Multimodal Social Media Analysis for Gang Violence Prevention

Philipp Blandfort,[1,2,*] Desmond Patton,[3] William R. Frey,[3] Svebor Karaman,[3] Surabhi Bhargava,[3] Fei-Tzin Lee,[3] Siddharth Varia,[3] Chris Kedzie,[3] Michael B. Gaskell,[3] Rossano Schifanella,[4] Kathleen McKeown,[3] Shih-Fu Chang[3]

[1] DFKI, Kaiserslautern, Germany

[2] TU Kaiserslautern, Kaiserslautern, Germany

[3] Columbia University, New York City, USA

[4] University of Turin, Turin, Italy

philipp.blandfort@dfki.de,{dp2787,w.frey,svebor.karaman,sb4019,fl2301,sv2504}@columbia.edu,kedzie@cs.columbia.edu,mbg2174@columbia.edu,schifane@di.unito.it,kathy@cs.columbia.edu,sc250@columbia.edu

ABSTRACT

Gang violence is a severe issue in major cities across the U.S. and recent studies [23] have found evidence of social media communications that can be linked to such violence in communities with high rates of exposure to gang activity. In this paper we partnered computer scientists with social work researchers, who have domain expertise in gang violence, to analyze how public tweets with images posted by youth who mention gang associations on Twitter can be leveraged to automatically detect psychosocial factors and conditions that could potentially assist social workers and violence outreach workers in prevention and early intervention programs. To this end, we developed a rigorous methodology for collecting and annotating tweets. We gathered 1,851 tweets and accompanying annotations related to visual concepts and the *psychosocial codes*: *aggression*, *loss*, and *substance use*. These codes are relevant to social work interventions, as they represent possible pathways to violence on social media. We compare various methods for classifying tweets into these three classes, using only the text of the tweet, only the image of the tweet, or both modalities as input to the classifier. In particular, we analyze the usefulness of mid-level visual concepts and the role of different modalities for this tweet classification task. Our experiments show that individually, text information dominates classification performance of the *loss* class, while image information dominates the *aggression* and *substance use* classes. Our multimodal approach provides a very promising improvement (18% relative in mean average precision) over the best single modality approach. Finally, we also illustrate the complexity of understanding social media data and elaborate on open challenges.

1 INTRODUCTION

Gun violence is a critical issue for many major cities. In 2016, Chicago saw a 58% surge in gun homicides and over 4,000 shooting victims, more than any other city comparable in size [13]. Recent data suggest that gun violence victims and perpetrators tend to have gang associations [13]. Notably, there were fewer homicides originating from physical altercations in 2016 than in the previous year, but we have little empirical evidence explaining why. Burgeoning social science research indicates that gang violence may be exacerbated by escalation on social media and the "digital street" [16] where exposure to aggressive and threatening text and images can lead to physical retaliation, a behavior known as "Internet banging" or "cyberbanging" [22].

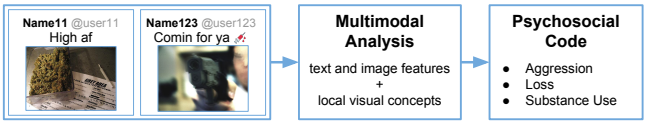

Figure 1: We propose a multimodal system for detecting psychosocial codes of social media tweets[1] related to gang violence.

Violence outreach workers present in these communities are thus attempting [21] to prioritize their outreach around contextual features in social media posts indicative of offline violence, and to try to intervene and de-escalate the situation when such features are observed. However, as most tweets do not explicitly contain features correlated with pathways of violence, an automatic or semi-automatic method that could flag a tweet as potentially relevant would lower the burden of this task. The automatic interpretation of tweets or other social media posts could therefore be very helpful in intervention, but quite challenging to implement for a number of reasons, e.g. the informal language, the African American Vernacular English, and the potential importance of context to the meaning of the post. In specific communities (e.g. communities with high rates of violence) it can be hard even for human outsiders to understand what is actually going on.

To address this challenge, we have undertaken a first multimodal step towards developing such a system that we illustrate in Figure 1. Our major contributions lie in innovative application of multimedia analysis of social media in practical social work study, specifically covering the following components:

- We have developed a rigorous framework to collect context-correlated tweets of gang-associated youth from Chicago containing images, and high-quality annotations for these tweets.
- We have teamed up computer scientists and social work researchers to define a set of visual concepts of interest.
- We have analyzed how the psychosocial codes *loss*, *aggression*, and *substance use* are expressed in tweets with images

[*] During some of this work Blandfort was staying at Columbia University.

[1] Note that the "tweets" in Figure 1 were created for illustrative purpose using Creative Commons images from Flickr and are NOT actual tweets from our corpus. Attributions of images in Figure 1, from left to right: "IMG_0032.JPG" by sashimikid, used under CC BY-NC-ND 2.0, "gun" by andrew_xjy, used under CC BY-NC-ND 2.0.

and developed methods to automatically detect these codes, demonstrating a significant performance gain of 18% by multimodal fusion.
- We have trained and evaluated detectors for the concepts and psychosocial codes, and analyzed the usefulness of the local visual concepts, as well as the relevance of image vs. text for the prediction of each code.

2 RELATED WORK

The City of Chicago is presently engaged in an attempt to use an algorithm to predict who is most likely to be involved in a shooting as either a victim or perpetrator [2]; however, this strategy has been widely criticized due to lack of transparency regarding the algorithm [30, 31] and the potential inclusion of variables that may be influenced by racial biases present in the criminal justice system (e.g. prior convictions) [1, 20].

In [9], Gerber uses statistical topic modeling on tweets that have geolocation to predict how likely 20 different types of crimes are to happen in individual cells of a grid that covers the city of Chicago. This work is a large scale approach for predicting future crime locations, while we detect codes in individual tweets related to future violence. Another important difference is that [9] is meant to assist criminal justice decision makers, whereas our efforts are community based and have solid grounding in social work research.

