Content-Driven Detection of Cyberbullying on the Instagram Social Network

Haoti Zhong, Hao Li
Dept. of Electrical Eng.
Pennsylvania State University
hzz133@psu.edu, hul175@psu.edu

Anna Squicciarini
Information Sciences and Technology
Pennsylvania State University
acs20@psu.edu

Sarah Rajtmajer
Dept. of Mathematics
Pennsylvania State University
smr48@psu.edu

Christopher Griffin
Dept. of Mathematics
United States Naval Academy
griffinch@ieee.org

David Miller
Dept. of Electrical Engineering
Pennsylvania State University
djmiller@engr.psu.edu

Cornelia Caragea
Dept. of Computer Science
University of North Texas
ccaragea@unt.edu

Abstract

We study detection of cyberbullying in photo-sharing networks, with an eye on developing early-warning mechanisms for the prediction of posted images vulnerable to attacks. Given the overwhelming increase in media accompanying text in online social networks, we investigate use of posted images and captions for improved detection of bullying in response to shared content. We validate our approaches on a dataset of over 3000 images along with peer-generated comments posted on the Instagram photo-sharing network, running comprehensive experiments using a variety of classifiers and feature sets. In addition to standard image and text features, we leverage several novel features including topics determined from image captions and a pretrained convolutional neural network on image pixels. We identify the importance of these advanced features in assisting detection of cyberbullying in posted comments. We also provide results on classification of images and captions themselves as potential targets for cyberbullies.

1 Introduction

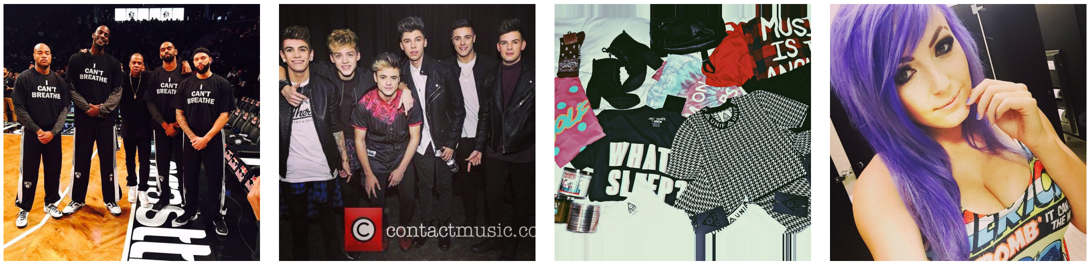
A growing body of research into cyberbullying in online social networks has been catalyzed by increasing prevalence and deepening consequences of this type of abuse. To date, automated detection of cyberbullying has focused on analyses of text in which bullying is suspected to be present. However, given the increase in media accompanying text in online social networks, an increasing number of cyberbullying incidents are linked with photos and media content, which are often used as targets for harassment and stalking.

For instance, in Instagram, a highly popular online photo-sharing platform, bullying is becoming a serious concern. Recent statistics indicate that anywhere between 9% and 25% of users claim to have been bullied on Instagram, with the problem even more prevalent on Twitter and Facebook [Cyberbullying Research Center, 2016]. Considering the pervasiveness and danger increasingly represented by bullying online, bully detection is of interest to a cross-sectional community of social and computer scientists. In particular, detecting instances of cyberbullying through analysis of media content is an important and challenging task, as the connection between a bullied image and its context is unclear. Yet, insight into the characteristics of shared content may prove extremely useful in eventual development of warning mechanisms designed to prevent cyberbullying.

In this work, we develop methods for detecting cyberbullying in commentaries following shared images on Instagram. In addition to image-specific and text features extracted from comments and from image captions, we leverage several novel features including topics determined from image captions and outputs of a pretrained convolutional neural network applied to image pixels. We identify the importance of these advanced features in detecting occurrences of cyberbullying in posted comments. We also provide results on classification of images and captions themselves as potential targets for cyberbullies. Leveraging features of the posted images and captions as well as the comments themselves, we are able to classify comments that contain bullying with over 93% accuracy. Moreover, we lay the foundation for identifying posted content which may be particularly vulnerable to bullying, noting the difficulty of the problem space and suggesting pointers for next steps. Using a pre-trained convolutional neural network and topics generated from image captions we provide the first meaningful attempt to anticipate instances of bullying in response to posted images, achieving 68.55% overall accuracy.

