Understanding and Discovering Deliberate Self-harm Content in Social Media

Yilin Wang[1]*, Jiliang Tang[2], Jundong Li[1], Baoxin Li[1], Yali Wan[3],
Clayton Mellina[4], Neil O'Hare[4], Yi Chang[5]
[1] Arizona State Univerity, AZ, USA [2] Michigan State University, MI, USA
[3] University of Washington, WA, USA [4]Yahoo Research, CA, USA
[5]Huawei Research America, CA, USA
{ ywang370,jundong.li,baoxin.li }@asu.edu, tangjili@msu.edu
yaliwan@uw.edu, yichang@acm.org, {nohare,clayton}@yahoo-inc.com

ABSTRACT

Studies suggest that self-harm users found it easier to discuss self-harm-related thoughts and behaviors using social media than in the physical world. Given the enormous and increasing volume of social media data, on-line self-harm content is likely to be buried rapidly by other normal content. To enable voices of self-harm users to be heard, it is important to distinguish self-harm content from other types of content. In this paper, we aim to understand self-harm content and provide automatic approaches to its detection. We first perform a comprehensive analysis on self-harm social media using different input cues. Our analysis, the first of its kind in large scale, reveals a number of important findings. Then we propose frameworks that incorporate the findings to discover self-harm content under both supervised and unsupervised settings. Our experimental results on a large social media dataset from Flickr demonstrate the effectiveness of the proposed frameworks and the importance of our findings in discovering self-harm content.

Keywords

Mental Health; User Modeling; Computational Health; Multimodal Data Mining; Social Media Mining

1. INTRODUCTION

A central challenge in public health revolves around how to identify individuals who are at risk for taking their own lives. Deliberate self-injury is a behavior that some people use to cope with difficulties or painful feelings, and it has become the second leading cause of death for young people aged 15 to 19 years, and the tenth leading cause of death among those aged 10 to 14 [31]. It has been reported by the National Alliance on Mental Illness [1] that there are around 2 million young adults and teenagers who have injured themselves. Another research from Britain [17] reported that among 400 pupils aged 14∼16, more than 6.5% confirmed they harmed themselves in the last year. Self-harm prevention is challenging since it is a multi-faceted problem, with different categories of self-harm behaviors due to different social/personal reasons, pathogenesis, and/or underlying illnesses [15].

Existing efforts toward discovering and caring self-harm people. especially adolescents, have primarily relied on self report or friends/family [31, 17, 15]. However, such efforts face tremendous methodological challenges. First, self-harm people often find it difficult to discuss their feelings [14] and that is why they use self-harm to express their emotions. Most self-harm people suffer depression, anxiety or other mental health issues[2] which make the self-harm behavior difficult to be discovered by their friends/families [21]. Second, although it is estimated that 7%-14% of adolescents may inflict self-harm at some time in their lives, and 20%-45% of older adolescents have been reported to have suicidal thoughts at some time[16], the relatively rare occurrence of completed self-harm treatment and the stigma associated with self-harm reports have made studies challenging and expensive to conduct. In addition, extremely long follow-up intervals are typically required for effective study. As consequence, there are limited research efforts on examining factors associated with the development of future self-harm thoughts among self-harm-prone people [16].

Nowadays people are increasingly using social media platforms, such as Twitter and Flickr, to share their thoughts and daily activities. The ubiquity of smart phones/tablets has also made such sharing easy and often instantaneous. As a result, social media provides a means to capture behavioral attributes that are relevant to an individual's thinking, mood, personal and social activities, and so on. In the physical world, people in need of help on mental issues usually do not know who to ask for help, and they often afraid that their trust could be betrayed, or they fear that asking for help may lead to more problems for themselves [19, 16]. On the other hand, they could be very active and open on social media when it comes to communication of the self-harm problem [10]. Given the enormous volume of social media data that is

*This work was done when the first author was on an internship at Yahoo! Research.

WWW 2017, April 3–7, 2017, Perth, Australia.
ACM 978-1-4503-4913-0/17/04.
http://dx.doi.org/10.1145/3038912.3052555

[1]www.nami.org
[2]http://www.mayoclinic.org/diseases-conditions/self-injury/symptoms-causes/dxc-20165427

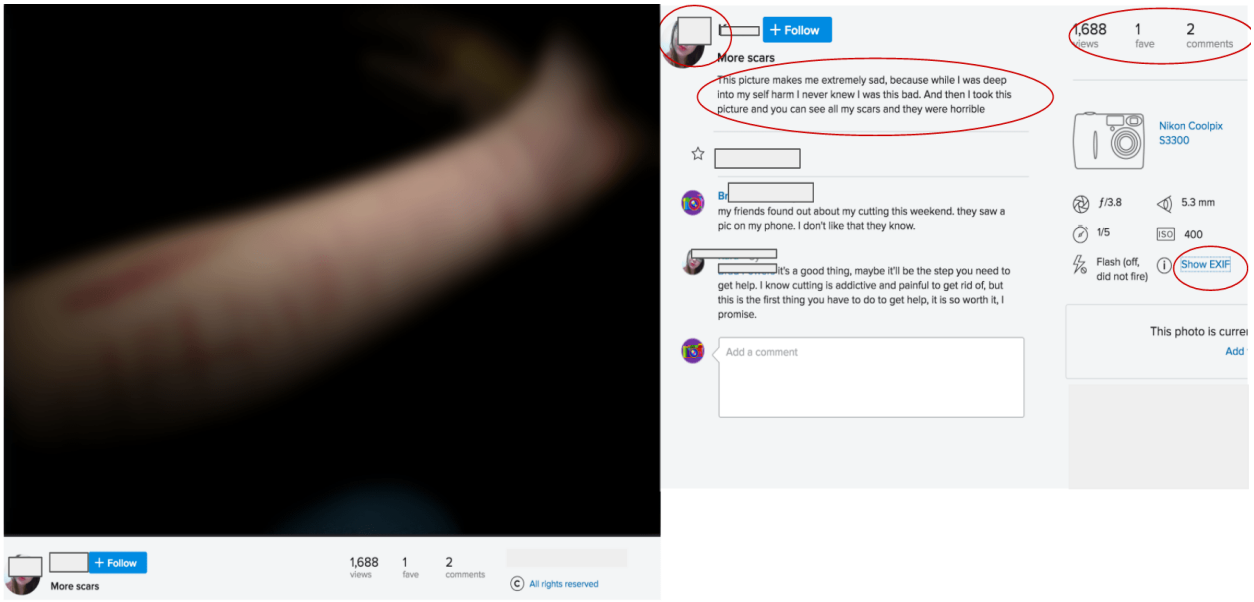


Figure 1: An example of self-harm posts from Flickr. Due to the privacy issue, we blurred the visual content in this post.

created daily, a crucial step to enable the voices of self-harm users to be heard is to identify self-harm content that could be buried by the vast amount of normal content. A self-harm post from Flickr is demonstrated in Figure 1, which consists of multiple sources including text, photo, temporal information and meta information of its owner (highlighted by red circles in Figure 1). It appears that, while this problem has not been well-studied before, the rich information in social media posts may provide unprecedented opportunities for us to understand self-harm content.