Within text classification, researchers have attempted to extract social events from web data including detecting police killings [14], incidents of gun violence [25], and protests [11]. However, these works primarily focus on extracting events from news articles and not on social media and have focused exclusively on the text, ignoring associated images.

The detection of local concepts in images has made tremendous progress in recent years, with recent detection methods [5, 10, 18, 28, 29] leveraging deep learning and efficient architecture enabling high quality and fast detections. These detection models are usually trained and evaluated on datasets such as the PascalVOC [8] dataset and more recently the MSCOCO [17] dataset. However, the classes defined in these datasets are for generic consumer applications and do not include the visual concepts specifically related to gang violence, defined in section 3.2. We therefore need to define a lexicon of gang-violence related concepts and train own detectors for our local concepts.

The most relevant prior work is that of [4]. They predict aggression and loss in the tweets of Gakirah Barnes and her top communicators using an extensive set of linguistic features, including mappings of African American vernacular English and emojis to entries in the Dictionary of Affective Language (DAL). The linguistic features are used in a linear SVM to make a 3-way classification between loss, aggression, and other. In this paper we additionally predict the presence of substance use, and model this problem as three binary classification problems since multiple codes may simultaneously apply. We also explore character and word level CNN classifiers, in addition to exploiting image features and their multimodal combinations.

3 DATASET

In this section we detail how we have gathered and annotated the data used in this work.

3.1 Obtaining Tweets

Working with community social workers, we identified a list of 200 unique users residing in Chicago neighborhoods with high rates of violence. These users all suggest on Twitter that they have a connection, affiliation, or engagement with a local Chicago gang or crew. All of our users were chosen based on their connections to a seed user, Gakirah Barnes, and her top 14 communicators in her Twitter network[2]. Gakirah was a self-identified gang member in Chicago, before her death in April, 2014. Additional users were collected using snowball sampling techniques [3]. Using the public Twitter API, in February 2017 we scraped all obtainable tweets from this list of 200 users. For each user we then removed all retweets, quote tweets and tweets without any image, limiting the number of remaining tweets per user to 20 to avoid most active users being overrepresented. In total the resulting dataset consists of 1,851 tweets from 173 users.

3.2 Local Visual Concepts

To extract relevant information in tweet images related to gang violence, we develop a specific lexicon consisting of important and unique visual concepts often present in tweet images in this domain. This concept list was defined through an iterative process involving discussions between computer scientists and social work researchers. We first manually went through numerous tweets with images and discussed our observations to find which kind of information could be valuable to detect, either for direct detection of "interesting" situations but also for extracting background information such as affiliation to a specific gang that can be visible from a tattoo. Based on these observations we formulated a preliminary list of visual concepts. We then collectively estimated utility (how useful is the extraction of the concept for gang violence prevention?), detectability (is the concept visible and discriminative enough for automatic detection?), and observability for reliable annotation (can we expect to obtain a sufficient number of annotations for the concept?), in order to refine this list of potential concepts and obtain the final lexicon.

Our interdisciplinary collaboration helped to minimize the risk of overseeing potentially important information or misinterpreting behaviors that are specific to this particular community. For example, on the images we frequently find people holding handguns with an extended clip and in many of these cases the guns are held at the clip only. The computer scientists of our team did not pay much attention to the extended clips and were slightly confused by this way of holding the guns, but then came to learn that in this community an extended clip counts as a sort of status symbol,

[2]Top communicators were statistically calculated by most mentions and replies to Gakirah Barnes.

[3]Attributions of Figure 2, from left to right: "GUNS" by djlindalovely, used under CC BY-NC-ND 2.0, "my sistah the art gangstah" by barbietron, used under CC BY-NC 2.0, "Money" by jollyuk, used under CC BY 2.0, "IMG_0032.JPG" by sashimikid, used under CC BY-NC-ND 2.0, "#codeine time" by amayzun, used under CC BY-NC-ND 2.0, "G Unit neck tattoo, gangs Trinidad" by bbcworldservice, used under CC BY-NC 2.0. Each image has been modified to show the bounding boxes of the local concepts of interest present in it.

(a) handgun, long gun (b) person, hand gesture (c) money (d) marijuana, joint (e) lean (f) person, tattoo

Figure 2: Examples of our gang-violence related visual concepts annotated on Creative Commons[3] images downloaded from Flickr.

hence this way of holding is meant to showcase a common status symbol. Such cross-disciplinary discussions lead to inclusion of concepts such as *tattoos* and separation of concepts to *handgun* and *long gun* in our concept lexicon.

From these discussions we have derived the following set of local concepts (in image) of interest:

- General: *person*, *money*
- Firearms: *handgun*, *long gun*
- Drugs: *lean*, *joint*, *marijuana*
- Gang affiliation: *hand gesture*, *tattoo*

This list was designed in such a way that after the training process described above, it could be further expanded (e.g. by specific hand gestures or actions with guns). We give examples of our local concepts in Figure 2.

3.3 Psychosocial Codes

Prior studies [4, 23] have identified *aggression*, *loss* and *substance use* as emergent themes in initial qualitative analysis that were associated with Internet banging, an emerging phenomenon of gang affiliates using social media to trade insults or make violence threats. Aggression was defined as posts of communication that included an insult, threat, mentions of physical violence, or plans for retaliation. Loss was defined as a response to grief, trauma or a mention of sadness, death, or incarceration of a friend or loved one. Substance use consists of mentions, and replies to images that discuss or show any substance (e.g. marijuana or a liquid substance colloquially referred to as "lean", see example in Figure 2) with the exception of cigarettes and alcohol.

The main goal of this work is to automatically detect a tweet that can be associated with any or multiple of these three psychosocial codes (*aggression*, *loss* and *substance use*) exploiting both textual and visual content.

3.4 Annotation

The commonly used annotation process based on crowd sourcing like Amazon Mechanical Turk is not suitable due to the special domain-specific context involved and the potentially serious privacy issues associated with the users and tweets.

