Multimodal Classification of Moderated Online Pro-Eating Disorder Content

Stevie Chancellor*
Georgia Tech
Atlanta, GA USA
schancellor3@gatech.edu

Yannis Kalantidis
Yahoo Research
San Francisco, CA USA
ykalant@image.ntua.gr

Jessica A. Pater
Georgia Tech
Atlanta, GA USA
pater@gatech.edu

Munmun De Choudhury
Georgia Tech
Atlanta, GA USA
munmund@gatech.edu

David A. Shamma
Centrum Wiskunde & Informatica (CWI)
Amsterdam, Netherlands
aymans@acm.org

ABSTRACT
Social media sites are challenged by both the scale and variety of deviant behavior online. While algorithms can detect spam and obscenity, behaviors that break community guidelines on some sites are difficult because they have multimodal subtleties (images and/or text). Identifying these posts is often regulated to a few moderators. In this paper, we develop a deep learning classifier that jointly models textual and visual characteristics of pro-eating disorder content that violates community guidelines. Using a million Tumblr photo posts, our classifier discovers deviant content efficiently while also maintaining high recall (85%). Our approach uses human sensitivity throughout to guide the creation, curation, and understanding of this approach to challenging, deviant content. We discuss how automation might impact community moderation, and the ethical and social obligations of this area.

ACM Classification Keywords
H.5.m. Information Interfaces and Presentation (e.g. HCI): Miscellaneous

Author Keywords
content moderation; social media; pro-eating disorder; deep learning; computer vision; deviant behavior; Tumblr

INTRODUCTION
With an increasing volume of data, online social platforms like Facebook and Tumblr face an increasing task of identifying content that violates policies and guidelines. Potential violations include accidentally posting copyrighted material [34], deceptive online reviews [64], and online harassment [61], among others. In this paper, we refer to violations of explicit rules and community guidelines as *deviant* content. When managing deviant content, platforms trade-off scalability and contextualization. To handle data volume, platforms have adopted some automated measures, most famously to identify spam [31]. Research has classified complex deviant behaviors as well, like trolling and abusive language [1, 21]. Yet, detecting these kinds of subtleties at scale has been elusive—most platforms rely on human moderators to review content.

Our area of interest is the Tumblr pro-eating disorder (pro-ED) community. Tumblr is a microblogging social network founded in 2007 focusing on short-form content sharing. Tumblr's community guidelines prohibit the glorification of self-harm, including promoting eating disorders and their accompanying lifestyles. This includes "content that urges or encourages others to... cut or injure themselves; [or to] embrace anorexia, bulimia, or other eating disorders" [85]. Posts in pro-ED communities can break these guidelines by encouraging eating disorders as a healthy lifestyle alternative, glorifying low weight and thinness [25]. They share extremely low calorie diets, pictures of underweight individuals, and stories of dangerous behaviors like dramatic weight control measures.

The Tumblr pro-ED community is a case study in detecting complex deviant content. Pro-ED communities have challenged social media platforms because of their pervasiveness despite explicit moderation against the community [17]. Not all content in pro-ED communities is dangerous [14] nor do all images qualify for removal—the same photo of a model could be used for a pro-ED blog or a fitness blog for motivation. However, for the content that is dangerous, psychological research shows that there are unique social contagion-like effects on those who are exposed to this content [10]. These posts might also encourage self-harm in users by letting them "live through" others' self-harm experiences [56].

In spite of these risks to the broader Tumblr community, deviant pro-ED content is rarely reported to moderators. This is because pro-ED communities are often "hidden in plain

*This work was conducted while author was at Yahoo.

CHI 2017, May 06-11, 2017, Denver, CO, USA
 ISBN 978-1-4503-4655-9/17/05 ...$15.00
DOI: http://dx.doi.org/10.1145/3025453.3025985

sight", adopting the use of atypical language or tags [17]. This reduces the chance that outsiders to these communities will encounter and report them. Further, individuals within these communities, looking for support in maintaining disordered behaviors, are unlikely to recognize dangerous pro-ED content as violations of community guidelines. Compounding these concerns about deviant behavior are the use of images in pro-ED communities. These images show explicit negative emotions and graphic content of thinness, motivation for starvation, and even pictures of self-injury [66]. This presents two unique challenges for existing moderation strategies. First, state-of-the-art approaches in automated detection of deviant content primarily rely on textual signals [20, 21]. When these methods are applied in the context of pro-ED, they are likely to miss the rich signals that are distinctively conveyed through images [39]. Second, for human moderators who must interact with this emotionally challenging and sensitive content, it may require domain-specific knowledge as well as an "emotional shield" [4, 19].

Prior work in HCI, mental wellness, and prediction has focused on the identification of specific mental wellness challenges like depression[26] or assessing more general wellness states of those who have posted to pro-ED communities[14]. This work is distinct from these works in three major ways. First, the prediction task in our work is identifying deviant content, not a post or a user who engages with this content. Second, we offer a novel methods contribution by bringing deep learning to social computing to understand more than just text content. Third, this paper actively scaffolds another methods contribution in HCI to building context-sensitive deep learning approaches to identifying deviant content.

In this paper, we address these challenges in detection and moderation of sensitive deviant content like pro-ED. We propose, develop, and evaluate a supervised learning model that distinguishes between deviant pro-ED content and acceptable content. Our proposed model is *multimodal*—it uses both textual and visual characteristics of pro-ED content. For this, we leverage recent advances in computer vision [72] and large-scale text mining [63]. Our model directly incorporates *human assessments* and works to support moderation pipelines with domain knowledge and reduced emotional stress.

To build this multimodal human-machine hybrid classifier, we first curate a dataset of nearly a million pro-ED posts on Tumblr. We then bring in human sensitivity by deploying an iterative expert rating task that identifies thousands of pro-ED posts as potential community guideline violations. Using this annotated data for training, we then evaluate a state-of-the-art Deep Neural Network classifier [54]. This classifier has high performance in detecting deviant pro-ED content—accuracy: 89% and F1: 65%. Importantly, we show that this model performs comparably well in classifying our expert rated posts and the posts actually removed by moderators for breaking community guidelines. We finally discuss the design, ethical, and social implications of our approach for both existing and next generation content moderation systems.

Our project considered ethics at all steps, and we looked to prior work in how to handle sensitive, removed content [5, 15]. All data, including posts that were later removed, were publicly available when we obtained the dataset. The authors had no research interactions with users, nor did they interact with Tumblr by reporting posts. Researchers contacted several users to obtain passive consent to use their photos in the paper (ref. Supplemental Materials). This means that our research did not qualify for ethics board review. To preserve the security and privacy of our data, all photos were stored and accessed from secured, firewalled servers. We do not report any results with personally identifying information, including identifying pictures, tags, or usernames. Some images are graphic.

RELATED WORK

Our work builds on research in HCI and social computing, including online deviance and content moderation and sensitive domains like pro-eating disorder.

Defining Deviant Behavior

We define *deviant behavior* as actions that are socially outside the norms for a community or group of people [3]. Behaviors can be violations of explicit rules and laws or social standards and norms. Deviant behaviors are contextual—whether a behavior is considered "bad" is dependent on the circumstances, the community, and the timing of that particular activity.

Many stakeholders develop and negotiate what is deviant on a particular website. These norms may be contained in explicit Terms of Service agreements, community guidelines, site-wide usage policies, and even unwritten expectations of a community; what becomes deviant is negotiated and constantly iterated on by these stakeholders. This can include the platform owners of the site, the moderators and administrators, and the individual users themselves [51]. Many studies have explored why deviant behavior exists in online communities, using psychological explanations of deviant behavior [35, 78, 79] as well as perspectives on the community itself [30, 41].

We study deviant behavior in a sensitive community: the pro-ED community. These behaviors are deviant as they break Tumblr's Community Guidelines, but there are unique concerns with this kind of content. Pro-ED behaviors have actual contagion effects; that is, when others see content promoting an eating disorder, these negative emotions can spread to others [60]. This is amplified by the fact that looking at self-harm lets users "live through" others' self-harm experiences [56].

Quantitative Studies of Online Deviant Behaviors

The HCI community has investigated deviant behaviors in online communities for over two decades [11, 77], and recently this topic has gained renewed attention. In this section, we focus on quantitative and machine learning methods to detect and predict online deviant behaviors.