In this paper, we aim to understand and discover self-harm content in social media. To achieve this goal, we need to (1) reveal the distinct characteristics of self-harm content from normal content; and (2) leverage these characteristics to build models to automatically discover self-harm content. We conduct a comprehensive analysis on self-harm social media content using textual, owner-related, temporal and visual information, and our major understandings are summarized as: (1) The language of self-harm content has different structures compared with normal content, and the self-harm content expresses much more negative sentiments; (2)On average, owners of self-harm content are likely to have more activities, more social responses and less online friends compared to owners of normal content; (3) Posting time of self-harm content presents hourly patterns different from those of normal content, and self-harm content is likely to be posted during the night especially late night; and (4) Photos in self-harm content are more gloomy and tend to focus on the salient body image patterns. In summary, the key contributions of this work are:

- **Findings**: A first and comprehensive study on deliberate self-harm posts on social media by analyzing more than 1 billion posts on Flickr. We find that the self-harm users have different patterns on social media platform on: language structure and usage, online activity, temporal variation and visual content preference.
- **Applications**: We develop a scalable framework that can discover self-harm content automatically for both supervised and unsupervised scenarios. The features from our findings are used to boost the prediction performance. The solution we developed may be used for public health monitoring and/or directly helping the self-harm users by providing advices when they post self-harm related content.

The rest of the paper is organized as follows. In Section 2, we conduct analysis to understand self-harm content. Section 3 details the proposed framework. In Section 4, we present detailed experiments with discussion. We briefly review related work in Section 5. Section 6 briefly talks about privacy and ethics of the studied problem. Section 7 concludes the work with future research directions.

2. DATA ANALYSIS

In this section, we first introduce the dataset for our study and then perform analysis to understand self-harm-related social media content.

2.1 Data

In this investigation, we use data from Flickr which is one of the largest image hosting websites owned by Yahoo! Inc. In Flickr, users can upload images along with short textual descriptions and tags as a post to share with others. In addition to posting some content, users can also engage in different interest groups.

Since the user's contact information is private, in order to avoid user bias, we collect data from Flickr by checking the visual content of the posts. For our initial data collection, we adopted an approach used in prior work on examination of eating disorders and anorexia in social media sites [1]. We first examine more than 1 billion public Flickr posts and select those public posts that annotated with "selfharm" and "selfinjury" tags. It results $15,729$ posts from $3,328$ distinct users. Then five experienced researchers manually check

2000 random selected Flickr posts that annotated with self-harm or self-injury tags. Based on the snowball sampling approach [13] during the inspection phase, we obtain an initial 30 tags of the highest frequencies along with selfharm content[3]. Some examples include "selfhate", "suicide", "depressed", etc. By removing common tags such as wounds, scares, cut, we use 15 seed tags as shown in Table 1 to further retrieve posts from Flickr. In this stage, we want collect self-harm data as a complementary of first stage and it may help to make the bias as small as possible. For example "secretsociety123" and its variations are widely used by self-harm users [30] .

During the process, we collect 383,614 Flickr posts from 63,949 users. According to the findings from prior work on expression of self-harm tendencies in social media, frequently used tags can be a strong indication that a user has mental issues [1, 30, 49]. Therefore, to obtain a set of relatively reliable self-harm users, we remove users who use selfharm related tags in less than five posts[4], resulting in 93,286 self-harm posts from 20,495 potential self-harm users. Also, we collect a set of 19,720 users from YFCC dataset [39], which is a 100 million open access dataset published by Flickr. We check all the historical posts of these 19,720 users to ensure that their posts do not contain any self-harm related content. We refer these users as normal users. We randomly sample 93,286 normal posts from these normal users for the following analysis. For each user, we crawled some statistical information such as the number of total posts and their user profiles; while for each post, we collected its associated information including the photos, the textual descriptions, user id of the owner, tags, and timestamps when the photo was taken and when the post was uploaded. Finally, we evaluate whether posts in the dataset contain signs of self-harm. Five experienced researchers familiar with social media and selfharm content evaluate the the correctness of the aforementioned method. In particular, each researcher randomly checked the posts and found that 95% of the posts with 'self-harm' and 'selfinjury' are correctly identified; while 83% of other tags are correctly identified. The Cohen kappa coefficient [5] is 0.85 which suggests the high rate of agreement on our data collection method.

eatingdisorder	suicide	anxious	anorexia
mental-illness	depressed	killme	depression
selfhate	anamia	anxious	secretesociety123
bruised	bulimia	bleeding	

Table 1: A set of extended tags that help identify selfharm posts.

2.2 Understanding Self-harm Content

A typical flickr post contains information from four dimensions including textual, owner, visual and temporal information. Therefore we analyze self-harm content from these four perspectives.

Textual Analysis. Linguistic style in texts is related to an individual's underlying psychological and cognitive states. It can reveal cues about their social coordination [35, 7, 38, 45, 44]. Therefore we compute the distributions of nouns, verbs and adverbs in texts of social media posts, including the descriptions, titles and comments via the CMUTweetTagger [12]. Also, we calculate readability scores to estimate the complexity and readability[5] of texts. Individuals in self-harm or depression conditions trend to use negative words or express negative sentiments. Therefore, we compute the sentiment polarities of texts based on an off-the-shelf manually labeled sentiment lexicon, i.e., MPQA subjective lexicon, for self-harm and normal content, respectively. The comparison between self-harm and normal content is shown in Table 2.

	Self-harm	Normal
Linguistic		
Nouns	0.158	0.268
Verbs	0.127	0.021
Adjective	0.035	0.084
Adverbs	0.032	0.023
readability	0.41	0.69
Sentiment		
Positive	0.06	0.29
Neutral	0.15	0.53
Negative	0.79	0.18

Table 2: Textual Analysis (the number stands for the ratio).

Theme	**Token**
Expression/ Symptom	anamia, anorexia, suicide, alone, stress, pretty, harms, stress, pain, angry, addiction, failure, beautiful, peace, illness, bulimic, individual, depressive, disorder
Disclosure	cuts, help, kill, live, die, plans, inflicted, treatments, eating, celebrates, suffer, saveme, triggers
Relationship/Noun	365days, razor, scar , blood, arms, wrist, band, knife, bathroom, bath, tattoo, girls, woman, boyfriend, people, body, night

Table 3: Unigrams from self-harm lexicon that appear with high frequencies in the self-harm content.