Therefore, we adapted and modified the Digital Urban Violence Analysis Approach (DUVAA) [4, 24] for our project. DUVAA is a contextually-driven multi-step qualitative analysis and manual labeling process used for determining meaning in both text and images by interpreting both on- and offline contextual features. We adapted this process in two main ways. First, we include a step to uncover annotator bias through a baseline analysis of annotator perceptions of meaning. Second, the final labels by annotators undergo reconciliation and validation by domain experts living in Chicago neighborhoods with high rates of violence. Annotation is provided by trained social work student annotators and domain experts, community members who live in neighborhoods from which the Twitter data derives. Social work students are rigorously trained in textual and discourse analysis methods using the adapted and modified DUVAA method described above. Our domain experts consist of Black and Latino men and women who affiliate with Chicago-based violence prevention programs. While our domain experts leverage their community expertise to annotate the Twitter data, our social work annotators undergo a five stage training process to prepare them for eliciting context and nuance from the corpus.

We used the following tasks for annotation:

- In the *bounding box annotation task*, annotators are shown the text and tweet of the image. Annotators are asked to mark all local visual concepts of interest by drawing bounding boxes directly on the image. For each image we collected two annotations.
- To reconcile all conflicts between annotations we implemented a *bounding box reconciliation task* where conflicting annotations are shown side by side and the better annotation can be chosen by the third annotator.
- For *code annotation*, tweets including the text, image and link to the original post, are displayed and for each of the three codes *aggression*, *loss* and *substance use*, there is a checkbox the annotator is asked to check if the respective code applies to the tweet. We collected two student annotations and two domain expert annotations for each tweet. In addition, we created one extra code annotation to break ties for all tweets with any disagreement between the student annotations.

Our social work colleagues took several measures to ensure the quality of the resulting dataset during the annotation process. Annotators met weekly as a group with an expert annotator to address any challenges and answer any questions that came up that week. This process also involved iterative correction of reoccurring annotation mistakes and infusion of new community insights provided by domain experts. Before the meeting each week, the expert annotator closely reviewed each annotator's interpretations and labels to check for inaccuracies.

Concepts/Codes	Twitter	Tumblr	Total
handgun	164	41	205
long gun	15	105	116
joint	185	113	298
marijuana	56	154	210
person	1368	74	1442
tattoo	227	33	260
hand gesture	572	2	574
lean	43	116	159
money	107	138	245
aggression	457 (185)	-	457 (185)
loss	397 (308)	-	397 (308)
substance use	365 (268)	-	365 (268)

Table 1: Numbers of instances for the different visual concepts and psychosocial codes in our dataset. For the different codes, the first number indicates for how many tweets at least one annotator assigned the corresponding code, numbers in parentheses are based on per-tweet majority votes.

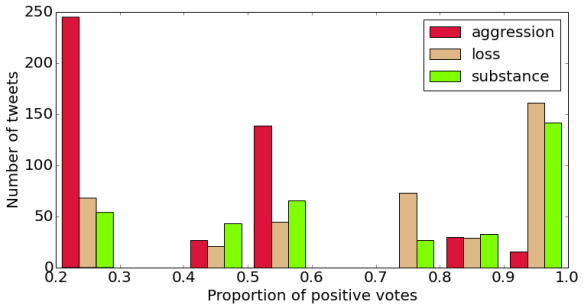

Figure 3: Annotator consensus for all psychosocial codes. For better visibility, we exclude tweets that were unanimously annotated as not belonging to the respective codes. Note that for each tweet there are 4 or 5 code annotations.

During the annotation process, we monitored statistics of the annotated concepts. This made us realize that for some visual concepts of interest, the number of expected instances in the final dataset was comparatively small.[4] Specifically, this affected the concepts *handgun*, *long gun*, *money*, *marijuana*, *joint*, and *lean*. For all of these concepts we crawled additional images from Tumblr, using the public Tumblr API with a keyword-based approach for the initial crawling. We then manually filtered the images we retrieved to obtain around 100 images for each of these specific concepts. Finally we put these images into our annotation system and annotated them w.r.t. all local visual concepts listed in Section 3.2.

3.5 Statistics

The distribution of concepts in our dataset is shown in Table 1. Note that in order to ensure sufficient quality of the annotations, but also due to the nature of the data, we relied on a special annotation process and kept the total size of the dataset comparatively small.

Figure 3 displays the distributions of fractions of positive votes for all 3 psychosocial codes. These statistics indicate that for the code *aggression*, disagreement between annotators is substantially higher than for the codes *loss* and *substance use*, which both display a similar pattern of rather high annotator consensus.

3.6 Ethical considerations

The users in our dataset comprise youth of color from marginalized communities in Chicago with high rates of gun violence. Releasing the data has the potential to further marginalize and harm the users who are already vulnerable to surveillance and criminalization by law enforcement. Thus, we will not be releasing the dataset used for this study. However, to support research reproducibility, we will release only the extracted linguistic and image features without revealing the raw content; this enables other researchers to continue research on training psychosocial code detection models without compromising the privacy of our users. Our social work team members initially attempted to seek informed consent, but to no avail, as participants did not respond to requests. To protect users, we altered text during any presentation so that tweets are not searchable on the Internet, excluded all users that were initially private or changed their status to private during the analysis, and consulted Chicago-based domain experts on annotation decisions, labels and dissemination of research.

[4]We were aiming for at least around 100-200 instances for training plus additional instances for testing.

4 METHODS FOR MULTIMODAL ANALYSIS

In this section we describe the building blocks for analysis, the text features and image features used as input for the psychosocial code classification with an SVM, and the multimodal fusion methods we explored. Details of implementation and analysis of results will be presented in Sections 5 and 6.

4.1 Text features

As text features, we exploit both sparse linguistic features as well as dense vector representations extracted from a CNN classifier operating at either the word or character level.

