton, Kathleen McKeown, Owen Rambow, and Jamie Macbeth. 2016. Using natural language processing and qualitative analysis in gang violence: A collaboration between social work researchers and data scientists. In *Proceedings of Bloomberg Data for Good Exchange*.

Ellie Pavlick and Chris Callison-Burch. 2016. The gun violence database. In *Proceedings of Bloomberg Data for Good Exchange*.

Mario Piergallini, A Seza Dogruöz, Phani Gadde, David Adamson, and Carolyn Penstein Rosé. 2014. Modeling the use of graffiti style features to signal social relations within a multi-domain learning paradigm. pages 107–115.

David C. Pyrooz, Scott H. Decker, and Richard K. Moule Jr. 2015. Criminal and routine activities in online settings: Gangs, offenders, and the internet. *Justice Quarterly*, 32(3):471–499.

Steven M Radil, Colin Flint, and George E Tita. 2010. 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*, 100(2):307–326.

Ines Rehbein. 2013. Fine-grained POS tagging of German tweets. In *Language Processing and Knowledge in the Web*, pages 162–175. Springer.

John Russell Rickford. 1999. *African American vernacular English: Features, evolution, educational implications*. Wiley-Blackwell.

Sara Rosenthal and Kathleen McKeown. 2013. Sentiment detection of subjective phrases in social media. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 478–482, Atlanta, Georgia, June. Association for Computational Linguistics.

Sara Rosenthal, Preslav Nakov, Alan Ritter, Veselin Stoyanov, Svetlana Kiritchenko, and Saif Mohammad. 2015. Semeval-2015 task 10: Sentiment analysis in Twitter. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, Denver,CO.

Jose Manuel Delgado Valdes, Jacob Eisenstein, and Munmun De Choudhury. 2015. Psychological effects of urban crime gleaned from social media. In *Proceedings of the International Conference on Web and Social Media (ICWSM)*, Menlo Park, California, May. AAAI Press.

Cynthia Whissell. 2009. Using the revised dictionary of affect in language to quantify the emotional undertones of samples of natural language. *Psychological Reports*, 105:509–521.

Sanjaya Wijeratne, Derek Doran, Amit Sheth, and Jack L Dustin. 2015. Analyzing the social media footprint of street gangs. In *Intelligence and Security Informatics (ISI), 2015 IEEE International Conference on*, pages 91–96. IEEE.