From Table 2, we can observe that self-harm content tends to include more verbs and adjectives/adverbs than nouns which is very consistent with suicidal word usage [38]. The average readability score of self-harm content is lower than normal content. The poor linguistic structure usage and language suggest the decreased cognitive functioning and coherence [33]. Further, in addition to less usage of nouns, we note that a large portion of negative sentiment words are used in self-harm content. Such observation shows lower interests in objects and things from owners of self-harm content. It is also well known to appear for suicide users [32].

There is no lexicon to understand the usage of self-harm related terms in social media. Therefore, we build a lexicon of terms that are likely to appear in the texts of self-harm content. We first extract each term in texts and after post-processing each term, we calculate its vector via word2vec [28] and cluster all the terms. Thereafter, we de-

[3]The post contains intentional, direct injuring of body tissue content

[4]If only few self-harm related posts, the user could post them by chance

[5]https://pypi.python.org/pypi/textstat/0.2

ploy the lexicon to calculate the frequencies of terms in self-harm content. In Table 3, we report sample unigrams from the self-harm lexicon. From Table 3, we observe that most captured expression/symptom terms indicate actions on eating habits, relations with others and sleeping. These are known to be correlated with sensitive disclosures [19]. Owners of self-harm content more frequently use action words such as "help", "treatments", and "plans", and entity words such as "people", "girls", and "woman". These observations suggest that the self-harm users turn to social media to communicate and share the experiences with others in order to seek for help or attract attention from others.

Tags are a special type of textual information. We visualize most frequently used tags in self-harm content as shown in Figure 2. Despite the fact that most of the tags are self-harm related, some tags such as "secret society 123", "triggerwarning", and "svv", are not explicitly related to self-harm but indicate self-harm content. Similar to eating disorder [1], self-harm users are likely to use some group tags that are merely used and can only be understood by themselves. To further verify these tags, we search these tags on Instagram and we find each of these tags returns lots of self-harm content[6].

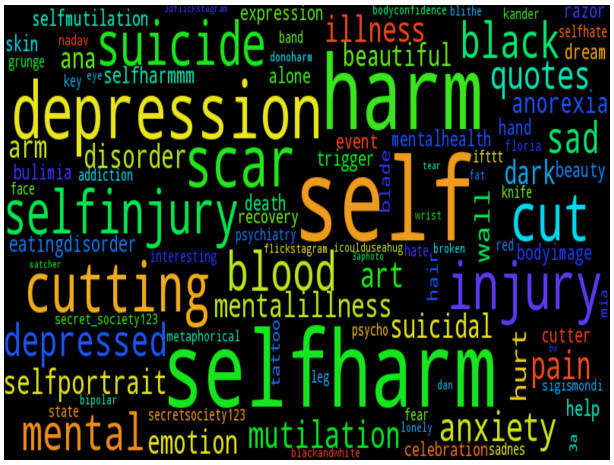


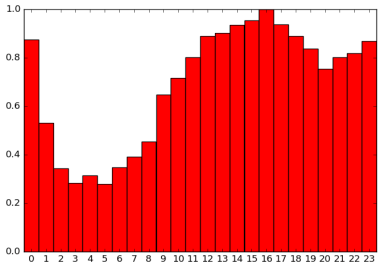
Figure 2: Tag Cloud for Self-harm Content

Owner Analysis. The owners of self-harm content provide crucial context to understand self-harm content. We analyze behaviors of owners from the following perspectives [7] – volume, proportion of reply, number of favorites, and number of friends. The volume is defined as the normalized number of posts per day by the owner. Proportion of reply, number of favorites and number of friends from a user suggest the level of social interactions with other users. The results are shown in Table 4.

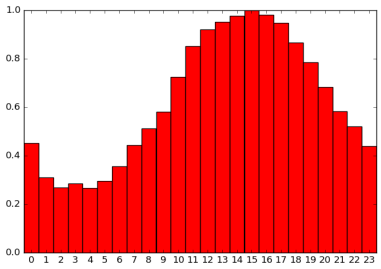
	Volume	% of reply	# favorite	# friends
Self-harm	7.76	0.15	0.56	296.89
Normal	3.79	0.11	0.23	477.57

Table 4: Owner Analysis.

[6]The retrieving results with the tag "svv",for example, can be found via `https://www.instagram.com/explore/tags/svv/`



(a) Self-harm related Content



(b) Normal Content.

Figure 3: Temporal analysis. Y axis represents the normalized portions of data volume of each hour. X axis is the time segment, which ranges from [0 23].

From the table, we first note that the average volume, proportion of reply and number of favorite of owners with self-harm content are much higher than those of normal content. High volume indicates that potential self-harm users are likely to be more active than normal users in social media – they desire the public to hear their voices for help and are likely to use social media to express the emotion and satisfy self-esteem. High proportion of reply and number of favorites suggest that content from potential self-harm users are likely to attract more social responses. In addition, the owners of self-harm content are likely to have fewer number of friends than normal users.

Temporal Analysis. People with mental issues could suffer from insomnia; and they may present different temporal patterns from normal users in terms of their on-line activities. For each self-ham post, we first obtain the local time information on when the post is published, and then count the number of self-harm posts in each hour of a day. The number distributions of self-harm content over hours of a day is demonstrated in Figure 3a. Following a similar process, the distributions of normal content is shown in Figure 3b. Note that the numbers in the Figures are normalized to (0,1] for better visualization. Note that, in this study, we cluster the posts by examining the EXIF data from the user upload images, which accurately records the time when the image is taken.

For normal content, in general, they are more likely to be published during the daytime instead of night. In particular, (1) fewer number is published later in the night (i.e., post-midnight) and early in the morning; (2) the number generally increases through the day; and (3) afternoon and early night show peaks. For self-harm content, a large number is posted during nights especially late in the night (22pm

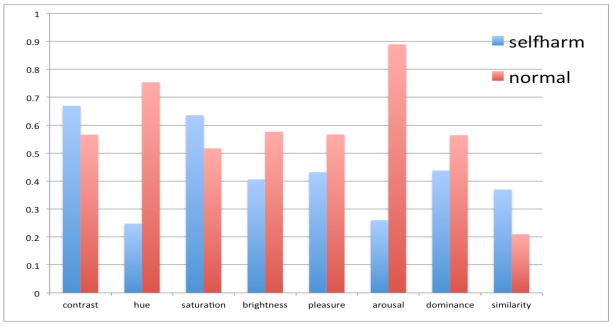


Figure 4: Visual analysis

to 1am), while fewer number in the morning (7am to 8 am). As mentioned earlier, people with mental issues could suffer from insomnia and their mood tends to worsen during the night [26].