Classification tasks on deviant behaviors fall into a few specific areas. One is predicting when someone may be engaging in cyberbullying or online harassment [42, 68]. Other work focuses on identifying trolling and antisocial behaviors. This uses the unique linguistic features of trolling news comments [20, 21], finding anti-social behavior in online games [9], or using psychological cues to understand the habits of trolls [12]. Research on identifying Wikipedia vandals incorporates user analysis [1, 44] and meta data like revision history [84] into their prediction task. Finally, there is work that uses the act

of deletion as the response variable in the prediction [22, 74]. Close to our work is Chancellor et al. [15] that built a classifier to predict whether Instagram posts would be removed from the platform; however, the work did not detect deviance from the perspective of moderators.

These works offers empirical evidence that social platforms provide signals of deviant behaviors using primarily text analysis. In pro-ED communities, images are a strong visual signal used for self-disclosure [66]. Images are a powerful means of expression that can covey feelings and emotions too complex for written communication [39]. They can also be a tool for identity management and communication techniques [75]. Therefore, we believe images are an important signal in the pro-ED community. Our work examines how the integration of both images and text can predict deviant pro-ED content, improving on the capacities of either separately.

Pro-ED Communities

Pro-eating disorder (or pro-ED) communities are groups that promote eating disorders as an alternative lifestyle choice rather than the dangerous mental disorder that they are. Users in these communities share diet advice, inspirational images of thin bodies, and general support to maintain low body weight and thinness. In this section, we discuss the unique appearances of pro-ED communities online and why they have challenged social networking platforms.

Pro-ED communities are a subset of and blend easily into eating disorder communities at large [17]. Importantly, not all posts about eating disorders are dangerous. For example, someone can post about their struggles with an eating disorder without encouraging others to adopt their behaviors. Because of this, the content around eating disorder communities online is contextually complex. On Instagram, users can move between promoting eating disorders and struggling to stop their behaviors on the same user account [14]. These nuances have led platforms like Tumblr to establish community guidelines that distinguish between these kinds of content [85].

Despite these guidelines, these communities persist on social platforms [57]. Qualitative research on three social networks show users appropriate pro-ED channels for negative health outcomes [66]. On Instagram, pro-ED communities adopt more linguistically variant tags after platform-enforced moderation, and these new communities are more likely to discuss suicidal and self-harm ideas [17]. Additionally, there is often collision between the pro-recovery community (that sites often want to support) and the pro-ED community [16, 25] that occasionally play out on the site [88]. Given the efforts of platforms to curb pro-ED, we believe an algorithmic approach boosted by knowledge of community moderation practices complements current approaches to handling pro-ED content.

Moderation Strategies in Online Communities

Since the emergence of online communities, moderation strategies have been a critical area of research [11, 24, 50, 77]. One moderation style allows users to moderate their own content. On one extreme, 4chan content is unmoderated, and the community has informal moderation of good content by "bumping" desirable threads and using the "sage" command to comment without bumping [8]. Many sites build explicit social moderation systems, where users can publicly vote for or against content on the site. This is used on sites like Reddit, MetaFilter, Yik Yak, Stack Exchange, and Slashdot. The effectiveness of these social moderation strategies are mixed and are often dependent on user engagement with these systems [38, 55].

Another style of moderation uses outright banning strategies. In this case, platforms either ban keywords about deviant behaviors, ban users, or delete entire communities. In many cases, outright banning of keywords limits the dissemination of certain kinds of content, like spam or pornography. However, these kind of moderation strategies can have unintended consequences. On Instagram, researchers studied behavior patterns following banning of tags in pro-ED communities and found that users developed new lexical variants to avoid these keyword bans [17]. On Reddit, the site banned several hateful subreddits, r/fatpeoplehate and r/coontown (a subreddit dedicated to degrading black people), and the community reacted strongly both for and against the banning [13].

Finally, another strategy uses human moderators to evaluate content. Almost all sites have volunteer or paid moderators to police content. Not all moderator actions are negative—some approaches help editors find beautiful weather photos [73], where others help find good social media content [2] and high-quality news comments [29, 65]. Between these three approaches, human decision-making is the common thread; these examples motivate us to use human judgments in our multimodal classification method.

DATA

In this section we present our data collection method in three steps: (1) building a master dataset of photo posts associated with eating disorders; (2) identifying content removed by Tumblr moderators for guideline violations; and (3) building a dataset of Tumblr posts semantically close but unrelated to eating disorders. Figure 1 shows our data-gathering strategy.

Overview of Tumblr

Tumblr is a microblogging and social networking site founded in 2007. Users post multimedia content to short-form personal blogs. The site has social components, like following and messaging other blogs and reblogging content. As of September 2016, Tumblr had over 300 million blogs and 45 million daily posts, making it one of the most popular social network sites[1].

Posts on Tumblr can be placed into 9 post types: `text`, `photo`, `video`, `audio`, `quote`, `question`, `link`, `caption`, or `GIF`[2]. About 75% of Tumblr posts are photo posts [18]. Tumblr does not have explicit community structures, like dedicated forums, where users would specifically post content about eating disorders. Instead, communities on Tumblr form around hashtags [43].

Building the Master Pro-ED Dataset

Our master dataset contains about 877,000 public photo posts from the Tumblr pro-ED community between November 2015

[1] http://www.tumblr.com/press
[2] https://www.tumblr.com/docs/en/api/v2

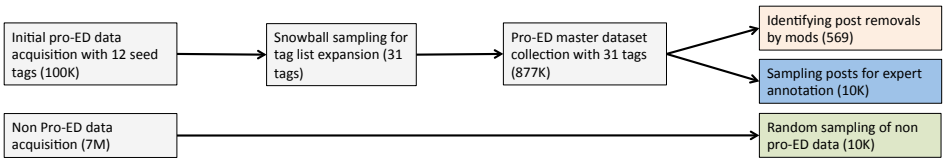


Figure 1: Schematic flowchart of our data collection strategy.

and August 2016. To assemble this dataset, we used the following approach:

Seed Data Collection. First, we crawled one month of public Tumblr posts to create a set of eating disorder and pro-ED tags. We referred to the list of tag suggestions made in the works of Chancellor et al [17] and De Choudhury [25]; both developed a lexicon of pro-ED related tags for Instagram and Tumblr, respectively. Compiling the two studies, we developed a short list of 12 known eating disorder and pro-ED tags[3] including #anorexia, #eating disorder, #pro ana, and #thighgap. Because Tumblr allows for whitespaces in their tags, we also included spaces that naturally separate words or phrases. For example, we included both #eating disorder and #eatingdisorder in our data crawl. This initial crawl with 12 seed tags returned about 100,000 seed posts from May 2016.

eating disorder	thinspo	thynspo	bonespo
eatingdisorder	proana	pro ana	thighgap
mia	ednos	anorexia	pro anorexia

Table 1: Examples of our tags used to crawl pro-ED posts.

Snowball Sampling. Using our seed tags, we used the methods in [17] to snowball and select additional eating disorder and pro-ED tags in our 100,000 photo posts from May 2016. From this list, we filtered for all tags that co-occurred with our initial set of 12 seed tags above a certain probability threshold (2%). We excluded tags related to recovery (*e.g.*, #ed recovery) or that were too general (*e.g.*, #ugly, #fat or #sad teen). This produced 31 unique tags for our final data collection. Some example tags are given in Table 1.

Final Data Collection. Next, we used the 31 tags to get a sample of pro-ED photo posts shared between November 2015 and August 2016. We ran a list-based filter on these posts to exclude posts with salacious tags (*e.g.*, #boobs) or recovery-related tags—recovery tags are known to promote healthy behaviors [16], and therefore unlikely to break Tumblr's guidelines. This master dataset contains about 877,000 photo posts from November 2015 to August 2016. For each post, we gathered its metadata and image. Summary statistics and graphs are provided in Table 2 and in Figure 2. We indicate the 25 most frequent tags in Table 3.

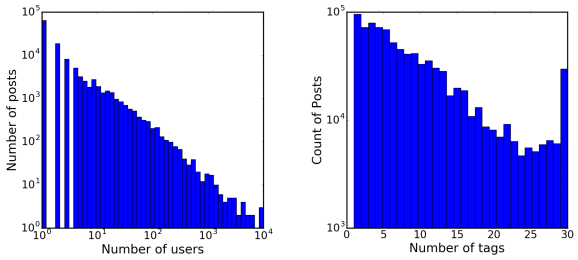
Compiling Posts Removed By Tumblr Moderators

The second step of our data collection is finding photo posts removed by Tumblr moderators for violating Tumblr's community guidelines. Recall that Tumblr community guidelines

[3]The 12 seed tags were #anorexia, #anoreixa, #eating disorder, #eatingdisorder, #ednos, #proana, #pro ana, #thinspo, #thynspo, #thighgap, #thigh gap.