2 Related Work

A body of work is emerging around the problem of cyberbullying, from various disciplines. Recent papers in developmental psychology and sociology have characterised the profiles and motivations of offenders, and have discussed possible strategies of prevention and intervention [Berson *et al.*, 2002; Hinduja and Patchin, 2013]. Of note, these studies



(a) Cyberbullying (b) Cyberbullying (c) No cyberbullying (d) No cyberbullying

Figure 1: Example of images subject to cyberbullying and not subject to cyberbullying.

highlight the influence of both peers and authorities on encouraging or mitigating cyberbullying behaviors. These facts motivate development of novel approaches to automated detection of cyberbullying in online social networks. Intervention will require identification of instantiations of the problem and, ideally, may follow from early warning mechanisms when particularly vulnerable content is posted.

Within computer science, researchers have developed methods to automatically detect cyberbullying, mostly focusing on text mining (e.g. [Yin *et al.*, 2009; Dinakar *et al.*, 2011; Kontostathis *et al.*, 2013; Chen *et al.*, 2012]). Framing the problem slightly differently, others [Dadvar *et al.*, 2013] have aimed to detect the cyberbullies themselves, leveraging additional user features (e.g., geoposition) as well as hybrid machine learning/expert systems. In existing approaches, little (if any) attention is paid to context such as the victims' profiles, targeted posted content and the nature of users' interactions, which may all be crucial in triggering and fostering bullying behavior [Sabella *et al.*, 2013; Berson *et al.*, 2002; Hinduja and Patchin, 2013]. [Yin *et al.*, 2009] is the most similar to our work in that they take a supervised learning approach to detect cyberbullying using content and sentiment features, as well as contextual features of the considered documents.Authors define context by two metrics, both assessing the similarity of a given post to the other posts in its immediate vicinity. Our work is distinct in several ways. Because we address cyberbullying of images in online social networks, our context is provided by features of the image itself, posted captions, and the posting user. We combine analysis of the text potentially containing abuse with these contextual features, using a combination of supervised and unsupervised learning approaches.

3 Problem Statement

We explore the relationships between text and visual content with respect to cyberbullying. Specifically, we aim to understand whether there is a correlation between shared media in the form of posted images and captions, and the occurrence of cyberbullying events. The strength of such relationships and their ability to inform possible future instances of cyberbullying are at the core of our problem. This work is motivated by the following two primary questions.

- Classic natural language processing techniques have been shown to work well for post-hoc detection of bullying in text. Can we further increase the accuracy in detecting bullying of shared images in the Instagram social network by leveraging contextual clues such as images features, image caption, and user metadata, including the number of follows/-ers
- Is it possible to anticipate instances of cyberbullying on a piece of shared content based on some combination of contextual features, i.e., features of the posted image itself, together with the caption and user metadata?

We hypothesize that the answers to these two questions are related and affirmative. That is, we aim to detect whether certain classes of images may be considered more controversial, and whether these images and corresponding captions may more readily incite attacks. We hypothesize that leveraging these data should both aid in identifying the presence of cyberbullying in posted comments and the prediction of this kind of abuse given its context.

Our problem is non-trivial and presents unique challenges. A clear pattern binding users' comments and the original thread may not exist. Figure 1 reports examples of representative images in our Instagram dataset illustrating this point (a detailed description of the dataset is provided in Section 4). Bullying comments were found in the commentary following Figure 1(a-b). One might expect this in photo (a) for several reasons. The basketball players pictured are celebrities, inherently prone to scrutiny. In this photo, they are wearing t-shirts to make comment on a controversial social issue. The bullying of (b) is more surprising. No individual or individual behavior is highlighted as a potential target for abuse in this seemingly harmless shot. However, the caption in this photo gives insight into why it was vulnerable to attack: `"The lads before our G.A.Y. gig last night :D ".` Referencing the sexual orientation of the boys in the photo, the caption provides a more risque context than evident from the image alone. Likewise, the two images in Figure 1(c-d) are examples of pictures which were not bullied, as expected for (c) but more surprisingly for (d), which features a single individual and bare skin.