Visual analysis. Color patterns are important cues to understand the emotion and affective value of a picture [21]. Therefore, we first compute a global contrast metric [4] that: (1) provides saliency information from the distinguish ability of colors based on the magnitude of the average luminance; and (2) exposes the image regions that are more likely to grasp the attention of the human eyes. Then we extract the average of the Hue, Saturation, Brightness (H,S,V) channels. By combining average Saturation (S) and Brightness (V) values, we also extract three indicators of emotional dimensions, i.e., pleasure, arousal and dominance, as suggested by previous work on affective image analysis [27]. Last, we extract some local image features including SIFT, LBP and GIST [40], which are widely used in image matching and visual search related tasks. Based on these features, we calculate the average similarities for photos in self-harm and normal content, separately. The comparison results are presented in Figure 4. Note that each score in the figure is normalized to (0,1] for visualization.

In general, photos in the self-harm content have lower average values in brightness, pleasure, arousal and dominance. As suggested by the previous findings [37, 44], lower values tend to express more negative sentiments. The higher global contrast demonstrates that photos in self-harm content have higher saliency value regions, e.g., body parts, possibly for attracting attention [20]. Photos in self-harm content are much more similar to each other than those in normal content. This observation suggests that the visual content of self-harm images may have more unique patterns, such as leg and wrist, and are less diverse than that in normal content.

3. SELF-HARM CONTENT PREDICTION

Our findings in the previous section indicate that potential self-harm users inclined to express their feelings and emotions in social media, with the purpose of seeking for help and attention. Social media content is generated at an unprecedented rate, and self-harm content is likely to be buried by the majority of normal content. Hence a crucial step to help their voices be heard by the public is to identify self-harm content. A social media post consists of multiple types of information. As suggested by our previous analysis, each type provides useful and complementary patterns to characterize self-harm content. Therefore combining multiple sources could provide a more comprehensive view about social media posts and has the potential to improve performance.

Typically supervised methods[47, 25, 44] can achieve better performance because the label information can guide the learning performance. However, most social media posts are unlabeled and annotating their labels is expensive and time consuming. Therefore an unsupervised method is also desired. In the following subsections, we will introduce frameworks to discover self-harm content automatically with and without labeled data. Before presenting the details, we first introduce the notations and definitions we will use in the proposed frameworks.

Let $\mathcal{P} = \{p_1, p_2, p_3, ..., p_n\}$ be a set of posts where n is the number of social media posts. Assume that the set of posts i.e., $\mathcal{P}$, can be represented by m heterogeneous feature spaces corresponding to m available sources. For Flickr posts in the studied dataset, m is 4 including textual, owner, temporal and visual sources. Let $\mathcal{F} = \{f_1, f_2, f_3, ..., f_m\}$ be a set of m feature spaces where $f_i \in \mathbb{R}^{l_i}$ denotes the feature space for the i-th source and l_i is the number of features in f_i. We use $\mathcal{X} = \{X_i \in \mathbb{R}^{n \times l_i}\}_{i=1}^m$ as the set of data matrices and X_i is the matrix representation of the i-th source. Note that for each source, we extract a set of features based on the previous analysis to augment the set of traditional features because these features cannot be captured by traditional ones but have abilities to discriminate self-harm content from normal content. For instance, we augment traditional word embedding features for the textual source by extracting features such as linguistic style and sentiments according to the textual analysis. More details about the traditional features can be found in the experiments section.

3.1 A Supervised Self-harm Content Prediction Framework

Under the supervised setting, we assume the availability of the label information of posts in $\mathcal{P}$. Let $\mathbf{Y} \in \mathbb{R}^{n \times 2}$ denote the label information of the n posts in $\mathcal{P}$ where $\{\mathbf{Y}_{i1} = 1, \mathbf{Y}_{i2} = 0\}$ and $\{\mathbf{Y}_{i1} = 0, \mathbf{Y}_{i2} = 1\}$ if the post p_i is labeled as self-harm and normal content, respectively. We concatenate $\{X_i \in \mathbb{R}^{n \times l_i}\}_{i=1}^m$ into one matrix $\mathbf{X} \in \mathbb{R}^{n \times \sum_{i=1}^m l_i}$. The goal is to learn a function $\mathbf{W} \in \mathbb{R}^{\sum_{i=1}^m l_i \times 2}$ that can map $\mathbf{X}$ to $\mathbf{Y}$. In this work, the basic method learns $\mathbf{W}$ via solving the following least square problem as:

$$\min_{\mathbf{W}} \|\mathbf{XW} - \mathbf{Y}\|_F^2 \tag{1}$$

However, after concatenating m feature spaces, the feature dimension of $\mathbf{X}$ is $\sum_{i=1}^m l_i$ and $\mathbf{X}$ could be very high-dimensional. Therefore the basic method in Eq. (1) could suffer from the curse of dimensionality. Meanwhile not all features especially these traditional features are useful to distinguish self-harm content and normal content. Therefore it is desired to incorporate feature selection into the framework that is achieved via adding $\ell_{2,1}$-norm regularization on $\mathbf{W}$. With the feature selection component, the supervised self-harm content prediction framework SCP is to solve the following optimization problem:

$$\min_{\mathbf{W}} \|\mathbf{XW} - \mathbf{Y}\|_F^2 + \alpha \|\mathbf{W}\|_{2,1} \tag{2}$$

where $\|\mathbf{W}\|_{2,1}$ ensures that $\mathbf{W}$ is sparse in rows, making it particularly suitable for feature selection. The parameter α controls the sparsity of $\mathbf{W}$.

Taking the derivative of the objective function in Eq. (2) and setting it to be zero, we can obtain the closed-form solution for $\mathbf{W}$ as:

$$\mathbf{W} = (\mathbf{X}^\top \mathbf{X} + \alpha \mathbf{D})^{-1} \mathbf{X}^\top \mathbf{Y} \tag{3}$$

where $\mathbf{D}$ is a diagonal matrix with its j-th diagonal element as $\mathbf{D}(j,j) = \frac{1}{2\|\mathbf{W}(j,:)\|_2}$.

3.2 An Unsupervised Self-harm Content Prediction Framework

Under the unsupervised scenario[42], we do not have label information to guide the learning process. However, in our studied problem, we have multiple sources that could make it possible to develop advanced framework for self-harm content prediction. The immediate challenge is how to capture relations among sources. Since we have the same set of posts for different sources, hence no matter which source we rely on to cluster posts, we should obtain similar cluster affiliations. This intuition paves us a way to capture relations among sources by assuming that all sources share the same cluster affiliations. We assume that $\mathbf{Z} \in \mathbb{R}^{n \times 2}$ is the shared cluster indicator matrix. Each post belongs to only one cluster where $\mathbf{Z}(i,1) = 1$ if p_i belongs to the first cluster, otherwise $\mathbf{Z}(i,1) = 0$. Thus $\mathbf{Z}$ should satisfy the following constraints:

$$\mathbf{Z}(i,:) \in \{0,1\}^k, \|\mathbf{Z}(i,:)\|_0 = 1, \quad 1 \le i \le n. \tag{4}$$

where $\|*\|_0$ is the vector zero norm, which counts the number of non-zero elements in the vector.