Unique Blogs	118106	Unique Tags	297,597
Average Posts/Blogs	7.43	Average Tags/Post	8.85
Median Posts/Blogs	1	Median Tags/Post	6
Standard Deviation	84.47	Standard Deviation	7.56

Table 2: Summary statistics of the master dataset of 877,998 Tumblr posts. This includes how many posts Tumblr blogs generate as well as the use of tags per post.



(a) (b)

Figure 2: (a) Distribution of posts over users, and (b) distribution of tags over posts in the master dataset.

prohibit the glorification of self-harm and eating disorders like anorexia or bulimia [85]. When Tumblr staff removes a post for breaking community guidelines, they overwrite the original image with a new image that says that Tumblr removed the original for violating community guidelines seen in Figure 3. This action indicates if a post was removed by a moderator.

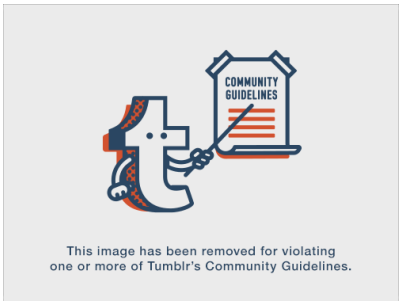


Figure 3: Default image used by Tumblr to substitute images in posts that violate community guidelines.

To gather moderator-removed posts, we first downloaded all images in our master dataset. This download happened in June and August 2016 (to get July and August's photo data). Using a strategy similar to [15], we re-downloaded the same set of images in September 2016. To detect images that were taken down for violating community guidelines, we tracked images with changes in file size and used a fast visual similarity approach [47] to find images identical to the default image in Figure 3. We found 569 posts removed by Tumblr moderators between June and September 2016.

thinspo	skinny	ana	thin	anorexia
thinspiration	mia	ed	bulimia	eating disorder
depression	fitspo	thigh gap	anorexic	depressed
suicide	proana	sad	anxiety	suicidal
self harm	motivation	pro ana	goals	tw

Table 3: Top 25 tags in our master dataset.

We found that the proportion of photo posts removed due to breaking the guidelines is low. We have several hypotheses that might explain this phenomenon. Even with the most aggressive moderator removal, social media platforms cannot examine all content that should be removed from their platforms. Most platforms rely on reported data from users to drive their curation and content moderation efforts [19]. Second, in communities like pro-ED, posts glorifying self-harm behaviors like purging or extreme food restriction are less likely to be flagged by community members who are deliberately seeking this content. These communities are often "hidden in plain sight" where anyone can find them, but their tags prevent most outside parties from discovering the community [17].

Collecting Non Pro-ED Data

Finally, we collected a sample of Tumblr photo posts that are semantically close to pro-ED images but are topically unrelated. We first manually inspected a sample of 200 posts in the master dataset that were not removed due to community guideline violations. We obtained tags in these posts unrelated to deviant pro-ED behaviors [15]. Examples of these tags include: #outfit, #fashion, #selfie, #fitness, and #fitchicks. Using these tags, we gathered a large set of over seven million posts from August 2016 and excluded posts with any of the 31 pro-ED tags from the master dataset. From this, we randomly sampled about 10,000 public photo posts.

METHODS

In this section, we describe how the data was used to detect deviant pro-ED content on Tumblr as well as the details of our deep learning classification framework.

Methodological Challenges in Detecting Deviant Content

In sensitive communities like pro-ED, it is not surprising that there are only a few examples of deviant content that were *actually moderated*. This slows down getting sufficient gold standard labels (or ground truth) to train a robust classifier.

There is also additional complexity in this classification task. As shown in prior work [14], humans are good at identifying the obvious extremes of what abides by community guidelines (*e.g.*, fitness posts versus pro-ED diet advice). However, certain pro-ED posts are challenging for humans and moderators alike (*e.g.*, thinspiration versus pro-ED diet advice). With access to only the posts that moderators removed, there is no way to identify the posts that moderators found challenging but eventually allowed to stay on the platform.

At first glance, these challenges might be addressed by crowdsourcing the ground truth labels, a technique used in many supervised learning problems [53, 76]. However, due to the inherent subtleties in pro-ED posts, there is a major resource cost as well as an emotional cost to identifying sensitive material [66]. It may also be challenging to find crowdworkers with significant skill to accurately assess these posts [14].

To tackle these challenges, our classification framework learns from ground truth assessments of deviant pro-ED content from domain experts. They have knowledge of both Tumblr's community practices, guidelines, and the pro-ED community. Our experts labeled content into three categories: (1) posts that clearly do not violate the guidelines; (2) those that do conform to guidelines, but are difficult to assess; and (3) deviant pro-ED content or potential violations. This solution addresses both the need for skill and the human sensitivity we want to bring into our method. Training and testing on challenging examples also provides insights into our classifier's performance over a range of deviant or non-deviant posts. This approach more closely replicates how we expect our human-machine moderation system to be developed by others.

Iterative Expert Rating Task for Ground Truth

In this section, we describe our rating task to use human judgment to develop a ground truth dataset.

Rater Credibility and Expertise. Our 3 raters are authors on this paper and are social computing researchers who have considerable research experience in quantitative and qualitative studies of mental health, specifically pro-ED. One rater was employed at Yahoo, the parent company of Tumblr.

Developing a Rationale for Assessing Guideline Violations. We wanted to objectively capture what might constitute removal from Tumblr's platform for "promoting or glorifying eating disorders or self harm." [85] Defining this involves context around the post as well as the community itself—recall that posting to a pro-ED tag is not necessarily grounds for removal on Tumblr. As the guidelines specify, this kind of content must, "urge or encourage others to: cut or injure themselves; embrace anorexia, bulimia, or other eating disorders; or commit suicide." [85] Therefore, we holistically evaluated a post to decide if it might be deviant content. We interpreted promotion either as encouraging the maintenance of an eating disorder or promoting related actions. The researchers used the posts' tags, image, the caption, and the author's username.

In the first iteration, the three raters independently rated 250 public Tumblr posts randomly sampled from the master dataset (ref. Data section). Ratings were based on a three-point scale: photo posts that would potentially be removed due to guideline violation (rating 3); posts that might be a challenging case of moderation but are actually not guideline violations (rating 2); and posts that should continue to remain on Tumblr (rating 1).

Then, the raters resolved rating differences and designed a shared rulebook that included their rationale to assess posts:

- *Rating 3:* This category contained posts where the whole post promoted negative behaviors or actions. This included: pro-eating disorder diet advice, extreme calorie restriction posts, or starvation; posts promoting beauty or transformation through thinness or starvation; glorification of thin body parts associated with eating disorders (collar bones, hip bones, disproportionately slender legs or arms, visibly prominent rib cages, etc.); and body checks.

Acronym	Description	#Posts
GV	*Guideline Violations:* Posts that have been removed from Tumblr by moderators	569
PGV	*Potential Guideline Violations:* Posts that were annotated to be potentially violating community guidelines	1,207
GCS	*Guideline Conforming—Simple:* Posts that were annotated to conform to community guidelines. Also includes the posts acquired by filtering for fitness and appearance related tags	11,545
GCH	*Guideline Conforming—Hard:* Posts annotated to be challenging in their assessment of adherence to community guidelines	2,087

Table 4: Description and statistics of different datasets used in classification.

	Training	Validation	Testing
Positives	80% PGV (966)	20% PGV (241)	GV (569)
Negatives	56% GCS (6465), 56% GCH (1169)	14% GCS (1616), 14% GCH (292)	30% GCS (3464), 30% GCH (626)

Table 5: Composition of training, validation, and test data. Numbers in brackets indicate the number of posts.

- *Rating 2:* This category contained posts with a mismatch between tags/captions and the images. This also included discussion about mental illness and eating disorders without actively promoting them. This included: general discussion about mental health or eating disorders; pro-ED tags with a safe photograph; mental illness discussion; and comments about potential relapsing to pro-ED lifestyles.
- *Rating 1:* These were posts that did not show pro-ED sentiments at all. This included general "thinspiration" or fitness motivation; fitness posts; photos of models; posts encouraging recovery or support; posts unrelated to eating disorders.