4 The Instagram Dataset

To generate a reliable dataset, we crawled publicly visible accounts on the popular Instagram social platform through the site's official API. Instagram affords a unique opportunity to collect posted images, captions and corresponding commentaries within a clearly-defined, directed network of followers/followees. Additionally, the popularity of Instagram and the known presence of cyberbullying on the network [Cyberbullying Research Center, 2016] make it a convincing choice for validation.

We collected 9000 images. In order to obtain contextual information about users' activities and profiles, along with each image, we collected the user-created image caption, specific information about the user who posted the content (username, total post count, number of followees and number of followers), and the text of the 150 most recently-posted comments (or fewer, in cases where the total number of comments for an image was less than 150). In all, we obtained approximately 500,000 comments. We note that our images and corresponding metadata were selected randomly from a list of popular images on the site at the time of the crawl.

The dataset was trimmed to 3000 images by removing images with non-English language comments and preserving from this subset the set of images having the greatest number of comments. These images were then labeled in two different iterations, using Mechanical Turk workers. First, we presented both images and comments, and asked labelers to identify whether the image was bullied based on the image's commentary. Next, labelers were asked to label each comment individually as either bullying or non-bullying.

- *Image labeling:* Images were presented to labelers with their corresponding comments. Labelers were asked to look at the image, read through the comments and answer two multiple-choice questions. First, we asked whether the comments included any bullying, and second, in case an instance of bullying was present, we asked whether that bullying seemed to be due to the content of the image. Each image with comments was presented to three distinct labelers, and we considered an image as having been bullied if 2 or 3 labelers responded affirmatively to either one or both questions. All other images were labeled non-bullied. In total, 560 images were considered bullied and 2540 were not. Among those bullied, 19.2% were said to be bullied due to the controversial nature of the image, 21.13% due to the appearance of the subjects of the image, 3% because of the private nature of the image, while the remainder were said to be targeted for "other" reasons (e.g., popularity of the posting user, subjects of the image).
- *Comment Labeling:* We asked users to label a subset of the comments, 30 comments each taken from 1120 images. Labelers had access to the image, the image's commentary, and indicated whether or not each comment represented bullying.

5 Feature Vector Construction

To investigate our first research question, we attempted classification of bullying based both on comments, image content, and contextual features. As discussed in Figures 1(a-d), there are certain images that *seem* more likely to be bullied, but it is hard to capture exactly what it is about those photos which make them so. It may be something explicitly identifiable in the content of the image, or it may be contextual. It would be extremely difficult for any machine learning algorithm to deduce the provocative nature of Figure 1(a), given the text on the players' t-shirts and its reference to the highly public death of Eric Garner. We suggest that leveraging a combination of text-based, image-based and meta- features will provide the strongest predictive power in this context. In the sequel, we discuss a variety of these features, both basic and sophisticated, classic and novel and evaluate their predictive power on our Instagram dataset.

5.1 Feature Set for Comments on Posted Content

A set of features was generated for each comment in the dataset. In preprocessing, comments were first cleared of "@" mentions, non-Enligsh words and emojis in order to permit their submission to the feature-generating processes described below.

Bag of Words To capture the main topics and jargon used in the commentaries, we analyzed word frequency, using a Bag of Words model. The "Bag of words" model (BoW) [Harris, 1954] is a baseline text feature wherein the given text is represented as a multiset of its words, disregarding grammar and word order. Multiplicity of words are maintained and stored as a word frequency vector. We applied standard word stemming and stoplisting to reduce the dictionary size, then created a word vector in which each component represents a word in our dictionary and its value corresponds to its frequency in the text. Finally, we create a word vector, where each component represents a word in the dictionary we have generated and its value corresponds to its frequency.