With the shared cluster indicator matrix $\mathbf{Z}$, we are further allowed to take advantages of information from multiple sources. First, we assume that similar data instances should have similar cluster indicators and then $\mathbf{Z}$ can be learned by spectral clustering:

$$\min_{\mathbf{Z}} \quad Tr(\mathbf{Z}^T \mathbf{L}_i \mathbf{Z}) \tag{5}$$

where $\mathbf{L}_i = \mathbf{V}_i - \mathbf{S}_i$ is a Laplacian matrix and $\mathbf{V}_i$ is a diagonal matrix with its elements defined as $\mathbf{V}_i(j,j) = \sum_{K=1}^{n} \mathbf{S}_i(K,j)$. $\mathbf{S}_i \in \mathbb{R}^{n \times n}$ denotes the similarity matrix based on $\mathbf{X}_i$ via a RBF kernel in this work.

Similar to the supervised framework SCP, we can learn a function $\mathbf{W}$ with the help of the shared cluster indicator matrix $\mathbf{Z}$. With these two model components, the proposed unsupervised self-harm content prediction framework USCP is to solve the following optimization problem:

$$\begin{aligned} \min_{\mathbf{W},\mathbf{Z}} \quad & \sum_{i=1}^{m} \lambda (Tr(\mathbf{Z}^T \mathbf{L}_i \mathbf{Z})) + \alpha \|\mathbf{X}\mathbf{W} - \mathbf{Z}\|_F^2 + \beta \|\mathbf{W}\|_{2,1} \\ \text{subject to} \quad & s_i \in \{0,1\}^n \\ & \|\mathbf{Z}(i,:)\|_0 = 1, i \in \{1,2,3...n\}, \\ & \mathbf{Z}(i,j) \in \{0,1\}, j \in \{1,2,...k\} \end{aligned} \tag{6}$$

The constraints in Eq. (6) is mixed vector zero norm with integer programming, making the problem hard to solve [9]. First, we need to relax the constraints on the cluster indicator matrix. By relaxing the value in $\mathbf{Z}$ to a continuous nonnegative value, we convert the constraints into:

$$\mathbf{Z}^T \mathbf{Z} = \mathbf{I}, \quad \mathbf{Z} \ge 0 \tag{7}$$

the constraints in Eq. (7) can ensure that there is only one non-negative value in each row of $\mathbf{Z}$.

With the relaxation, USCP is to solve the following optimization problem:

$$\begin{aligned} \min_{\mathbf{W},\mathbf{Z}} \quad & \sum_{i=1}^{m} \lambda_i (Tr(\mathbf{Z}^T \mathbf{L}_i \mathbf{Z})) + \alpha \|\mathbf{X}\mathbf{W} - \mathbf{Z}\|_F^2 + \beta \|\mathbf{W}\|_{2,1} \\ \text{subject to} \quad & \mathbf{Z}^T \mathbf{Z} = \mathbf{I}, \quad \mathbf{Z} \ge 0 \end{aligned} \tag{8}$$

We adopt an alternating optimization to solve the optimization problem of USCP and update $\mathbf{W}$ and $\mathbf{Z}$ iteratively and alternately. Since optimizing $\mathbf{W}$ is the same as that in Eq. (2), we focus on how to update $\mathbf{Z}$ in the following part. Fixing $\mathbf{W}$, $\mathbf{Z}$ can be obtained via the following optimization problem:

$$\begin{aligned} \min_{\mathbf{Z}} \quad & \sum_{i=1}^{m} \jmath(\mathbf{Z}) = \lambda_i (Tr(\mathbf{Z}^T \mathbf{L}_i \mathbf{Z})) + \alpha \|\mathbf{X}\mathbf{W} - \mathbf{Z}\|_F^2 \\ \text{subject to} \quad & \mathbf{Z}^T \mathbf{Z} = \mathbf{I}, \quad \mathbf{Z} \ge 0 \end{aligned} \tag{9}$$

The Lagrangian function of Eq. (9) is:

$$\begin{aligned} \mathbf{Z} = & Tr(\mathbf{Z}^T \mathbf{M} \mathbf{Z}) + Tr(\Gamma(\mathbf{Z}^{\mathbf{Z}} - \mathbf{I}) \\ & - Tr(\Lambda \mathbf{Z}) + \alpha Tr(-2\mathbf{A}^\top \mathbf{Z} + \mathbf{Z}^T \mathbf{Z}) \end{aligned} \tag{10}$$

where we use $\mathbf{M} = \sum_{i=1}^{m} \lambda_i \mathbf{L}_i$, and $\mathbf{A} = \mathbf{X}\mathbf{W}$. Γ and Λ are Lagrangian multipliers. Due to the space limit, we omit the derivations to optimize Eq. (10), and more details can be found in [9]. The provided updating rule for $\mathbf{Z}$ is as following:

$$\mathbf{Z}(p,q) \leftarrow \mathbf{Z}(p,q) \sqrt{\frac{(\mathbf{M}^- \mathbf{Z} + \alpha \mathbf{A}^+ + \mathbf{Z}\Gamma^-)(p,q)}{(\mathbf{M}^+ \mathbf{Z} + \alpha \mathbf{A}^- + \mathbf{Z} + \mathbf{Z}\Gamma^+)(p,q)}} \tag{11}$$

where $\mathbf{X}^+(p,q) = (|\mathbf{X}(p,q)| + \mathbf{X}(p,q))/2$, $\mathbf{X}^-(p,q) = (-|\mathbf{X}(p,q)| + \mathbf{X}(p,q))/2$, $\mathbf{X} = \mathbf{X}^+ + \mathbf{X}^-$, and $\Gamma = \alpha(\mathbf{Z}^T \mathbf{A} - \mathbf{I}) - \mathbf{Z}^T \mathbf{M} \mathbf{Z}$

With the updating rules of $\mathbf{Z}$ and $\mathbf{W}$, we present the detailed algorithm to optimize Eq. (8) in Algorithm 1.

Algorithm 1 Pseudo code of the proposed USCP

Input: $\{\mathbf{X}_i, \lambda_i\}, \alpha, \beta$
Output: the cluster label for each instance
for i = 1 to m: **do**
  Constructing Laplacian Matrix $\mathbf{L}_i$
end for
while Not converge **do**
  Update $\mathbf{W}$ by Eq. 3
  Compute $\mathbf{A} = \mathbf{X}\mathbf{W}$
  Compute $\Gamma = \alpha(\mathbf{Z}^T \mathbf{A} - \mathbf{I}) - \mathbf{Z}^T \mathbf{M} \mathbf{Z}$
  Update $\mathbf{Z}$ using Eq. (11)
end while
for i = 1 to n: **do**
  Max pooling in $\mathbf{Z}$ to find the cluster label for each instance
end for

4. EXPERIMENTS

In this section, we conduct experiments which (a) quantify the effectiveness of the proposed frameworks, and (b) validate the importance of findings from data analysis in discovering self-harm content. We begin by introducing experimental settings.