With this rulebook, the researchers rated an additional set of 50 posts to test the convergence of their rating system. Using Fleiss' interrater reliability metric κ, we found high agreement ($\kappa = 0.7$) between the three raters. Additionally, we measured how well our ratings identified posts that should be removed (rating 3) and the posts with clear signals of guideline conformity (rating 1) and found that interrater reliability was higher ($\kappa = .8$). In our initial sample of 50 posts following rater calibration, 15/50 of posts received at least one 3, 13 had two 3s, and 8 had three 3s. Greater agreement for distinctive post categories further bolsters the rigor of our rating task.

For the final step in our rating task, another 5000 posts were randomly selected from our master dataset (without replacement from the previous rating samples), and two raters assigned a single score to every post. Prior work shows that the most dangerous content in these communities is relatively uncommon [14] and we predicted it would be our smallest category. To boost the size of this category for the classification task, we identified the top 25 tags on other posts labeled as a 3 in the initial set of 5000. Using these tags to filter, we identified a final sample of 960 photo posts from our master dataset—this narrowed our search space of finding posts most likely to be a 3 so we could quickly bulk up our set of posts with a rating of 3. The same raters rated these additional posts. In total, 5960 posts were rated.

In Figure 4 we provide examples of rated content. These pictures are publicly available, but to respect the users, we obtained passive online consent to use their photos. Image A was rated 1 and had fitness and wellness tags. Image B shows a drawing and had mental illness tags. Image C is a very slim person and had tags that encouraged starvation.

(a) Rating: 1 (b) Rating 2 (c) Rating 3

Figure 4: Example images rated 1, 2, 3 for their likelihood of violating community guidelines. Passive online consent was sought from the account owners (see Supplementary Materials for verbiage of the consent we sought).

Constructing Training, Validation, and Test Datasets

Bringing our data collection and rated posts together, we discuss the construction of the training, validation, and test sets.

Positive Examples/Class 1: Our positive examples (deviant content) for *training* and *validation* included posts that rated as a 3—posts that potentially violate the community guidelines (referred to as PGV or Potential Guideline Violations now on). We split this dataset 80% for *training* and 20% for *validation*. Our positive examples for *testing* included posts that were taken down from Tumblr by the moderators (referred to as GV, or Guideline Violations). Using this approach, we can judge the performance of our classifier against existing moderation.

Negative Examples/Class 0: Our negative examples contained three sets of data: (1) photo posts gathered in the Data section unrelated to eating disorders; (2) the annotated posts that were rated a 1; (3) the annotated posts rated a 2. Because the first two sets of negative posts are likely to have distinctive visual and textual markers in contrast to pro-ED posts, we refer to them as GCS (or Guideline Conforming—Simple). The third set of posts are challenging or contextually complex, so we refer to them as GCH (or Guideline Conforming—Hard). We randomly split this combined set of negative examples into 70% for *training* and *validation* and 30% for *testing*. In Table 4 we summarize the data in our classification tasks and their corresponding counts. In Table 5, we outline the dataset compositions used in training, validation, and testing.

Classification Framework

We present a state-of-the-art multimodal classifier to identify deviant pro-ED content. We discuss our method for obtaining visual features from the images, how we represent their tags, and finally a supervised deep learning framework that incorporated these features in a jointly trained classifier.

Representing Visual Content of Image Posts

Image content representation has evolved greatly with a shift from local features and bag-of-words representations to ag-

gregated features and global image representations [27, 40, 45, 58, 67, 81]. With the recent advances in GPUs and the amount of training data available, the computer vision community has revisited Deep Convolutional Neural Networks (CNNs), *i.e.*, neural networks with many hidden layers and millions of parameters. When trained on big image databases like ImageNet [72], CNNs have been shown to "effortlessly" improve previous state-of-the-art results in many computer vision applications, *e.g.*, image classification [54] and visual search [6, 48, 82]. The aggregated approaches as well as the CNN-based ones all produce a *global* image signature, *i.e.*, a high-dimensional feature vector in Euclidean space.

We extracted visual features of our image posts by starting with the publicly available pre-trained AlexNet model from the Caffe deep learning framework [46]. We experimented with two types of features from this CNN. The $4k$-dimensional $fc7$ features, *i.e.*, the features after the second fully connected layer (we refer the reader to [54] for more details). Limited by the amount of available annotated data for our problem and in pursuit of a more compact representation, we use dimensionality reduction to get the features down to $d = 128$ dimensions. To reduce dimensionality, we use principle component analysis, learned using $fc7$ features from a public 100 Million YFCC100M image set [37, 80]. We define the visual features for an image i as $\mathscr{V}_i \in \mathbb{R}^D$, where $D = \{128, 4096\}$.

Text Embeddings and Aggregation

To extract text features, we used the *skip-gram* model [62] and *word2vec* [63] training. Recent work has used word2vec for tag prediction [28, 87]. The principal force behind the skip-gram model is the use of *context as supervision*. As a learning objective in a generic text representation scenario, skip-gram tries to maximize classification of a word based on another word in the same sentence. Here, we do not target tag or word prediction but choose to *aggregate* the individual learned tag representations of posts given by word2vec into a single compact post representation.

We create our tag contexts through *tag co-occurrencs* and form our supervision signal from all possible pairs of tags that appear in a post. Given a set $T_i = \{t_1, \ldots, t_T\}$ of T_i tags for post i we construct the set of co-occurring tag pairs $\mathscr{P}_i$: $\{(t_j, t_k) \forall j, k \in T_i\}$ with the objective to maximize the average log probability spanning all tag pairs in i [62, 63]. We form our tag dictionary $\mathscr{D}$ using the most common (top-$40K$) tags from our training dataset. After learning the embeddings, each tag in our dictionary is now represented by a dense compact vector, that lies in a space $\mathscr{E} \in \mathbb{R}^E$ of dimensionality $E = 128$, where tags that often co-occur are *close*. Examples of nearest neighbors in the embedded space are shown in Table 6. The word2vec model produces semantically relevant tags in our dataset for ED-related terms: for example, #anorexic co-occurs most strongly with #starving and #anamia, two tags related to pro-ED actions and ideations.

More formally, the model maps every tag $t_j \in \mathscr{D}$ to an embedding vector $e_j \in \mathscr{E}$ and now an image. Photo post i is represented by the set $\mathscr{T}_i = \{e_1, \ldots, e_{T_d}\}$ where now only a subset T_d of all post tags that are present in the dictionary are kept. In order to have one compact tag representation per post, we aggregated the tag embeddings using average pooling, *i.e.* by averaging all vectors per image.

Tag	Closest tags in embedded space
anorexic	starving, thinspriation, overweight, anamia, thinspi
thigh	gap, gab, highs, thighspo, hips
anorexia	bulima, okay, bulimia, anorexyc, skynnny
harm	self, self-harm, injury, tw:, disappointment
ana	collarbones, eddies, anarexic, skinnygirl, anotexia
ed	manorexia, Ed, relapse, warriors, disorder
love	couple, passion, life, followers, kiss
thinspo	thinspire, anabuddy, pretty, thin, thinspiration
bulimic	bulimia, depressed, starve, bulimix, purge

Table 6: Top 5 nearest neighbors in the embedded space.

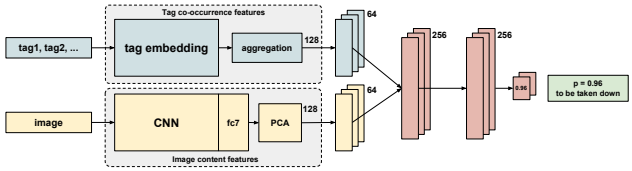


Figure 5: Our multimodal deep neural network architecture.

Learning a Deep Multimodal Neural Network (DNN)

Given a post i with visual features $\mathscr{V}_i$ and textual features $\mathscr{T}_i$, we want to use both to learn whether or not this is deviant pro-ED content. Formulating this as a classification problem, we jointly learn from both modalities. Our joint model outputs a function $f(\mathscr{V}_i, \mathscr{T}_i)$ that models the probability $p(y|i)$ of whether a post is deviant.

Multimodal models have been used for tag prediction [28, 87], image captioning [52] and categorization [36, 70]. Our model follows the same basic structure of [36], where two streams, one per modality, are first independent and then concatenated to learn a joint embedding before the final classifier. However, we divert from the architecture of [36] by adding a fully connected layer on top of the tag embeddings as well as adding multiple layers after concatenation. An overview of the model is shown in Figure 5. Layers in the CNN block follow the AlexNet model [54] while all added layers are fully connected. Although our model can be trained end-to-end, we chose to only learn the last layers of each modality jointly with all subsequent multimodal layers. We trained the model using the Adagrad [32] optimizer and the softmax cross entropy as our loss function. Although the depth and width choices for our model are limited due to limited ground truth data, we experimented with deep architectures with up to 3 joint layers.