Linguistic features. To obtain the linguistic features, we used the feature extraction code of [4] from which we obtained the following:

- Unigram and bigram features.
- Part-of-Speech (POS) tagged unigram and bigram features. The POS tagger used to extract these features was adapted to this domain and cohort of users.
- The minimum and maximum pleasantness, activation, and imagery scores of the words in the input text. These scores are computed by looking up each word's associated scores in the Dictionary of Affective Language (DAL). Vernacular words and emojis were mapped to the Standard American English of the DAL using a translation phrasebook derived from this domain and cohort of users.

CNN features. To extract the CNN features we train binary classifiers for each code. We use the same architecture for both the word and character level models and so we describe only the word level model below. Our CNN architecture is roughly the same as [15] but with an extra fully connected layer before the final softmax. I.e.,

the text is represented as a sequence of embeddings, over which we run a series of varying width one-dimensional convolutions with max-pooling and a pointwise-nonlinearity; the resultant convolutional feature maps are concatenated and fed into a multi-layer perceptron (MLP) with one hidden layer and softmax output. After training the network, the softmax layer is discarded, and we take the hidden layer output in the MLP as the word or character feature vector to train the psychosocial code SVM.

4.2 Image features

We here describe how we extract visual features from the images that will be fed to the psychosocial code classifier.

Local visual concepts. To detect the local concepts defined in section 3.2, we adopt the Faster R-CNN model [29], a state-of-the-art method for object detection in images. The Faster R-CNN model introduced a *Region Proposal Network* (RPN) to produce region bounds and objectness score at each location of a regular grid. The bounding boxes proposed by the RPN are fed to a Fast R-CNN [10] detection network. The two networks share their convolutional features, enabling the whole Faster R-CNN model to be trained end-to-end and to produce fast yet accurate detections. Faster R-CNN has been shown [12] to be one of the best models among the modern convolutional object detectors in terms of accuracy. Details on the training of the model on our data are provided in Section 5.2. We explore the usefulness of the local visual concepts in two ways:

- For each *local visual concept* detected by the faster R-CNN, we count the frequency of the concept detected in a given image. For this, we only consider predictions of the local concept detector with a confidence higher than a given threshold, which is varied in experiments.
- In order to get a better idea of the potential usefulness of our proposed local visual concepts, we add one model to the experiments that uses *ground truth local concepts* as features. This corresponds to features from a perfect local visual concept detector. This method is considered out-of-competition and is not used for any fusion methods. It is used only to gain a deeper understanding of the relationship between the local visual concepts and the psychosocial codes.

Global features. As *global image features* we process the given images using a deep convolutional model (Inception-v3 [32]) pretrained on ImageNet [6] and use activations of the last layer before the classification layer as features. We decided not to update any weights of the network due to the limited size of our dataset and because such generic features have been shown to have a strong discriminative power [27].

4.3 Fusion methods for code detection

In addition to the text- and image-only models that can be obtained by using individually each feature described in Sections 4.1 and 4.2, we evaluate several tweet classification models that combine multiple kinds of features from either one or both modalities. These approaches always use features of all non-fusion methods for the respective modalities outlined in Sections 4.1 and 4.2, and combine information in one of the following two ways:

- *Early fusion*: the different kinds of features are concatenated into a single feature vector, which is then fed into the SVM. For example, the text-only early fusion model first extracts linguistic features and deploys a character and a word level CNN to compute two 100-dimensional representations of the text, and then feeds the concatenation of these three vectors into an SVM for classification.
- *Late fusion* corresponds to an ensemble approach. Here, we first train separate SVMs on the code classification task for each feature as input, and then train another final SVM to detect the psychosocial codes from the probability outputs of the previous SVMs.

5 EXPERIMENTS

Dividing by twitter users[5], we randomly split our dataset into 5 parts with similar code distributions and total numbers of tweets. We use these splits for 5-fold cross validation, i.e. all feature representations that can be trained and the psychosocial code prediction models are trained on 4 folds and tested on the unseen 5th fold. All reported performances and sensitivities are averaged across these 5 data splits. Statements on statistical significance are based on 95% confidence intervals computed from the 5 values on the 5 splits.

We first detail how the text and image representations are trained on our data. We then discuss the performance of different uni- and multimodal psychosocial code classifiers. The last two experiments are designed to provide additional insights into the nature of the code classification task and the usefulness of specific concepts.

5.1 Learning text representations

Linguistic features. We do not use all the linguistic features described in Section 4.1 as input for the SVM but instead during training apply feature selection using an ANOVA F-test that selects the top 1,300 most important features. Only the selected features are provided to the SVM for classification. We used the default SVM hyperparameter settings of [4].

CNN features. We initialize the word embeddings with pretrained 300-dimensional *word2vec* [19] embeddings.[6] For the character level model, we used 100-dimensional character embeddings randomly initialized by sampling uniformly from $(-0.25, 0.25)$. In both CNN models we used convolutional filter windows of size 1 to 5 with 100 feature maps each. The convolutional filters applied in this way can be thought of as word (or character) ngram feature detectors, making our models sensitve to chunks of one to five words (or characters) long. We use a 100-dimensional hidden layer in the MLP. During cross-validation we train the CNNs using the Nesterov Adam [7] optimizer with a learning rate of .002, early stopping on 10% of the training fold, and dropout of .5 applied to the embeddings and convolutional feature maps.

5.2 Learning to detect local concepts

Our local concepts detector is trained using the image data from Twitter and Tumblr and the corresponding bounding box annotations. We use the Twitter data splits defined above and similarly

[5]We chose to do the split on a user basis so that tweets of the same user are not repeated in both training and test sets.
[6]https://code.google.com/p/word2vec/

define five splits for the Tumblr data with similar distribution of concepts across different parts. We train a Faster R-CNN[7] model using a 5-fold cross validation, training using 4 splits of the Twitter and Tumblr data joined as a training set. We evaluate our local concepts detection model on the joined test set, as well as separately on the Twitter and Tumblr test set, and will discuss its performance in section 6.1.