Offensiveness Following previous work [Kontostathis *et al.*, 2013] indicating that the occurrence of second person pronouns in close proximity to offensive words is highly indicative of cyberbullying, we use an "offensiveness level" (OFF) feature [Chen *et al.*, 2012]. We first use a parser to capture the grammatical dependencies within a sentence. Then for each word in the sentence, a word offensiveness level is calculated as the sum of its dependencies' intensity levels. We define the offensiveness level of a sentence:

$$O_s = \sum_w O_w \sum_{j=1}^{k} d_j$$

where $O_w = 1$ if word w is an offensive word, and 0 otherwise. For word w, there are k word dependencies, and $d = 2$ if dependent word j is a user identifier, $d = 1.5$ if it is an offensive word, and 1 otherwise.

Word2Vec Word2Vec is a state-of-art model for computing a continuous vector representation of individual words[Mikolov *et al.*, 2013], commonly used to calculate word similarity or predict the co-occurrence of other words in a sentence. Here we generate a Word2Vec comment feature vector by concatenating each word's vector, based on the observation that performing simple algebraic operations on these result in similar words' vectors. For testing purposes, we apply pre-trained vectors trained on data from the Google News dataset to generate the comment features.

5.2 A Feature Set for Posted Content

As discussed, our goal is to incorporate not only text features but also image features for detecting cyberbullying.

Feature	**Overall Accuracy**	Precision	Recall	F1-measure
BoW	76.74%	71.37%	82.11%	0.7636
OFF	74.53%	52.00%	97.05%	0.6771
Word2Vec	81.21%	85.47%	76.95%	0.8099
BoW, OFF	87.00%	82.74%	91.26%	0.8679
BoW, OFF, Word2Vec	89.31%	91.68%	0.8695%	0.8926
Captions, OFF, BoW, Word2Vec	**95.00%**	**94.74%**	95.26%	**0.9500**
CNN-Cl, OFF, BoW	86.90%	83.79%	90.00%	0.8678
CNN-Cl, Captions	84.53%	84.11%	84.95%	0.8453
CNN-Cl, Captions, OFF, BoW	93.21%	92.21%	94.21%	0.9320

Table 1: Classification results using SVM with an RBF kernel, given various (concatenated) feature sets. BoW=Bag of Words; OFF=Offensiveness score; Captions=LDA-generated topics from image captions; CNN-Cl=Clusters generated from outputs of a pre-trained CNN over images.

Our analysis of image content incorporated standard image-specific features (i.e., SIFT, color histogram), many of which have been successfully used in other work for similar non-descriptive research questions (e.g., [Datta *et al.*, 2006; Zerr *et al.*, 2012]). We additionally consider more sophisticated features extracted with deep learning and leveraged using unsupervised clustering methods.

Deep Learning for Clustering of Images Deep neural networks have proven extremely powerful for a wide variety of image processing tasks [Krizhevsky *et al.*, 2012]. We extracted image features using a pre-trained Convolutional Neural Network (CNN) [Theano Development Team, 2016], which is the benchmark standard for image classification and object detection tasks [Razavian *et al.*, 2014]. Deep Learning Networks (DL) refer to architectures which have more than 2 hidden layers (i.e., traditional MLPs). The deep learning mode we adopt [Krizhevsky *et al.*, 2012] has been shown to provide strong results in image classification challenges like the Imagenet [Deng *et al.*, 2009] competition. Deep learning algorithms are able to perform object recognition in realistic settings if a large number of samples are provided; sometimes millions of training images are needed. Given the limited number of images in our dataset, we cannot train our own deep learning network. Instead, we use a pre-trained network as a reasonable alternative. We use the implementation of Caffee [Jia *et al.*, 2014]. The first five layers of this network extract features by convolution of a set of image filters. Classification is influenced by the last three layers, with the 8th layer ultimately providing membership coefficients for 1000 topics in Imagenet, corresponding to detected objects.