4.1 Experiment Settings

Datasets. We perform the evaluation on the dataset used in the data analysis section. That dataset is balanced with equal size of self-harm and normal content. In reality, there could be more normal content than self-harm content. To consider this situation, we sample 850,000 more normal content from these normal users to construct an imbalanced dataset. We will assess the performance of self-harm content prediction on both balanced and imbalanced datasets under supervised and unsupervised settings.

The Finding Features. Our findings in the previous section contain multiple cues. For each finding, we regard as one feature source. (1) lingual feature (a vector of language structure ratios and normalized term frequencies in the lexicon) (2) owner feature (a vector of user information) (3) temporal feature(1-hot vector of time) (4) visual feature (a vector of averaged saliency value, averaged HSV value, averaged pleasure, arousal and dominance value, and normalized SIFT,LBP GIST feature[7])

Traditional Features. In addition to features extracted according to our findings in the data analysis section, we also follow the state-of-the-art methods to extract traditional features for textual and visual sources as follows:

- The textual features are extracted from texts in social media posts including descriptions, titles and comments. For each word, we first transform it to a 100-dimension vector representation using a pre-trained word2vec [28] model. The final feature representation is the sum of vector representation for all the words.

- The CNN features are the last layer of fully connected layer of the convolutional neutral network. In our experiment, we use AlexNet[23] pre-trained on ImageNet. The feature size is 4096.

Note that we use both the finding features and traditional features. Thus, the m is set to be 6 in both SCP and USCP.

4.2 Performance Comparisons for Supervised Self-harm Content Prediction

Evaluation metrics: In imbalanced datasets, the accuracy metric under supervised settings is well known to be misleading [7]. For example, given the massive data on social media, a trivial classifier that labels all the samples as non self-harm post can achieve very high accuracy. In self-harm content prediction, we aim to achieve high precision and recall over self-harm posts defined in terms of the confusion matrix of a classifier– $prescision = \frac{tp}{tp+fp}$, $recall = \frac{tp}{tp+fn}$ and $F1 = 2\frac{precision \cdot recall}{precision+recall}$. Usually precision and recall are combined into their harmonic mean, the Fmeasure; hence we will adopt F1-measure as one metric for the performance evaluation. As suggested in [1, 7], in some scenarios, we put more emphasis on precision because the most challenging task is to seek for some self-harm posts with high probability, even at the price of increasing false negatives. Hence, we also report the precision performance.

We compare the proposed supervised framework SCP with the following baselines:

[7]We use PCA to reduce the feature dimension of concatenated of SIFT,LBP and GIST features. The final dimension of these three features is 128

Algorithm	Balanced		Imbalanced	
	F1	precision	F1	precision
Word-embedding	57.9%	63.7%	37.9%	30.1 %
CNN-image	61.8%	64.5%	48.6%	44.7%
CNN+WE	68.3%	72.3%	53.1%	46.7%
SCP-lite	68.4%	73.1%	54.5%	47.9%
SCP	**72.1**%	**75.2**%	**56.7**%	**49.8**%

Table 5: Performance comparisons for supervised self-harm content prediction.

- *Word-embedding(WE)*: We represent each text as the sum of the embedding of the words it contains; and the prediction is based on a 2 layer convolutional neural network [22]. This method is one of the state-of-the-arts in textual classification tasks such as sentiment classification [22];

- *CNN-image*: It is one of the state-of-the-art model for image classification [23]. We use the same architecture except the softmax layer with self-harm and normal labels;

- *CNN+WE* : We combine CNN and word embedding features and the prediction is based on a linear regression model; and

- *SCP-lite*: A lite version of SCP which only considers traditional features; while ignoring features extracted by our findings.

We use 60% of the data as training and the remaining as testing, and parameters are determined via cross-validation. The comparison results are demonstrated in Table 5. We make the following observations:

- CNN+WE obtains much better performance than Word-embedding and CNN-image. This result suggests that textual and visual sources contain complementary information;

- By incorporating feature selection, SCP-lite performs slightly better than CNN+WE; and

- SCP outperforms SCP-lite in both balanced and imbalanced datasets. We conduct t-test on the results and the evidence from t-test indicates the improvement is significant. The remarkable improvement of SCP over SCP-lite is from the augmented features. These results demonstrate that (1) traditional features cannot fully cover our findings; and (2) features extracted based on our findings can boost the performance significantly.

4.3 Performance Comparisons for Unsupervised Self-harm Content Prediction

In this subsection, we evaluate the proposed unsupervised framework USCP. Following the common practice[41], we choose NMI and accuracy (ACC) to assess the clustering performance. The baseline methods are defined as follows:

- *CNN+kmeans*: We use pre-trained CNN features [23] and then perform kmeans for clustering;

- *Word-embedding+kmeans*: We use word embedding features and then perform kmeans for clustering;

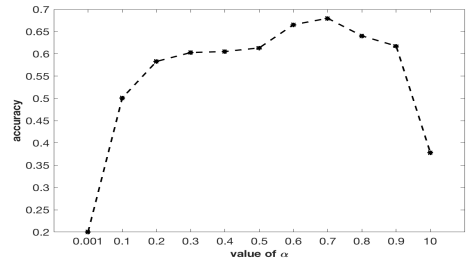
Algorithm	Balanced		Imbalanced	
	NMI	ACC	NMI	ACC
CNN+kmean	0.36	47.3%	0.15	15.3%
WE+kmeans	0.08	33.8%	0.04	10.3 %
CNN+WE+kmeans	0.46	56.2%	0.23	23.1%
USCP-lite	0.48	58.3%	0.26	24.3%
USCP	0.51	61.2%	0.31	27.4%

Table 6: Performance comparisons for unsupervised self-harm content prediction.

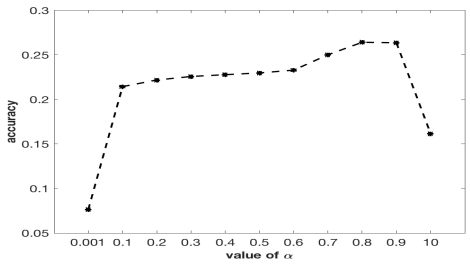
- *CNN+WE+kmeans*: We combine CNN and word embedding features and then perform kmeans for clustering; and
- *USCP-lite*: It is a variant of the proposed framework USCP that ignores features extracted according to our findings.