The detector follows the network architecture of VGG-16 and is trained using the 4-step alternating training approach detailed in [29]. The network is initialized with an ImageNet-pretrained model and trained for the task of local concepts detection. We use an initial learning rate of 0.001 which is reduced by a factor of 0.9 every 30k iterations and trained the model for a total of 250k iterations. We use a momentum of 0.8 and a weight decay of 0.001.

During training, we augment the data by flipping images horizontally. In order to deal with class imbalance while training, we weigh the classification cross entropy loss for each class by the logarithm of the inverse of its proportion in the training data. We will discuss in detail the performance of our detector in Section 6.1.

5.3 Detecting psychosocial codes

We detect the three psychosocial codes separately, i.e. for each code we consider the binary classification task of deciding whether the code applies to a given tweet.

For our experiments we consider a tweet to belong to the positive class of a certain code if at least one annotator marked the tweet as displaying that code. For the negative class we used all tweets that were not marked by any annotator as belonging to the code (but might belong or not belong to any of the two other codes). We chose this way of converting multiple annotations to single binary labels because our final system is not meant to be used as a fully automatic detector but as a pre-filtering mechanism for tweets that are potentially useful for social workers. Given that the task of rating tweets with respect to such psychosocial codes inevitably depends on the perspective on the annotator to a certain extent, we think that even in case of a majority voting mechanism, important tweets might be missed.[8]

In addition to the models trained using the features described in Section 4, we also evaluate two baselines that do not process the actual tweet data in any way. Our *random baseline* uses the training data to calculate the prior probability of a sample belonging to the positive class and for each test sample predicts the positive class with this probability without using any information about the sample itself. The other baseline, *positive baseline*, always outputs the positive class.

All features except the linguistic features were fed to an SVM using the RBF kernel for classifying the psychosocial codes. For linguistic features, due to issues when training with an RBF kernel, we used a linear SVM with squared hinge loss, as in [4], and C = 0.01, 0.03 and 0.003 for detecting *aggression*, *loss* and *substance use* respectively. Class weight was set to balanced, with all other parameters kept at their default values. We used the SVM implementation of the Python library scikit-learn [26]. This two stage approach of feature extraction plus classifier was chosen to allow for a better understanding of the contributions of each feature. We preferred SVMs in the 2nd stage over deep learning methods since SVMs can be trained on comparatively smaller datasets without the need to optimize many hyperparameters.

For all models we report results with respect to the following metrics: precision, recall and F1-score (always on positive class), and average precision (using detector scores to rank output). The former 3 measures are useful to form an intuitive understanding of the performances, but for drawing all major conclusions we rely on average precision, which is an approximation of the area under the entire precision-recall curve, as compared to measurement at only one point.

The results of our experiments are shown in Table 2. Our results indicate that image and text features play different roles in detecting different psychosocial codes. Textual information clearly dominates the detection of code *loss*. We hypothesize that loss is better conveyed textually whereas substance use and aggression are easier to express visually. Qualitatively, the linguistic features with the highest magnitude weights (averaged over all training splits) in a linear SVM bear this out, with the top five features for loss being i) *free*, ii) *miss*, iii) *bro*, iv) *love* v) *you*; the top five features for substance use being i) *smoke*, ii) *cup*, iii) *drank*, iv) *@mention* v) *purple*; and the top five features for aggression being i) *Middle Finger Emoji*, ii) *Syringe Emoji*, iii) *opps*, iv) *pipe* v) *2017*. The loss features are obviously related to the death or incarceration of a loved one (e.g. *miss* and *free* are often used in phrases wishing someone was freed from prison). The top features for aggression and substance use are either emojis which are themselves pictographic representations, i.e. not a purely textual expression of the code, or words that reference physical objects (e.g. *pipe, smoke, cup*) which are relatively easy to picture.

Image information dominates classification of both the *aggression* and *substance use* codes. Global image features tend to outperform local concept features, but combining local concept features with global image features achieves the best image-based code classification performance. Importantly, by fusing both image and text features, the combined detector performs consistently very well for all three codes, with the mAP over three codes being 0.60, compared to 0.51 for the text only detector and 0.49 for the image only detector. This demonstrates a relative gain in mAP of around 20% of the multimodal approach over any single modality.

5.4 Sensitivity analysis

We performed additional experiments to get a better understanding of the usefulness of our local visual concepts for the code prediction task. For sensitivity analysis we trained linear SVMs on psychosocial code classification, using as features either the local visual concepts detected by Faster R-CNN or the ground truth visual concepts. All reported sensitivity scores are average values of the corresponding coefficients of the linear SVM, computed across the 5 folds used for the code detection experiments. Results from this experiment can be found in Table 3.

From classification using ground truth visual features we see that for detecting *aggression*, the local visual concepts *handgun*

[7] We use the publicly available implementation from: https://github.com/endernewton/tf-faster-rcnn

[8] For future work we are planning to have a closer look at the differences between annotations of community experts and students and based on that treat these types of annotations differently. We report a preliminary analysis in that direction in Section 6.2.