We leveraged these learned memberships by clustering the images in our balanced dataset (1900 images), with the goal of grouping similar images, which could correspond to similar bullying signatures. Let M be an 1000x1900 matrix where M_{ij} measures the strength of membership of image j to topic i. This is qualitatively similar to a term-document matrix in Latent Semantic Indexing [Berry *et al.*, 1995]. Applying a singular value decomposition, we construct:

$$M = \mathbf{U} \cdot \mathbf{\Sigma} \cdot \mathbf{V}^T$$

where U is the topic matrix, V is the image matrix and Σ is the matrix of singular values. By keeping only the top $k = 200$ singular values, we reduce the dimensionality of the image vectors from 1000 to 200. There is no exact way to choose k, so we chose the value near a knee in the singular value curve. To construct clusters, we create an auxiliary graph structure whose vertices correspond to images. An edge is present if the cosine distance between the corresponding column vectors of $\mathbf{V}$ is sufficiently close:

$$1 - \frac{\mathbf{v}_i \cdot \mathbf{v}_j}{||\mathbf{v}_i||||\mathbf{v}_j||} < 0.05.$$

Clusters are defined as connected components in the resulting graph. A final cluster indicator vector is used as the feature for the classifier.

A Feature Set for Photo Captions Complementary to image features, we considered metadata originating from the image captions. We used Latent Dirichlet allocation (LDA) to analyze captions' text and extract their main topics [Blei *et al.*, 2003]. In LDA, each document is viewed as a mixture of various topics, and each topic defines a distribution over the words. This is similar to probabilistic latent semantic analysis (pLSA) [Hofmann, 1999], except that in LDA the topic distribution is assumed to have a Dirichlet prior. LDA is also not as prone to overfitting as pLSA. We identify 50 topics over the set of all captions, and use the proportionality of their respective presence in each caption as its feature vector. Presence of a topic in a given caption is determined as a relationship between the words appearing in the caption and a set of top ten keywords associated with that topic.

User Characteristics We considered several characteristics of individual users in the dataset: number of posts; followed-bys; replies to this post; average total replies per follower. We note that none of these features proved to be powerful in the course of our analyses. We conjecture the reason being the thread-like nature of Instagram, at least with respect to the posts we mined, which were largely from celebrity accounts. We suggest that these features might perform well in other datasets with more typical user accounts, wherein the number of followers and similar information may provide a profile for users more or less likely to be targets of bullying.

6 Post-Hoc Bullying Detection

We carried out a comprehensive set of experiments using a number of different classifiers for supervised learning, lever-

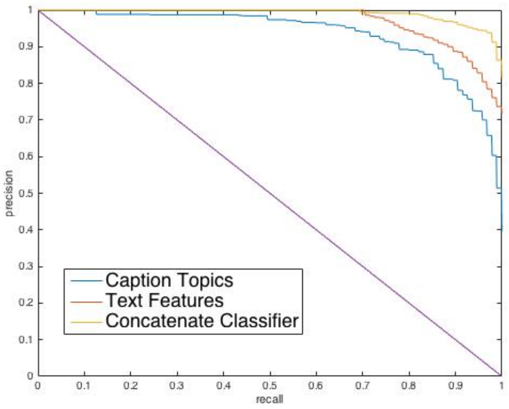


Figure 2: Precision-Recall Curve for Comment Bullying

aging combinations of the text- and image-based features described in Section 5 for the detection of occurrences of cyberbullying in peer-generated comments following posted images on the Instagram dataset.

In the family of supervised learning models, each model performs well for particular scenarios and poorly for others. Heterogeneity of data, data redundancy interactions among features are considerations when selecting a method. We experimented using a multi-layer perceptron, a Bayesian classifier and a Support Vector Machine (SVM). Our best results were obtained using an SVM with a radial basis function (RBF) kernel in OpenCV. Hyperparameters of the SVM were optimized using the cross-validation estimate of the validation set error. We first balanced the dataset using subsampling (950 bullying comments, 950 non-bullying). We used the standard k-fold validation technique ($k = 10$) to train and evaluate generalization accuracy.