Since kmeans can only obtain local optimal solutions, we repeat experiments for these baselines based on kmeans 10 times and report the average performance. The performance comparisons are shown in Table 6. From the table, we make similar observations as the supervised self-harm content prediction experiments: (1) textual and visual sources are complementary to each other; and (2) the features extracted based on our findings can significantly improve the prediction performance under the unsupervised setting.

Parameter Analysis. There is one important parameter α for the proposed unsupervised framework USCP. The parameter controls the contribution of the model component capturing relations among sources. Next we study the impact of the component on the proposed framework by investigating how the performance changes with different values of α. We vary α as $\{0.001, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 10\}$. The performance variance w.r.t. α in terms of ACC is shown in Figure 5. Note that we do not show performance in terms of NMI since we make similar observations. In general, with the increase of α, the performance tends to first increase and then decrease. In particular, (1) the performance increases a lot when α is increased from 0.001 to 0.1 that indicates the importance of capturing relations among sources; (2) when α in certain regions, the performance is relatively stable that can ease the process of parameter selection in practice; and (3) when α increases to 10, the performance degrades significantly since the term capturing relations among sources will dominate the learning process that will lead to overfitting.



(a) Balanced Data



(b) Imbalanced Data

Figure 5: Parameter analysis for the proposed unsupervised framework USCP.

5. RELATED WORK

Selfharm research from psychology and medicine: Some work [48, 29] from psychology and medicine have been done on understanding and characterizing the deliberate self-harm patients. In [15], it investigates 8,950 deliberate self-harm (DSH) patients from 1990 to 2000 in Oxford, UK to capture their behavior trends. It shows that from 1997 to 2000, gender and age became a large portion of DSH – DSH rates in female and aged in 15 to 24 and 34 to 54 have been significantly increased. The major reasons of DSH are alcohol abuse, violence and misusing drugs. In [2], the authors reported that DSH helps the patients escape or regulate the emotions and most self-injurious behaviors are along with cognitive disabilities. In recent years [6, 10, 34, 30, 11, 18]more and more attention has been paid on social media platforms and studies [34] have shown that self-harm and suicide can be prevented from social supports from other social media users. However, the limitation of these studies is that they are typically based on surveys and self-reports about emotion and results a relative amount of the data. Most assessments are designed to collect the data about DSH experiences over long periods of time (1 to 5 years). Few studies are on the short term since the resources and invasiveness are required to observe individuals' behaviors over days and months.

Social media mining and health: The explosive growth of web and social media has records millions of user data, how to utilized such large scale data for human behavior sensing is a crucial challenge for the research communities. [3] and [24] have studied how to scale the prediction model on the time-evolving data . [44, 46, 43] explore how to use the multimodal data in the web for sentiment analysis. In the last few years, the interests of studying public health in social media are keep growing in the research community. [36] explored how to find diseases based on the posts in Twitter. [1] studied the eating-disorder community on Tumblr and finds that the tags for eating-disorder community are keep evolving. In [7, 8], authors investigated the patterns of activities for depression groups on web by analyzing the posts from Twitter and Reddit,. However, research on self-harm understanding in social media is still in its infancy.

6. PRIVACY AND ETHICS

We use all the public data and personally identifiable information was removed in the dataset. The content was de-identified and paraphrased before being reported in the paper for exemplary purpose

7. CONCLUSION

In this paper, we aim to understand and discover self-harm content in social media since social media has become increasingly popular for self-harm users to discuss their problems. We conducted the first comprehensive analysis on self-harm content with data from the social media site Flickr. Our analysis suggests that characteristics of self-harm content are different from those of normal content, from textual, owner, temporal and visual perspectives. These findings have potentials to help us distinguish self-harm content from others, and we have thus developed frameworks by incorporating these findings to discover self-harm content automatically. Empirical results demonstrate that (1) the proposed frameworks can accurately identify self-harm content under both supervised and unsupervised settings; and (2) our findings play an important role in boosting the prediction performance.

There are several interesting directions for further investigations. First, we would like to extend our proposed models to the semi-supervised setting because in reality we can obtain a small amount of labeled data but need to deal with a large amount of unlabeled data[47]. Second, while the findings on self-harm content motivated us to develop approaches for identifying posts related to self-harm, it is interesting to further understand how such post-level analysis can be extended to automatically identify the self-harm users. Third, social networks are pervasively available in social media and it could be promising to study the impact of peer influence on self-harm user behaviors and leverage social networks to improve predictive tasks in self-harm research.

8. ACKNOWLEDGMENT

Yilin Wang and Baoxin Li was supported in part by an ARO grant (#W911NF1410371) and an ONR grant (#N00014-15-1-2722). Any opinions expressed in this material are those of the authors and do not necessarily reflect the views of ARO or ONR.

9. REFERENCES

[1] S. Chancellor, Z. J. Lin, and M. D. Choudhury. "this post will just get taken down": Characterizing removed pro-eating disorder social media content. In *CHI, San Jose, CA, USA, May 7-12, 2016*.

[2] A. L. Chapman, K. L. Gratz, and M. Z. Brown. Solving the puzzle of deliberate self-harm: The experiential avoidance model. *Behaviour research and therapy*, 44(3):371–394, 2006.

[3] X. Chen and K. S. Candan. Lwi-svd: low-rank, windowed, incremental singular value decompositions on time-evolving data sets. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 987–996. ACM, 2014.

[4] M.-M. Cheng, N. J. Mitra, X. Huang, P. H. Torr, and S.-M. Hu. Global contrast based salient region detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(3):569–582, 2015.

[5] J. Cohen. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, 70(4):213, 1968.

[6] K. Daine, K. Hawton, V. Singaravelu, A. Stewart, S. Simkin, and P. Montgomery. The power of the web: a systematic review of studies of the influence of the internet on self-harm and suicide in young people. *PloS one*, 8(10):e77555, 2013.

[7] M. De Choudhury, M. Gamon, S. Counts, and E. Horvitz. Predicting depression via social media. In *ICWSM*, page 2, 2013.

[8] M. De Choudhury, E. Kiciman, M. Dredze, G. Coppersmith, and M. Kumar. Discovering shifts to suicidal ideation from mental health content in social media. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 2098–2110. ACM, 2016.

[9] C. H. Ding, T. Li, and M. I. Jordan. Convex and semi-nonnegative matrix factorizations. *IEEE transactions on pattern analysis and machine intelligence*, 32(1):45–55, 2010.

[10] M. P. Dyson, L. Hartling, J. Shulhan, A. Chisholm, A. Milne, P. Sundar, S. D. Scott, and A. S. Newton. A systematic review of social media use to discuss and view deliberate self-harm acts. *PLoS one*, 11(5):e0155813, 2016.

[11] C. Eichenberg, J. Dyba, and M. Schott. Bindungsstile, nutzungsmotive und internetsucht. *Psychiatrische Praxis*, 44(01):41–46, 2017.