Modality	Features	Fusion	Aggression				Loss				Substance use				mAP
			P	R	F1	AP	P	R	F1	AP	P	R	F1	AP	
-	- (random baseline)	-	0.25	0.26	0.26	0.26	0.17	0.17	0.17	0.20	0.18	0.18	0.18	0.20	0.23
-	- (positive baseline)	-	0.25	1.00	0.40	0.25	0.21	1.00	0.35	0.22	0.20	1.00	0.33	0.20	0.22
text	linguistic features	-	0.35	0.34	0.34	0.31	0.71	0.47	0.56	0.51	0.25	0.53	0.34	0.24	0.35
text	CNN-char	-	0.37	0.47	0.39	0.36	0.75	0.66	0.70	**0.77**	0.27	0.32	0.29	0.28	0.45
text	CNN-word	-	0.39	0.46	0.42	0.41	0.71	0.65	0.68	**0.77**	0.28	0.30	0.29	0.31	0.50
text	all textual	early	0.40	0.46	0.43	0.42	0.70	0.73	0.71	**0.81**	0.25	0.37	0.30	0.30	0.51
text	all textual	late	0.43	0.41	0.42	0.42	0.69	0.65	0.67	**0.79**	0.29	0.37	0.32	0.32	0.51
image	inception global	-	0.43	0.64	0.51	**0.49**	0.38	0.57	0.45	0.43	0.41	0.62	0.49	**0.48**	0.47
image	Faster R-CNN local (0.1)	-	0.43	0.64	0.52	**0.47**	0.28	0.56	0.37	0.31	0.44	0.30	0.35	0.37	0.38
image	Faster R-CNN local (0.5)	-	0.47	0.48	0.47	0.44	0.30	0.39	0.33	0.31	0.46	0.12	0.19	0.30	0.35
image	all visual	early	0.49	0.62	0.55	**0.55***	0.38	0.57	0.45	0.44	0.41	0.59	0.48	**0.48**	0.49
image	all visual	late	0.48	0.51	0.49	**0.52**	0.40	0.51	0.44	0.43	0.47	0.52	0.50	**0.51***	0.49
image+text	all textual + visual	early	0.48	0.51	0.49	**0.53**	0.72	0.73	0.73	**0.82***	0.37	0.53	0.43	**0.45**	**0.60**
image+text	all textual + visual	late	0.48	0.44	0.46	**0.53**	0.71	0.67	0.69	**0.80**	0.44	0.43	0.43	**0.48**	**0.60***

Table 2: Results for detecting the psychosocial codes: aggression, loss and substance use. For each code we report precision (P), recall (R), F1-scores (F1) and average precision (AP). Numbers shown are mean values of 5-fold cross validation performances. The highest performance (based on AP) for each code is marked with an asterisk. In bold and red we highlight all performances not significantly worse than the highest one (based on statistical testing with 95% confidence intervals).

Concept	Aggression			Loss			Substance use		
	0.1	0.5	GT	0.1	0.5	GT	0.1	0.5	GT
handgun	0.73	0.93	1.05	0.06	0.10	0.06	0.06	0.09	0.11
long gun	0.26	0.91	1.30	-0.17	0.14	0.14	0.42	0.04	-0.47
joint	0.42	-0.08	0.05	-0.15	0.00	0.10	0.25	1.3	1.41
marijuana	0.17	0.18	0.12	-0.19	-0.45	-0.35	0.93	1.29	1.47
person	0.34	-0.01	-0.17	0.11	0.10	0.12	0.04	0.28	-0.01
tattoo	-0.11	-0.09	0.01	-0.02	0.03	-0.03	0.04	0.06	-0.02
hand gesture	0.20	0.67	0.53	-0.01	0.12	0.05	0.01	0.06	-0.02
lean	-0.07	0.03	-0.28	-0.20	-0.06	-0.14	0.68	0.59	1.46
money	-0.06	0.06	-0.02	0.00	-0.01	-0.01	0.18	-0.04	-0.19
F1	0.51	0.46	0.65	0.37	0.33	0.38	0.34	0.17	0.76
AP	0.41	0.39	0.54	0.29	0.28	0.30	0.33	0.27	0.72

Table 3: Sensitivity of visual local concept based classifiers w.r.t. the different concepts. For each of the three psychosocial codes, we include two versions that use detected local concepts ("0.1" and "0.5", where the number indicates the detection score threshold) and one version that uses local concept annotations as input ("GT").

and *long gun* are important, while for detecting *substance use*, the concepts *marijuana*, *lean*, *joint* are most significant. For the code *loss*, *marijuana* as the most relevant visual concept correlates negatively with *loss*, but overall, significance scores are much lower.

Interestingly, the model that uses the higher detection score threshold of 0.5 for the local visual concept detection behaves similarly to the model using ground truth annotations, even though the classification performance is better with the lower threshold. This could indicate that using a lower threshold makes the code classifier learn to exploit false alarms of the concept detector.

However, it needs to be mentioned that sensitivity analysis can only measure how much the respective classifier uses the different parts of the input, given the respective overall setting. This can give you useful information about which parts are *sufficient* for obtaining comparable detection results, but there is no guarantee that the respective parts are also *necessary* for achieving the same classification performance. [9]

For this reason, we ran an ablation study to get quantitative measurements on the necessity of local visual concepts for code classification.

5.5 Ablation study

In our ablation study we repeated the psychosocial code classification experiment using ground truth local visual concepts as features, excluding one concept at a time to check how this affects overall performance of the model.

We found that for *aggression*, removing the concepts *handgun* or *hand gesture* leads to the biggest drops in performance, while for *substance use*, the concepts *joint*, *marijuana* and *lean* are most important. For *loss*, removal of none of the concepts causes any significant change. See Table 4 for further details.

6 OPEN CHALLENGES

In this section, we provide a more in-depth analysis of what makes our problem especially challenging and how we plan to address those challenges in the future.

6.1 Local concepts analysis

We report in Table 5 the average precision results of our local concept detection approach on the "Complete" test set, i.e. joining data from both Twitter and Tumblr, and separately on the Twitter and Tumblr test sets. We compute the average precision on each test fold separately and report the average and standard deviation values

[9] For example, imagine that two hypothetical concepts A and B correlate perfectly with a given class and a detector for this class is given both concepts as input. The detector could make its decision based on A alone, but A is not really necessary since the same could be achieved by using B instead.

Removed concept	Aggression		Substance use	
	F1	AP	F1	AP
handgun	**-0.10**	**-0.15**	**-0.01**	0.01
long gun	-0.01	**-0.01**	-0.00	-0.00
joint	0.00	-0.00	**-0.35**	**-0.28**
marijuana	0.00	0.00	**-0.09**	**-0.09**
person	**-0.01**	-0.01	-0.01	-0.00
tattoo	0.00	0.00	0.01	-0.00
hand gesture	**-0.13**	**-0.09**	0.00	0.00
lean	-0.00	0.00	**-0.07**	**-0.07**
money	0.00	0.00	0.00	0.00

Table 4: Differences in psychosocial code detection performance of detectors with specific local concepts removed as compared to a detector that uses all local concept annotations. (Numbers less than 0 indicate that removing the concept reduces the corresponding score.) Bold font indicates that the respective number is significantly less than 0. For the code loss none of the numbers was significantly different from 0, hence we decided to not list them in this table.