We examine baseline text features (Bag of Word, Word2Vec and Offensiveness) for the comments and examine whether their replacement by or combination with meta-features (LDA-derived caption topics, Deep Learning) improves performance. Table 1 gives the performance of individual and concatenated feature sets in various combinations. Bag of Words and Offensiveness level perform well as baseline methods; Word2Vec outperform other features and proved to be a strong baseline. The concatenated BoW, OFF and Word2Vec feature set proves to be the most powerful combination of comment-based features.

As hypothesized, the addition of imaged-based features improves accuracy. While CNN-Cl is not very strong in this context, LDA-based caption tags reveal themselves as a powerful feature. As an ensemble, the concatenated feature set BoW, OFF, Word2Vec, Captions provides our strongest result at 95.00%. Intuitively, we may interpret this result as an indicator that an image's caption may trigger bullying in respondents, reveal valuable information about the nature of the image, or likewise about the nature of the user who has posted the image. Figure 2 illustrates the precision-recall curves for Captions and Comment Features (concatenated BoW, OFF, Word2Vec). While both have a similar Area Under the precision recall curve, their shape is not the same and our final concatenated classifier's performance is superior to that of the individual classifiers.

7 Anticipating Cyberbullying of Shared Images

With regard to our second research question, our task is the classification of a given image as either *bully-prone* or *nonbully-prone*. That is to say, we seek to calculate the probability that an image, along with the context of its posting (characteristics of the user posting the image, and the caption he appends) is a trigger for cyberbullying events. Theoretically, we may expect this result to depend on three feature sets: characteristics of the user who has posted the image, which we denote U_i (*e.g.*, age, gender), image content C_i, and image metadata, as defined by the owner-posted tags and captions S_i. These distinct feature sets permit three dimensions of analysis. That is, who is the bullied user? How does the substance of his shared image incite a negative reaction? And, how does the image metadata either catalyze or inform the occurrence of abuse in peer responses? Note that anticipation of cyberbullying in response to shared media is a non-trivial task. There is strong evidence that contextual factors, such as users' behavior outside of the social network, influence cyberbullying behavior [Snell and Englander, 2012]. Here we aim to capture complex features of the posted image and caption as an important piece of that context.

7.1 Instagram-specific Visual and Metadata Features and Results

We now report results using features derived from the posted images and captions themselves, without comment features.

We use the LDA-generated caption topics (Captions) and the pre-trained CNN (see Section 5.2) for the images themselves. Because we are neither interested in clustering the images, nor in object identification, we take a different approach to the implementation of the CNN for this task. Previously, we clustered the output of the 8th layer of the CNN, representing membership coefficients in 1000 topics in Imagenet. In this image-classification task, we disregard the 8th layer and treat the output of the 7th layer of the network as a 4096-dimensional feature vector, describing high-level features of each image but not topics explicitly. We refer to these as Deep Learning features (DL).

Because the SVM attempts to maximize the margin between two classes, greater distance implies greater confidence. The precision-recall curve in Figure 3 illustrates the detection accuracy of the Captions and DL-FS features, alone and in ensemble. Random classification yields the diagonal line, illustrated. We obtain the precision-recall curve by varying the threshold of the classifiers. For SVM, the threshold is the bias in the support vector.

Note that the Captions-only classifier (blue line) outperforms DL confirming the importance of the direct textual input of the posting user as a trigger for cyberbuyllying events. The concatenated classifier closely follows the performance of DL. One explanation for this may be the very high feature dimension when compared to Captions. The ensem-

Feature	Overall Accuracy	Precision	Recall	F1-measure
CH	54.56%	51.79%	57.32%	0.5441
GIST	56.85%	55.89%	57.80%	0.5682
SIFT	56.88%	53.21%	60.54%	0.5663
GIST+SIFT (concatenate)	58.82%	59.27%	58.36%	0.5881
DL	59.82%	64.00%	55.64%	0.5953
GIST+SIFT, DLFS (Stacked)	60.64%	62.73%	58.54%	0.6056
DL-FS	61.19%	49.82%	72.55%	0.5907
Captions	68.09%	55.45%	80.73%	0.6574
Captions, DL (Stacked)	67.73%	67.64%	67.82%	0.6773
Captions, DL-FS (Stacked)	**68.55%**	62.91%	74.18%	0.6808