RESULTS

Evaluation Protocol and Methods Compared

We compare the Deep Neural Network (DNN) to text-only, image-only, and multimodal Support Vector Machines (SVMs) [23]. In all cases, we use the training, validation and testing datasets presented in Table 5 and the features presented in the previous section, *i.e.* CNN-based visual features $\mathscr{V}_i$ and textual features $\mathscr{T}_i$ from word2vec embeddings. For the multimodal SVM, the two are first concatenated. We experimented with both linear and Radial Basis Function (RBF)

Method	Tags	Vis	Metrics				
			Acc	*P*	*R*	*AUC*	*F1*
SVM	✓		0.89	0.72	0.41	0.61	0.53
SVM		✓	0.85	0.53	0.36	0.49	0.43
SVM	✓	✓	**0.86**	**0.54**	**0.81**	**0.69**	**0.65**
SVM-4k	✓	✓	0.86	0.56	0.50	0.57	0.53
DNN-l1-256	✓	✓	0.90	0.67	0.74	0.72	0.70
DNN-l2-128	✓	✓	0.88	0.57	0.82	0.71	0.67
DNN l2-256	✓	✓	**0.90**	**0.62**	**0.85**	**0.75**	**0.72**
DNN-l2-512	✓	✓	0.90	0.66	0.75	0.73	0.70
DNN l3-256	✓	✓	0.90	0.65	0.75	0.72	0.69

Table 7: Results for the validation dataset. Our best performing SVM and DNN are bolded.

Method	Tags	Vis	Metrics				
			Acc	*P*	*R*	*AUC*	*F1*
SVM	✓		0.88	0.50	0.25	0.42	0.33
SVM		✓	0.73	0.30	0.90	0.60	0.45
SVM	✓	✓	**0.86**	**0.46**	**0.85**	**0.66**	**0.59**
SVM-4k	✓	✓	0.88	0.50	0.46	0.51	0.48
DNN-l1-256	✓	✓	0.90	0.57	0.70	0.65	0.62
DNN-l2-128	✓	✓	0.87	0.49	0.82	0.67	0.61
DNN-l2-256	✓	✓	**0.89**	**0.52**	**0.85**	**0.70**	**0.65**
DNN-l2-512	✓	✓	0.88	0.52	0.60	0.58	0.56
DNN l3-256	✓	✓	0.89	0.54	0.71	0.64	0.61

Table 8: Results for the test dataset. Our best performing SVM and DNN are bolded.

kernels and found performance to be similar. We report results on the classifier trained with the Radial Basis Function (RBF) kernel (C=100, g=0.01). We also report results using the 4K-dimensional visual features for the SVM.

We experimented with different configurations for the DNN, varying the trainable parameters both through the depth (*i.e.* the number of layers added) and width (*i.e.* the size of each layer) of our model. For brevity, we will only present results for the top performing configurations at each depth in Tables 7 and 8. We refer to different configurations of our model as *DNN-lX-Y*, where X stands for the number of joint layers and Y for their size. During training we used a batch size of 64 and a learning rate of 0.01 in most cases. We set the probability threshold of the classifier to $p = 0.5$. We report accuracy (A), precision (P), recall (R), F1-measure and area under the curve (AUC), also known as average precision [7].

Classifier Performance on Validation and Test Data

In Tables 7 and 8, we present results for the SVMs and the DNN over the validation and test sets. We varied model parameters for all methods in the validation set and applied the best performing models at the test set. In this discussion, we report our best-performing models for the multimodal SVM and the DNN-l2-256 models measured by AUC.

In our validation step, the multimodal SVM outperforms both unimodal ones by 8–20%, and the deep models outperform the multimodal SVM by 2–6%. Interestingly, higher dimensionality features (the 4K SVM) did not result in higher performance for the SVM—we hypothesize that this model might be overfitting on the training data with the large feature space. However,

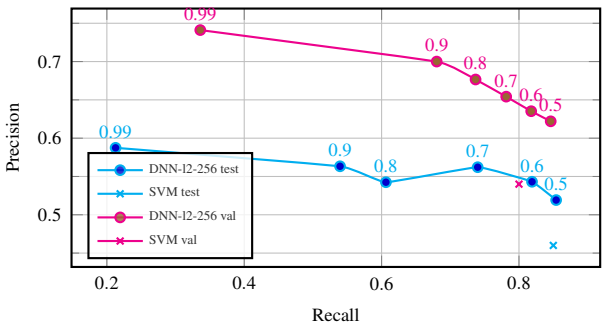


Figure 6: Precision-Recall curve for our classification with the best performing deep (DNN-l2-256) and SVM models.

increasing the deep model's parameter set size makes the model perform better on both the validation and the test sets (3–12% improvement over all other deep models). In Figure 6, we show precision versus recall curves when changing the probability threshold to identify deviant pro-ED content.

We report results on the test set. From Table 8, our DNN outperforms our state-of-the-art SVM. Our DNN l2–256 model achieves an accuracy of 0.89, a precision of 0.52, a recall of 0.85, and an AUC of 0.7. This is a 3% improvement in accuracy over the best SVM (0.86) and a 4% increase in AUC (SVM at 0.66). Our recall is particularly high, indicating that our method is robust against false negatives. What is contributing to the improvement of AUC is a 6% increase in our precision, a significant step in precision. That is, our DNN is able to capture fewer false positives while maintaining recall.

Finally, between the validation and test sets, we see one interesting result: for almost all of the SVM and the deep models, relative performance is preserved between the validation and testing phases. For the best performing SVM, the validation AUC is 0.69 ($F1 = 0.65$) and testing AUC is 0.66 ($F1 = 0.59$). For our DNN, the validation AUC is 0.75 ($F1 = 0.72$) and testing AUC is 0.70 ($F1 = 0.65$). This emphasizes the power of our human-machine method—our hand-curated, expertly rated dataset of pro-ED posts and the posts removed by moderators capture similar visual and textual cues.

Accuracy and Error Analysis

We present a qualitative analysis of example posts and their classification outcomes. Specifically, we identify posts with visual similarity but distinctive contexts and meanings with different classification outcomes. Since we conduct these analyses on images removed from the Tumblr platform, the actual images are not included; however we provide descriptions of their visual and textual content to assist in interpretation.

We begin with a *true positive* ($P(\text{Class}_1) = 0.67$), or example A. In the post, there is a black and white photograph of person laying down that emphasizes the midsection. The midsection of the person is particularly emaciated, and their hip bones and rib cage are very visible. In general, there is a strong emphasis on their bone structure. The image also emphasizes their thinness because the person is wearing a very baggy sweater. The face is not visible. The post has a few tags, like #thinspo. Our classifier correctly marked this post.

Example B is a post in a similar pose that is an example of a *false positive*. This was not taken down from Tumblr, but the classifier felt strongly that it should be removed ($P(\text{Class}_1) = 0.99$). This post is a black and white photo of an individual's midsection laying down. The person is slim but appears at a healthy body weight. The post is tagged with about seven tags, with #pro ana and #thinspo as the most concerning. One reason why the classifier might have been confused is that the pose and coloring is suggestive of common poses seen in the dataset—emphasis on a body part like the hips or ribs. We also noticed that black and white photos were often the most likely candidates for removal, perhaps because they are seen as more somber and sad [69]. However, the person in the image is at what appears to be a healthy bodyfat level, which is why it was likely not removed. The tags are not necessarily motivation for taking actions that promote an eating disorder.

Our next example C is a *true negative* ($P(\text{Class}_1) = 0.04$). This post was not removed from Tumblr, and our classifier agreed with this. This post is a black and white photograph of an individual seated outside smoking a cigarette. The whole body is visible, including the face. The person is slim and the pose highlights their slender legs. The post has several tags, including #fashion and #model. Although this person is slim and seems to be emphasizing her legs, the photo appears to be of a model relaxing. Because the tags confirm this contextualization, we believe that the classifier picked up on this context and was able to correctly identify this image.