Concept	Complete	Twitter	Tumblr
	AP ± SD	AP ± SD	AP ± SD
handgun	0.30 ± 0.07	0.13 ± 0.02	0.74 ± 0.11
long gun	0.78 ± 0.03	0.29 ± 0.41	0.85 ± 0.05
joint	0.30 ± 0.07	0.01 ± 0.01	0.57 ± 0.04
marijuana	0.73 ± 0.08	0.28 ± 0.17	0.87 ± 0.09
person	0.80 ± 0.03	0.80 ± 0.03	0.95 ± 0.03
tattoo	0.26 ± 0.06	0.08 ± 0.02	0.84 ± 0.06
hand gesture	0.27 ± 0.05	0.28 ± 0.04	0.83 ± 0.29
lean	0.78 ± 0.07	0.38 ± 0.15	0.87 ± 0.03
money	0.60 ± 0.02	0.35 ± 0.08	0.73 ± 0.05
mAP	0.54 ± 0.01	0.29 ± 0.05	0.81 ± 0.02

Table 5: Local concepts detection performance.

over the 5 folds. When looking at the results on the "Complete" test set, we see average precision values ranging from 0.26 on *tattoo* to 0.80 for *person* and the mean average precision of 0.54 indicating a rather good performance. This results on the "Complete" test set hides two different stories, however, as the performance is much lower on the Twitter test set (mAP of 0.29) than on the Tumblr one (mAP of 0.81).

As detailed in Section 3.4, we have crawled additional images, especially targeting the concepts with a low occurrence count in Twitter data as detailed in Table 1. However, crawling images from Tumblr targeting keywords related to those concepts lead us to gather images where the target concept is the main subject in the image, while in our Twitter images they appear in the image but are rarely the main element in the picture. Further manually analyzing the images crawled from Twitter and Tumblr, we have confirmed this "domain gap" between the two sources of data that can explain the difference of performance. This puts in light the challenges associated with detecting these concepts in our Twitter data. We believe the only solution is therefore to gather additional images from Twitter from similar users. This will be part of the future work of this research.

The local concepts are highly relevant for the detection of the codes *aggression* and *substance use* as it can be highlighted in the column GT in Table 3 and from the ablation study reported in Table 4. The aforementioned analysis of the local concepts detection limitation on the Twitter data explains why the performance using the detected concepts is substantially lower than when using ground truth local concepts. We will therefore continue to work on local concepts detection in the future as we see they could provide significant help in detecting these two codes and also because they would help in providing a clear interpretability of our model.

6.2 Annotation analysis

In order to identify factors that led to divergent classification between social work annotators and domain experts, we reviewed 10% of disagreed-upon tweets with domain experts. In general, knowledge of local people, places, and behaviors accounted for the majority of disagreements. In particular, recognizing and having knowledge of someone in the image (including their reputation, gang affiliation, and whether or not they had been killed or incarcerated) was the most common reason for disagreement between our annotators and domain experts. Less commonly, identifying or recognizing physical items or locations related to the specific cultural context of the Chicago area (e.g., a home known to be used in the sale of drugs) also contributed to disagreement. The domain experts' nuanced understanding of hand signs also led to a more refined understanding of the images, which variably increased or decreased the perceived level of aggression. For example, knowledge that a certain hand sign is used to disrespect a specific gang often resulted in increased perceived level of aggression. In contrast, certain hand gestures considered to be disrespectful by our social work student annotators (e.g., displaying upturned middle fingers) were perceived to be neutral by domain experts and therefore not aggressive. Therefore, continuous exchange with the domain experts is needed to always ensure that the computer scientists are aware of all these aspects when further developing their methods.

6.3 Ethical implications

Our team was approached by violence outreach workers in Chicago to begin to create a computational system that would enhance violence prevention and intervention. Accordingly, our automatic vision and textual detection tools were created to assist social workers in their efforts to understand and prevent community violence through social media, but not to optimize any systems of surveillance. This shift away from identifying potentially violent users to understanding pathways to violent online content highlights systemic gaps in economic, educational, and health-related resources that are often root causes to violent behavior. Our efforts for ethical and just treatment of the users who provide our data include encryption of all Twitter data, removal of identifying information during presentation of work (e.g., altering text to eliminate searchability), and the inclusion of Chicago-based community members as domain experts in the analysis and validation of our findings. Our long term efforts include using multimodal analysis to enhance

current violence prevention efforts by providing insight into social media behaviors that may shape future physical altercations.

7 CONCLUSION

We have introduced the problem of multimodal social media analysis for gang violence prevention and presented a number of automatic detection experiments to gain insights into the expression of *aggression*, *loss* and *substance use* in tweets coming from this specific community, measure the performance of state-of-the-art methods on detecting these codes in tweets that include images, and analyze the role of the two modalities text and image in this multimodal tweet classification setting.

We proposed a list of general-purpose local visual concepts and showed that despite insufficient performance of current local concept detection, when combined with global visual features, these concepts can help visual detection of *aggression* and *substance use* in tweets. In this context we also analyzed in-depth the contribution of all individual concepts.

In general, we found the relevance of the text and image modalities in tweet classification to depend heavily on the specific code being detected, and demonstrated that combining both modalities leads to a significant improvement of overall performance across all 3 psychosocial codes.

Findings from our experiments affirm prior social science research indicating that youth use social media to respond to, cope with, and discuss their exposure to violence. Human annotation, however, remains an important element in vision detection in order to understand the culture, context and nuance embedded in each image. Hence, despite promising detection results, we argue that psychosocial code classification is far from being solved by automatic methods. Here our interdisciplinary approach clearly helped to become aware of the whole complexity of the task, but also to see the broader context of our work, including important ethical implications which were discussed above.