Table 2: Feature combinations for the detection of images prone to cyberbullying; CH=Color Histogram; DL=Deep Learning; DL-FS=Deep Learning with Feature Selection

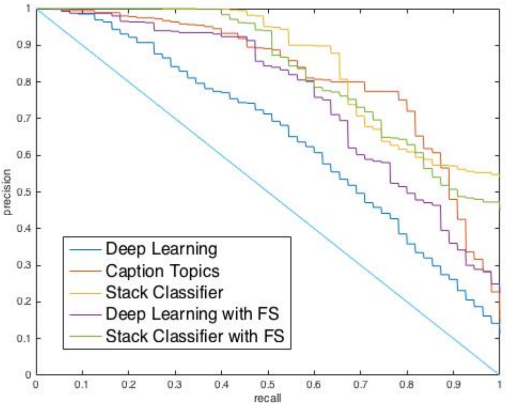


Figure 3: Precision-Recall for Image Bullying detection task

ble classifier alleviates this issue and slightly outperforms all other methods. The values of the area under the precision/recall curve are: concatenated classifier (0.6573), Captions (0.8209), stacked classifier (0.8537), DLFS (0.7601), stacked classifier with FS (0.8308).

7.2 Classification with Additional Visual Features

We compare the performance of our Captions and DL features separately and in combination with image features which have been shown to help extract the semantics of an image. Due to space limitations, we here discuss most important results and significant features.

Scale-Invariant Feature Transform Scale-invariant feature transform (SIFT) finds interesting points in an image and provides 128-dimensional descriptors for each one of them [Philbin *et al.*, 2007]. Based on these descriptors, we use a popular technique [Zerr *et al.*, 2012] to quantize the interesting points into "words" and build a bag of visual words. From these we create sparse, fixed-length vectors for each image that encode the number of occurrences of each word and perform k-means clustering to learn a visual codebook for vector quantization over the entire dataset.

Color Histogram (CH) Some colors, like skin color or outdoor colors (blue or green) may be helpful in discriminating certain type of bullied images (e.g. photos at the beach). To capture this, we consider Red, Green and Blue (RGB) intensities for a total of 3 categories of color histograms. Following the methodology of [Raja *et al.*, 1998], we divide each feature space into 256 bins representing distinct intensities on each spectrum and based on the member pixels in each bin, we create one final 768-dimension feature vector.

Gist of image(GIST) The "gist" of a scene is an abstract representation that spontaneously activates memory representations of scene categories. The GIST feature [Oliva and Torralba, 2001] represents five perceptual dimensions of an image, namely naturalness, openness, roughness, expansion, and ruggedness designed to capture the spatial structure of a scene. This low-dimensional, holistic representation of the image is informative of its probable semantic category.

SIFT, CH and GIST features were also submitted to feature selection as described for the DL feature. Obtained features were input to an SVM classifier with RBF kernel. Our results for various feature combinations are in Table 2. Note that our advanced features significantly outperfom baseline textual-based features independently and in combination, confirming the need of advanced image features to capture the subjective nature of images sensitive to cyberbullying.

8 Conclusions

We have studied the detection of cyberbullying in photo-sharing networks, with an eye on the development of early-warning mechanisms for identifying images vulnerable to attacks. In the context of photo-sharing, we have refocused this effort on features of the images and captions themselves, finding that captions in particular can serve as a surprisingly powerful predictor of future cyberbullying for a given image. This work is a foundational step toward developing software tools for social networks to monitor cyberbullying.

9 Acknowledgments

Griffin's, Squicciarini's and Rajtmajer's work was partly supported by the Army Research Office under grant W911NF-13-1-0271. Portions of Dr. Squicciarini and Caragea's work were also supported by National Science Foundation grant 1421776 and grant 1421970, respectively.

References

[Berry *et al.*, 1995] M. W. Berry, S. T. Dumais, and G. W. O'Brien. Using linear algebra for intelligent information retrieval. *SIAM Review*, 37(4):573–595, 1995.