The final example D is a *false negative* ($P(\text{Class}_1) = 0.29$). The classifier labeled this post as safe, however this was removed from the platform for violating guidelines. The image contains a color photo of an individual laying on a bed wearing a bra and jeans. The whole body is visible in the photo, including the face. It does not appear that the person has low body fat. At the bottom of the photo, the words "stay strong. starve on" are included. The photo has five tags including #thinspoo and #skinny. There are some elements of the post that the classifier may be misinterpreting. The vast majority of our positive examples were black and white photographs, and it is unusual to have a color photo in this dataset. Further, the photo shows a full body with a face. Most positive examples do not show a person's face. In this case, however, the text overlaid on the image indicates that this image promotes starvation. The deep learning model we built does not incorporate optical character recognition—we found so few images in our rated dataset had text that we did not implement it as a layer in our CNN. This might be why the classifier misclassified this post—it was not able to "read" the text on the photograph to discern that it was promoting starvation.

DISCUSSION

In this paper, we introduced a multimodal classification approach to detect deviant pro-ED Tumblr content. We built a dataset with a combination of human and machine curated examples. We found images are a powerful signal that work with text to characterizing subtle cues of posts in sensitive, complex domains like pro-ED. Importantly, we showed that a deep-learning, convolutional neural network is a powerful method to distinguish deviant content from other content; CNNs outperforms state-of-the art supervised machine learning techniques like SVMs. We demonstrate a novel methodological advancement in machine learning on a problem of broad and profound importance to HCI research and practice.

Our work showed that humans can carefully curate a dataset with similar classification performance to posts removed by Tumblr moderators. This is a powerful discovery for research in deviant behavior communities like pro-ED that suffer from small, challenging, or difficult-to-find datasets. Expert-created datasets can be used in these cases, and algorithms can be developed to detect this kind of content. We are optimistic that researchers of deviant behavior and practitioners may borrow from our method and can tackle problems in these spaces.

Designing Intelligent Moderation Systems

Our multimodal approach tempered with human sensitivity can be a flexible, yet powerful mechanism to address moderation challenges on social media. We developed this method with Tumblr in mind, where there are active human moderators that review user-reported content for community guidelines violations. In practice, we believe our method could be used in human-machine moderation systems, where it could improve moderation *scale*, *efficiency*, and *skill development*:

(1) Our method could be a broad "first pass" for human moderators to scale up their tasks. By surfacing posts that likely break the guidelines, our method could prune the search space of posts that need intervention. Moderators could also focus on violations that are not reported by the platform's users, e.g., due to the clandestine nature of the pro-ED community [17], or are reported once they reach many users.

(2) Building on the above design direction, our method could also be used in an online learning-based moderation system [59]. Moderators could set the decision threshold of the classifier per context and need. They could also adjust the desired balance between false positives and false negatives. This is an important consideration for moderation of sensitive content like pro-ED—moderators of different platforms may want more or less stringency to reflect what their community guidelines consider harmful behaviors. The moderators could provide feedback to our method so it can learn from images that it misclassifies (*i.e.*, re-tune the CNN). These systems have been extremely successful in online text classification and image retrieval [83]. We think this kind of moderation system would improve efficiency and also help the system 'evolve' over time by learning from misclassifications.

(3) Moderation of sensitive content like pro-ED needs a unique set of skills to assess community harm or guideline violations compared to other kinds of content (*e.g.*, salacious content). This skill set is often gained over time, potentially making new moderator training a time-consuming and emotionally intensive process [5]. We believe our method could design novel moderation training systems to help new moderators recognize objectively codified rules from prior deviant content.

Managing Impacts to Current Moderation Practices

Our proposed intelligent moderation systems could increase the posts drawn to the attention of moderators—this has both benefits and drawbacks. For communities that actively use human moderators, these systems inevitably increase the posts that a moderator will need to check. This will likely place

strain on existing moderation practices, might change labor expectations, hiring needs, or other human or capital consequences. There might also be unintended consequences on what were once manually moderated communities. Rampant automated moderation of content perceived to be deviant by a human-machine system could additionally discourage user participation on a platform. It also presents difficult, unique challenges in contexts where content incorrectly marked as deviant is accidentally taken down.

One way to manage these tensions is with the trade-off of precision and recall in the design of an algorithm. Our results had higher recall and were therefore more robust against false negatives. Said another way, our method would pull more potentially deviant content to the attention of moderators. This feature is valuable where our method works alongside the moderators; high recall will expose moderators to more content than they may normally find. However, in an automated context, designers might favor higher precision. This would prevent safe content from being removed from the platform and negatively affecting genuine user participation. Summarily, we believe developing these moderation systems need careful and considerate execution and implementation.

Ethical and Social Challenges

Our work underscores an important ethical and social question—do social networks have an obligation to curate this content on their platforms? Social networks, and communities more broadly, are inherently defined both by the content they permit on the platform as well as the content they remove.

One might argue that social networks should allow as much speech as possible. For issues like copyright infringement, social networks have a legal obligation to remove that content. However, for socially contentious subjects like pro-ED, social networks might not necessarily choose to ban them. In some cases, the community may feel that it is better for these people to identify and express themselves. Other work has also shown that, after banning certain tags on Instagram, the pro-ED community became more insular and focused on more dangerous ideas [17]. There is also promising research that suggests discussing dangerous ideas might help people disinhibit themselves from self-harm [33].

On the other hand, there are those that value moderation of deviant content for its ability to constrain sentiments that might harm individuals or communities. In particular, pro-ED behaviors show contagion-like effects [60], and research has shown that this discourse can encourage others to maintain or continue their dangerous behaviors [71]. Some might argue that, because the social network can shape discourse, this kind of content should be taken down because it is "toxic" to the community [49]. This argument could be extended to other kinds of content. Another argument for moderation is for user engagement—if the negative content causes people to leave or stop participating, the content should be removed.

This is by no means an exhaustive analysis of the benefits and drawbacks to content moderation, nor do we decide whether these platforms have these social obligations. In fact, this only scratches the surface of a complicated set of issues around content moderation, social networks, and deviant behavior like pro-ED. It will take collaborations from industry professionals, researchers, designers, psychologists, and other stakeholders to make decisions in this area.

Limitations and Future Work

One limitation is that we only used public data in this paper and our method did not learn from any private pro-ED content. Because there are few posts identified as actual moderator removals, we could not use them in training our classifiers. While our approach performed equally well on validation and test data, misclassifications may be attributed to the classifier's inability to learn from decisions that drive moderator removals. We could not identify posts that received the attention of a moderator but are still public.

These concerns could be solved in future work with mixed methods working with a social media platform. We could gather information of true positives (removals) and true negatives (posts that were flagged but stay public) to train our models. Researchers could also interview moderators to better understand removal procedures. This was successful in other algorithmic approaches that assist human editors in finding photos [73], and we see this as a promising next step.

Another limitation of our deep learning approach relates to interpretability of the results of the deep learning models. These deep learning models are significantly more powerful; however, they do not produce interpretable results about what features are most important. We recognize that this could present challenges to non-expert end users, like moderators, who could integrate the outputs of this method into their workflow. We believe ongoing research in machine learning interpretability [86] can help resolve these issues in the future.

Finally, despite satisfactory performance, we acknowledge the gap that exists between being able to identify potential guideline violations versus being able to cross-check those assessments with an actual moderator. A future extension could solicit feedback from moderators as an additional way to assess the effectiveness of our proposed method.

CONCLUSION

In this paper, we described a semi-automated method to identify pro-ED posts on Tumblr that violate their community guidelines. Our method jointly incorporated visual and text cues from photo posts to build a deep neural network based classifier to handle this task. Employing a hand-curated dataset of nearly a million pro-ED Tumblr posts, this classifier incorporated human sensitivity by consulting expert ratings on a sample of posts. Our results showed that this deep-learning model performed comparably to ground truth that included actually moderated Tumblr data. Through this work, we demonstrate the power that humans and machines can have when they are used together on a contextually-rich and complicated classification task as identifying deviant pro-ED content.

ACKNOWLEDGMENTS

This work was partly supported through a Yahoo Faculty Engagement Award to De Choudhury.

REFERENCES

1. B Thomas Adler, Luca De Alfaro, Santiago M Mola-Velasco, Paolo Rosso, and Andrew G West. 2011.

Wikipedia vandalism detection: Combining natural language, metadata, and reputation features. In *Computational linguistics and intelligent text processing*. Springer, 277–288.