ACKNOWLEDGMENTS

During his stay at CU, the first author was supported by a fellowship within the FITweltweit programme of the German Academic Exchange Service (DAAD). Furthermore, we thank all our annotators: Allison Aguilar, Rebecca Carlson, Natalie Hession, Chloe Martin, Mirinda Morency.

REFERENCES

[1] 2017. Click, Pre Crime, Chicago's crime-predicting software. (May 2017). http://www.bbc.co.uk/programmes/p052fll7

[2] 2017. Strategic Subject List | City of Chicago | Data Portal. (2017). https://data.cityofchicago.org/Public-Safety/Strategic-Subject-List/4aki-r3np Online; Accessed April 2018.

[3] R Atkinson and J Flint. 2001. Accessing Hidden and Hard-to-reach Populations: Snowball Research Strategies. *Social Research Update* 33 (2001). http://eprints.gla.ac.uk/37493/

[4] Terra Blevins, Robert Kwiatkowski, Jamie Macbeth, Kathleen McKeown, Desmond Patton, and Owen Rambow. 2016. Automatically processing Tweets from gang-involved youth: Towards detecting loss and aggression. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. 2196–2206.

[5] Jifeng Dai, Yi Li, Kaiming He, and Jian Sun. 2016. R-fcn: Object detection via region-based fully convolutional networks. In *Advances in neural information processing systems*. 379–387.

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. 2009. ImageNet: A large-scale hierarchical image database.. In *CVPR*. IEEE Computer Society, 248–255.

[7] Timothy Dozat. 2016. Incorporating nesterov momentum into adam. (2016).

[8] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. 2010. The Pascal Visual Object Classes (VOC) Challenge. *International Journal of Computer Vision* 88, 2 (June 2010), 303–338.

[9] Matthew S. Gerber. 2014. Predicting crime using Twitter and kernel density estimation. *Decision Support Systems* 61 (2014), 115 – 125. DOI : http://dx.doi.org/https://doi.org/10.1016/j.dss.2014.02.003

[10] Ross Girshick. 2015. Fast R-CNN. In *Computer Vision (ICCV), 2015 IEEE International Conference on*. IEEE, 1440–1448.

[11] Alex Hanna. 2017. MPEDS: Automating the Generation of Protest Event Data. (2017).

[12] Jonathan Huang, Vivek Rathod, Chen Sun, Menglong Zhu, Anoop Korattikara, Alireza Fathi, Ian Fischer, Zbigniew Wojna, Yang Song, Sergio Guadarrama, and others. 2017. Speed/accuracy trade-offs for modern convolutional object detectors. In *IEEE CVPR*.

[13] Max Kapustin, Jens Ludwig, Marc Punkay, Kimberley Smith, Lauren Speigel, and David Welgus. 2017. Gun Violence in Chicago, 2016. *Chicago, IL: University of Chicago Crime Lab* (2017).

[14] Katherine A Keith, Abram Handler, Michael Pinkham, Cara Magliozzi, Joshua McDuffie, and Brendan O'Connor. 2017. Identifying civilians killed by police with distantly supervised entity-event extraction. *arXiv preprint arXiv:1707.07086* (2017).

[15] Yoon Kim. 2014. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882* (2014).

[16] Jeffrey Lane. 2016. The digital street: An ethnographic study of networked street life in Harlem. *American Behavioral Scientist* 60, 1 (2016), 43–58.

[17] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*. Springer, 740–755.

[18] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. 2016. Ssd: Single shot multibox detector. In *European conference on computer vision*. Springer, 21–37.

[19] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed Representations of Words and Phrases and Their Compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2 (NIPS'13)*. Curran Associates Inc., USA, 3111–3119. http://dl.acm.org/citation.cfm?id=2999792.2999959

[20] A. Nellis, J.A. Greene, M. Mauer, and Sentencing Project (U.S.). 2008. *Reducing Racial Disparity in the Criminal Justice System: A Manual for Practitioners and Policymakers*. Sentencing Project. https://books.google.com/books?id=MQKznQEACAAJ

[21] Citizens Crime Commission of New York City. 2017. E-Responder: a brief about preventing real world violence using digital intervention. (2017). www.nycrimecommission.org/pdfs/e-responder-brief-1.pdf

[22] 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 (2013), A54–A59.

[23] Desmond U Patton, Jeffrey Lane, Patrick Leonard, Jamie Macbeth, and Jocelyn R Smith Lee. 2017. Gang violence on the digital street: Case study of a South Side Chicago gang member's Twitter communication. *new media & society* 19, 7 (2017), 1000–1018.

[24] Desmond Upton Patton, Kathleen McKeown, Owen Rambow, and Jamie Macbeth. 2016. Using Natural Language Processing and Qualitative Analysis to Intervene in Gang Violence: A Collaboration Between Social Work Researchers and Data Scientists. *arXiv preprint arXiv:1609.08779* (2016).

[25] Ellie Pavlick, Heng Ji, Xiaoman Pan, and Chris Callison-Burch. 2016. The Gun Violence Database: A new task and data set for NLP. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. 1018–1024.

[26] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[27] Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. 2014. CNN features off-the-shelf: an astounding baseline for recognition. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2014 IEEE Conference on*. IEEE, 512–519.

[28] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. 2016. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 779–788.

[29] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2017. Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence* 39, 6 (2017), 1137–1149.

[30] Christine Schmidt. 2018. Holding algorithms (and the people behind them) accountable is still tricky, but doable. (Mar 2018). http://nie.mn/2ucDuw2 Online; Accessed April 2018.

[31] Karen Sheley. 2017. Statement on Predictive Policing in Chicago. (Jun 2017). http://www.aclu-il.org/en/press-releases/statement-predictive-policing-chicago
[32] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. 2016. Rethinking the Inception Architecture for Computer Vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. 2818–2826. DOI: http://dx.doi.org/10.1109/CVPR.2016.308