[Berson *et al.*, 2002] Ilene R Berson, Michael J Berson, and Michael J Berson. Emerging risks of violence in the digital age: Lessons for educators from an online study of adolescent girls in the united states. *Journal of School Violence*, 1(2):51–71, 2002.

[Blei *et al.*, 2003] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

[Chen *et al.*, 2012] Ying Chen, Yilu Zhou, Sencun Zhu, and Heng Xu. Detecting offensive language in social media to protect adolescent online safety. In *International Conference on Privacy, Security, Risk and Trust (PASSAT)*, pages 71–80. IEEE, 2012.

[Cyberbullying Research Center, 2016] Cyberbullying Research Center. Cyberbullying on Instagram, 2016. http://cyberbullying.org/cyberbullying-on-instagram/.

[Dadvar *et al.*, 2013] Maral Dadvar, Dolf Trieschnigg, and Franciska Jong. Expert knowledge for automatic detection of bullies in social networks. 2013.

[Datta *et al.*, 2006] Ritendra Datta, Dhiraj Joshi, Jia Li, and James Z Wang. Studying aesthetics in photographic images using a computational approach. In *Computer Vision–ECCV 2006*, pages 288–301. Springer, 2006.

[Deng *et al.*, 2009] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Conference on Computer Vision and Pattern Recognition, 2009. CVPR 2009*, pages 248–255. IEEE, 2009.

[Dinakar *et al.*, 2011] Karthik Dinakar, Roi Reichart, and Henry Lieberman. Modeling the detection of textual cyberbullying. In *The Social Mobile Web*, 2011.

[Harris, 1954] Zellig S Harris. Distributional structure. *Word*, 1954.

[Hinduja and Patchin, 2013] Sameer Hinduja and Justin W Patchin. Social influences on cyberbullying behaviors among middle and high school students. *Journal of youth and adolescence*, 42(5):711–722, 2013.

[Hofmann, 1999] Thomas Hofmann. Probabilistic latent semantic analysis. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, pages 289–296. Morgan Kaufmann Publishers Inc., 1999.

[Jia *et al.*, 2014] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the ACM International Conference on Multimedia*, pages 675–678. ACM, 2014.

[Kontostathis *et al.*, 2013] April Kontostathis, Kelly Reynolds, Andy Garron, and Lynne Edwards. Detecting cyberbullying: query terms and techniques. In *Proceedings of the 5th annual acm web science conference*, pages 195–204. ACM, 2013.

[Krizhevsky *et al.*, 2012] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[Mikolov *et al.*, 2013] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[Oliva and Torralba, 2001] Aude Oliva and Antonio Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *Int. J. Comput. Vision*, 42(3):145–175, May 2001.

[Philbin *et al.*, 2007] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2007.

[Raja *et al.*, 1998] Y. Raja, S.J. McKenna, and Shaogang Gong. Tracking and segmenting people in varying lighting conditions using colour. In *Third IEEE International Conference on Automatic Face and Gesture Recognition*, pages 228–233, Apr 1998.

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

[Sabella *et al.*, 2013] Russell A Sabella, Justin W Patchin, and Sameer Hinduja. Cyberbullying myths and realities. *Computers in Human behavior*, 29(6):2703–2711, 2013.

[Snell and Englander, 2012] P.A. Snell and E.K. Englander. Cyberbullying victimization and behaviors among girls: Applying research findings in the field. *J. Soc. Sci.*, 6:510–514, 2012.

[Theano Development Team, 2016] Theano Development Team. "convolutional neural network example in theano", 2016. http://deeplearning.net/tutorial/lenet.html.

[Yin *et al.*, 2009] Dawei Yin, Zhenzhen Xue, Liangjie Hong, Brian D Davison, April Kontostathis, and Lynne Edwards. Detection of harassment on web 2.0. *Proceedings of the Content Analysis in the WEB*, 2:1–7, 2009.

[Zerr *et al.*, 2012] Sergej Zerr, Stefan Siersdorfer, Jonathon Hare, and Elena Demidova. Privacy-aware image classification and search. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '12, pages 35–44, New York, NY, USA, 2012. ACM.