2. Eugene Agichtein, Carlos Castillo, Debora Donato, Aristides Gionis, and Gilad Mishne. 2008. Finding high-quality content in social media. In *Proceedings of the 2008 international conference on web search and data mining*. ACM, 183–194.
3. Ronald L Akers. 1977. Deviant behavior: A social learning approach. (1977).
4. Nazanin Andalibi and Andrea Forte. 2016. Social Computing Researchers, Vulnerability, and Peer Support. In *Ethical Encounters in HCI: Research in Sensitive and Complex Settings Workshop at the Conference on Human Factors in Computing Systems*.
5. Nazanin Andalibi, Pinar Ozturk, and Andrea Forte. 2017. Sensitive Self-disclosures, Responses, and Social Support on Instagram: the case of #depression. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work & Social Computing*. Forthcoming.
6. Artem Babenko, Anton Slesarev, Alexandr Chigorin, and Victor Lempitsky. 2014. Neural Codes for Image Retrieval. In *ECCV*.
7. Ricardo A. Baeza-Yates and Berthier A. Ribeiro-Neto. 2011. *Modern Information Retrieval - the concepts and technology behind search, Second edition*. Pearson Education Ltd., Harlow, England.
8. Michael S Bernstein, Andrés Monroy-Hernández, Drew Harry, Paul André, Katrina Panovich, and Gregory G Vargas. 2011. 4chan and/b: An Analysis of Anonymity and Ephemerality in a Large Online Community.. In *ICWSM*.
9. Jeremy Blackburn and Haewoon Kwak. 2014. Stfu noob!: predicting crowdsourced decisions on toxic behavior in online games. In *Proceedings of the 23rd international conference on World wide web*. ACM, 877–888.
10. Dina LG Borzekowski, Summer Schenk, Jenny L Wilson, and Rebecka Peebles. 2010. e-Ana and e-Mia: A Content Analysis of Pro–Eating Disorder Web Sites. *American journal of public health* 100, 8 (2010), 1526.
11. Amy Bruckman, Pavel Curtis, Cliff Figallo, and Brenda Laurel. 1994. Approaches to managing deviant behavior in virtual communities. In *Conference Companion on Human Factors in Computing Systems*. ACM, 183–184.
12. Erin E Buckels, Paul D Trapnell, and Delroy L Paulhus. 2014. Trolls just want to have fun. *Personality and individual Differences* 67 (2014), 97–102.
13. Alissa Centivany and Bobby Glushko. 2016. 'Popcorn Tastes Good': Participatory Policymaking and Reddit's 'Amageddon". In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*.
14. Stevie Chancellor, Zhiyuan Lin, Erica L Goodman, Stephanie Zerwas, and Munmun De Choudhury. 2016b. Quantifying and Predicting Mental Illness Severity in Online Pro-Eating Disorder Communities. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*. ACM, 1171–1184.
15. Stevie Chancellor, Zhiyuan Jerry Lin, and Munmun De Choudhury. 2016a. "This Post Will Just Get Taken Down": Characterizing Removed Pro-Eating Disorder Social Media Content. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 1157–1162.
16. Stevie Chancellor, Tanushree Mitra, and Munmun De Choudhury. 2016c. Recovery Amid Pro-Anorexia: Analysis of Recovery in Social Media. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 2111–2123.
17. Stevie Chancellor, Jessica Pater, Trustin Clear, Eric Gilbert, and Munmun De Choudhury. 2016d. #thyghgapp: Instagram Content Moderation and Lexical Variation in Pro-Eating Disorder Communities. In *Proceedings of the 2016 Conference on Computer Supported Cooperative Work & Social Computing(CSCW)*. ACM.
18. Yi Chang, Lei Tang, Yoshiyuki Inagaki, and Yan Liu. 2014. What is tumblr: A statistical overview and comparison. *ACM SIGKDD Explorations Newsletter* 16, 1 (2014), 21–29.
19. Adrien Chen. 2014. The Laborers Who Keep Dick Pics and Beheadings Out of Your Facebook Feed. (2014). `https://www.wired.com/2014/10/content-moderation/`
20. Justin Cheng, Michael Bernstein, Cristian Danescu-Niculescu-Mizel, and Jure Leskovic. 2017. Anyone Can Become a Troll: Causes of Trolling Behavior in Online Discussions. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work & Social Computing*. Forthcoming.
21. Justin Cheng, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. 2015. Antisocial Behavior in Online Discussion Communities. In *International Conference on Weblogs and Social Media (ICWSM)*. AAAI.
22. Denzil Correa and Ashish Sureka. 2014. Chaff from the wheat: characterization and modeling of deleted questions on stack overflow. In *Proceedings of the 23rd international conference on World wide web*. ACM, 631–642.
23. Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning* 20, 3 (1995), 273–297.
24. John P Davis, Shelly Farnham, and Carlos Jensen. 2002. Decreasing online 'bad' behavior. In *CHI'02 Extended Abstracts on Human Factors in Computing Systems*. ACM, 718–719.
25. Munmun De Choudhury. 2015. Anorexia on Tumblr: A Characterization Study. In *Proc. Digital Health*.

26. Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. Predicting depression via social media. In *AAAI Conference on Weblogs and Social Media*.
27. J. Delhumeau, PH. Gosselin, H. Jegou, and P. Perez. 2013. Revisiting the VLAD Image representation. In *ACM Multimedia*.
28. Emily Denton, Jason Weston, Manohar Paluri, Lubomir Bourdev, and Rob Fergus. 2015. User conditional hashtag prediction for images. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 1731–1740.
29. Nicholas A Diakopoulos. 2015. The Editor's Eye: Curation and Comment Relevance on the New York Times. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*. ACM, 1153–1157.
30. Judith S Donath and others. 1999. Identity and deception in the virtual community. *Communities in cyberspace* 1996 (1999), 29–59.
31. Harris Drucker, Donghui Wu, and Vladimir N Vapnik. 1999. Support vector machines for spam categorization. *IEEE Transactions on Neural networks* 10, 5 (1999), 1048–1054.
32. John Duchi, Elad Hazan, and Yoram Singer. 2011. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research* 12, Jul (2011), 2121–2159.
33. T Emmens and A Phippen. 2010. Evaluating Online Safety Programs. *Harvard Berkman Center for Internet and Society.[23 July 2011]* (2010).
34. Casey Fiesler, Cliff Lampe, and Amy S Bruckman. 2016. Reality and perception of copyright terms of service for online content creation. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*. ACM, 1450–1461.
35. Jesse Fox and Margaret C Rooney. 2015. The Dark Triad and trait self-objectification as predictors of men's use and self-presentation behaviors on social networking sites. *Personality and Individual Differences* 76 (2015), 161–165.
36. Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Tomas Mikolov, and others. 2013. Devise: A deep visual-semantic embedding model. In *Advances in neural information processing systems*. 2121–2129.
37. Pierre Garrigues, Sachin Farfade, Hamid Izadinia, Kofi Boakye, and Yannis Kalantidis. 2016. Tag Prediction at Flickr: a View from the Darkroom. *arXiv preprint arXiv:1612.01922* (2016).
38. Eric Gilbert. 2013. Widespread underprovision on reddit. In *Proceedings of the 2013 conference on Computer supported cooperative work*. ACM, 803–808.
39. Val Gillies, Angela Harden, Katherine Johnson, Paula Reavey, Vicky Strange, and Carla Willig. 2005. Painting pictures of embodied experience: The use of nonverbal data production for the study of embodiment. *Qualitative research in psychology* 2, 3 (2005), 199–212.
40. Philippe-Henri Gosselin, Naila Murray, Hervé Jégou, and Florent Perronnin. 2014. Revisiting the Fisher vector for fine-grained classification. *Pattern Recognition Letters* 49 (2014), 92–98.
41. Lynne Hall and Carlisle E George. 1999. Law and Punishment in Virtual Communities. *Proceedings of Cybersociety* (1999).
42. Sameer Hinduja and Justin W Patchin. 2014. *Bullying beyond the schoolyard: Preventing and responding to cyberbullying*. Corwin Press.
43. Yuheng Hu, Lydia Manikonda, Subbarao Kambhampati, and others. 2014. What We Instagram: A First Analysis of Instagram Photo Content and User Types.. In *ICWSM*.
44. Sara Javanmardi, David W McDonald, and Cristina V Lopes. 2011. Vandalism detection in Wikipedia: a high-performing, feature-rich model and its reduction through Lasso. In *Proceedings of the 7th International Symposium on Wikis and Open Collaboration*. ACM, 82–90.
45. H. Jégou, M. Douze, C. Schmid, and P. Perez. 2010. Aggregating Local Descriptors into a Compact Image Representation. In *CVPR*.
46. Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. 2014. Caffe: Convolutional Architecture for Fast Feature Embedding. In *Proceedings of the 22Nd ACM International Conference on Multimedia (MM '14)*. ACM, New York, NY, USA, 675–678. DOI: http://dx.doi.org/10.1145/2647868.2654889
47. Yannis Kalantidis, Lyndon Kennedy, Huy Nguyen, Clayton Mellina, and David A Shamma. 2016a. LOH and behold: Web-scale visual search, recommendation and clustering using Locally Optimized Hashing. *ECCV VSM Workshop* (2016).
48. Yannis Kalantidis, Clayton Mellina, Flickr Vision, and Simon Osindero. 2016b. Cross-dimensional Weighting for Aggregated Deep Convolutional Features. In *VSM Workshop, ECCV*.
49. Ruogu Kang, Laura Dabbish, and Katherine Sutton. 2016. Strangers on Your Phone: Why People Use Anonymous Communication Applications. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*. ACM, 359–370.
50. Sara Kiesler, Robert Kraut, Paul Resnick, and Aniket Kittur. 2012. Regulating behavior in online communities. *Building Successful Online Communities: Evidence-Based Social Design. MIT Press, Cambridge, MA* (2012).

51. Amy Jo Kim. 2000. *Community building on the web: Secret strategies for successful online communities*. Addison-Wesley Longman Publishing Co., Inc.
52. Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. 2014. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539* (2014).
53. Aniket Kittur, Boris Smus, Susheel Khamkar, and Robert E Kraut. 2011. Crowdforge: Crowdsourcing complex work. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*. ACM, 43–52.
54. Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *NIPS*.
55. Cliff Lampe and Paul Resnick. 2004. Slash (dot) and burn: distributed moderation in a large online conversation space. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 543–550.
56. A Laye-Gindhu and KA Schonert-Reichl. 2005. Nonsuicidal self-harm among community adolescents: Understanding the "whats" and "whys" of self-harm. *Journal of Youth & Adolescence* 34, 5 (2005), 447–457.
57. Stephanie M. Lee. 2016. Why Eating Disorders Are So Hard For Instagram And Tumblr To Combat. (2016). `https://www.buzzfeed.com/stephaniemlee/why-eating-disorders-are-so-hard-for-instagram-and/-tumblr-to`
58. D.G. Lowe. 2004. Distinctive Image Features from Scale-Invariant Keypoints. *IJCV* 60, 2 (2004), 91–110.
59. Justin Ma, Lawrence K Saul, Stefan Savage, and Geoffrey M Voelker. 2009. Identifying suspicious URLs: an application of large-scale online learning. In *Proceedings of the 26th annual international conference on machine learning*. ACM, 681–688.
60. Jeanne B Martin. 2010. The development of ideal body image perceptions in the United States. *Nutrition Today* 45, 3 (2010), 98–110.
61. J. Nathan Matias, Amy Johnson, Whitney Erin Boesel, Brian Keegan, Jaclyn Friedman, and Charlie DeTar. 2015. Reporting, Reviewing, and Responding to Harassment on Twitter. *CoRR* abs/1505.03359 (2015). `http://arxiv.org/abs/1505.03359`
62. Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).
63. T Mikolov and J Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems* (2013).
64. Arjun Mukherjee, Vivek Venkataraman, Bing Liu, and Natalie S Glance. 2013. What yelp fake review filter might be doing?. In *ICWSM*.
65. Deokgun Park, Simranjit Sachar, Nicholas Diakopoulos, and Niklas Elmqvist. 2016. Supporting comment moderators in identifying high quality online news comments. In *Proc. Conference on Human Factors in Computing Systems (CHI)*.
66. Jessica A Pater, Oliver L Haimston, Nazanin Andalibi, and Elizabeth D Mynatt. 2016. "Hunger Hurts but Starving Works": Characterizing the Presentation of Eating Disorders Online. In *Proceedings of the 19th ACM conference on Computer Supported Cooperative Work & Social Computing (CSCW)*.
67. F. Perronnin, Y. Liu, J. Sanchez, and H. Poirier. 2010. Large-Scale Image Retrieval with Compressed Fisher Vectors. In *CVPR*.
68. Rebecca Rafferty and Thomas Vander Ven. 2014. "I Hate Everything About You": A Qualitative Examination of Cyberbullying and On-Line Aggression in a College Sample. *Deviant behavior* 35, 5 (2014), 364–377.
69. Andrew G Reece and Christopher M Danforth. 2016. Instagram photos reveal predictive markers of depression. *arXiv preprint arXiv:1608.03282* (2016).
70. Zhou Ren, Hailin Jin, Zhe Lin, Chen Fang, and Alan Yuille. 2016. Joint Image-Text Representation by Gaussian Visual-Semantic Embedding. In *ACM Multimedia*, Vol. 4.
71. Denise Restauri. 2012. "Tumblr to Pinterest to Instagram – The Self-Harm 'Thinspo' Community Is House-Hunting". (2012). `http://www.forbes.com/sites/deniserestauri/2012/04/16/tumblr-to-pinterest-to-instagram-self-harm-thinspo/-community-is-house-hunting`
72. Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, and others. 2014. Imagenet large scale visual recognition challenge. *arXiv preprint arXiv:1409.0575* (2014).
73. David A Shamma, Lyndon Kennedy, Jia Li, Bart Thomee, Haojian Jin, and Jeff Yuan. 2016. Finding Weather Photos: Community-Supervised Methods for Editorial Curation of Online Sources. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*. ACM, 86–96.
74. Manya Sleeper, Justin Cranshaw, Patrick Gage Kelley, Blase Ur, Alessandro Acquisti, Lorrie Faith Cranor, and Norman Sadeh. 2013. I read my Twitter the next morning and was astonished: A conversational perspective on Twitter regrets. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 3277–3286.
75. Brian K Smith, Jeana Frost, Meltem Albayrak, and Rajneesh Sudhakar. 2006. Facilitating narrative medical discussions of type 1 diabetes with computer visualizations and photography. *Patient Education and Counseling* 64, 1 (2006), 313–321.

76. Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y Ng. 2008. Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *Proceedings of the conference on empirical methods in natural language processing*. Association for Computational Linguistics, 254–263.
77. Janet Sternberg. 2012. *Misbehavior in cyber places: The regulation of online conduct in virtual communities on the Internet*. Rowman & Littlefield.
78. John Suler. 2004. The online disinhibition effect. *Cyberpsychology & behavior* 7, 3 (2004), 321–326.
79. John R Suler and Wende L Phillips. 1998. The bad boys of cyberspace: Deviant behavior in a multimedia chat community. *CyberPsychology & Behavior* 1, 3 (1998), 275–294.
80. Bart Thomee, Benjamin Elizalde, David A Shamma, Karl Ni, Gerald Friedland, Douglas Poland, Damian Borth, and Li-Jia Li. 2016. YFCC100M: The new data in multimedia research. *Commun. ACM* 59, 2 (2016), 64–73.
81. Giorgos Tolias, Yannis Avrithis, and Hervé Jégou. 2015. Image Search with Selective Match Kernels: Aggregation Across Single and Multiple Images. *International Journal of Computer Vision* (2015), 1–15.
82. Giorgos Tolias, Ronan Sicre, and Hervé Jégou. 2016. Particular object retrieval with integral max-pooling of CNN activations. In *ICLR*.
83. Simon Tong and Edward Chang. 2001. Support vector machine active learning for image retrieval. In *Proceedings of the ninth ACM international conference on Multimedia*. ACM, 107–118.
84. Khoi-Nguyen Tran and Peter Christen. 2015. Cross-language learning from bots and users to detect vandalism on wikipedia. *IEEE Transactions on Knowledge and Data Engineering* 27, 3 (2015), 673–685.
85. Tumblr. 2016. "Tumblr Community Guidelines". (2016). `https://www.tumblr.com/policy/en/community`
86. Alfredo Vellido, José David Martín-Guerrero, and Paulo JG Lisboa. 2012. Making machine learning models interpretable.. In *ESANN*, Vol. 12. Citeseer, 163–172.
87. Jason Weston, Sumit Chopra, and Keith Adams. 2014. # TagSpace: Semantic embeddings from hashtags. (2014).
88. Elad Yom-Tov, Luis Fernandez-Luque, Ingmar Weber, and P Steven Crain. 2012. Pro-Anorexia and Pro-Recovery Photo Sharing: A Tale of Two Warring Tribes. *J Med Internet Res* (2012).