[12] K. Gimpel, N. Schneider, B. O'Connor, D. Das, D. Mills, J. Eisenstein, M. Heilman, D. Yogatama, J. Flanigan, and N. A. Smith. Part-of-speech tagging for twitter: Annotation, features, and experiments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pages 42–47. Association for Computational Linguistics, 2011.

[13] L. A. Goodman. Snowball sampling. *The annals of mathematical statistics*, pages 148–170, 1961.

[14] K. L. Gratz, S. D. Conrad, and L. Roemer. Risk factors for deliberate self-harm among college students. *American journal of Orthopsychiatry*, 72(1):128, 2002.

[15] K. Hawton, J. Fagg, S. Simkin, E. Bale, and A. Bond. Trends in deliberate self-harm in oxford, 1985-1995. implications for clinical services and the prevention of suicide. *The British Journal of Psychiatry*, 171(6):556–560, 1997.

[16] K. Hawton and A. James. Suicide and deliberate self harm in young people. *Bmj*, 330(7496):891–894, 2005.

[17] K. Hawton, K. Rodham, E. Evans, and R. Weatherall. Deliberate self harm in adolescents: self report survey in schools in england. *Bmj*, 325(7374):1207–1211, 2002.

[18] K. Hawton, K. E. Saunders, and R. C. O'Connor. Self-harm and suicide in adolescents. *The Lancet*, 379(9834):2373–2382, 2012.

[19] D. J. Houghton and A. N. Joinson. Linguistic markers of secrets and sensitive self-disclosure in twitter. In *System Science (HICSS), 2012 45th Hawaii International Conference on*, pages 3480–3489. IEEE, 2012.

[20] L. Itti and C. Koch. Computational modelling of visual attention. *Nature reviews neuroscience*, 2(3):194–203, 2001.

[21] S. Kairam, J. Kaye, J. A. G. Gómez, and D. A. Shamma. Snap decisions? how users, content, and aesthetics interact to shape photo sharing behaviors. *CHI 2016*, 2016.

[22] Y. Kim. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*, 2014.

[23] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[24] X. Li, S. Huang, K. S. Candan, and M. L. Sapino. Focusing decomposition accuracy by personalizing tensor decomposition (PTD). In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, CIKM 2014, Shanghai, China, November 3-7, 2014*, pages 689–698, 2014.

[25] W. Lian, P. Rai, E. Salazar, and L. Carin. Integrating features and similarities: flexible models for heterogeneous multiview data. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, pages 2757–2763. AAAI Press, 2015.

[26] L. Lustberg and C. F. Reynolds. Depression and insomnia: questions of cause and effect. *Sleep medicine reviews*, 4(3):253–262, 2000.

[27] J. Machajdik and A. Hanbury. Affective image classification using features inspired by psychology and art theory. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 83–92. ACM, 2010.

[28] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[29] M. A. Milkie. Social comparisons, reflected appraisals, and mass media: The impact of pervasive beauty images on black and white girls' self-concepts. *Social Psychology Quarterly*, pages 190–210, 1999.

[30] M. A. Moreno, A. Ton, E. Selkie, and Y. Evans. Secret society 123: understanding the language of self-harm on instagram. *Journal of Adolescent Health*, 58(1):78–84, 2016.

[31] J. J. Muehlenkamp, L. Claes, L. Havertape, and P. L. Plener. International prevalence of adolescent non-suicidal self-injury and deliberate self-harm. *Child and Adolescent Psychiatry and Mental Health*, 6(1):1, 2012.

[32] J. P. Pestian, P. Matykiewicz, M. Linn-Gust, B. South, O. Uzuner, J. Wiebe, K. B. Cohen, J. Hurdle, and C. Brew. Sentiment analysis of suicide notes: A shared task. *Biomedical informatics insights*, 5(Suppl. 1):3, 2012.

[33] K. Petrie and R. Brook. Sense of coherence, self-esteem, depression and hopelessness as correlates of reattempting suicide. *British Journal of Clinical Psychology*, 31(3):293–300, 1992.

[34] J. Robinson, G. Cox, E. Bailey, S. Hetrick, M. Rodrigues, S. Fisher, and H. Herrman. Social media and suicide prevention: a systematic review. *Early Interv Psychiatry*, 2015.

[35] S. S. Rude, C. R. Valdez, S. Odom, and A. Ebrahimi. Negative cognitive biases predict subsequent depression. *Cognitive Therapy and Research*, 27(4):415–429, 2003.

[36] A. Sadilek, H. A. Kautz, and V. Silenzio. Modeling spread of disease from social interactions. 2012.

[37] S. Siersdorfer, E. Minack, F. Deng, and J. Hare. Analyzing and predicting sentiment of images on the social web. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 715–718. ACM, 2010.

[38] S. W. Stirman and J. W. Pennebaker. Word use in the poetry of suicidal and nonsuicidal poets. *Psychosomatic Medicine*, 63(4):517–522, 2001.

[39] B. Thomee, D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, and L.-J. Li. The new data and new challenges in multimedia research. *arXiv preprint arXiv:1503.01817*, 2015.

[40] T. Tuytelaars and K. Mikolajczyk. Local invariant feature detectors: a survey. *Foundations and trends® in computer graphics and vision*, 3(3):177–280, 2008.

[41] N. X. Vinh, J. Epps, and J. Bailey. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *JLMR*, pages 2837–2854, 2010.

[42] S. Wang, Y. Wang, J. Tang, C. Aggarwal, S. Ranganath, and H. Liu. Exploiting hierarchical structures for unsupervised feature selection. In *Proceedings of the 2017 SIAM International Conference on Data Mining*, 2017.

[43] S. Wang, Y. Wang, J. Tang, K. Shu, S. Ranganath, and H. Liu. What your images reveal: Exploiting visual contents for point-of-interest recommendation. In *Proceedings of the 26th International Conference on World Wide Web*, 2017.

[44] Y. Wang, Y. Hu, S. Kambhampati, and B. Li. Inferring sentiment from web images with joint inference on visual and social cues: A regulated matrix factorization approach. In *Ninth International AAAI Conference on Web and Social Media*, 2015.

[45] Y. Wang and B. Li. Sentiment analysis for social media images. In *Proceedings of the 2015 IEEE International Conference on Data Mining Workshop (ICDMW)*, pages 1584–1591. IEEE Computer Society, 2015.

[46] Y. Wang, S. Wang, J. Tang, H. Liu, and B. Li. Unsupervised sentiment analysis for social media images. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2015.

[47] Y. Wang, S. Wang, J. Tang, H. Liu, and B. Li. Ppp: Joint pointwise and pairwise image label prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6005–6013, 2016.

[48] J. M. G. Williams and M. Williams. *Cry of pain: Understanding suicide and self-harm*. Penguin Group USA, 1997.

[49] E. Yom-Tov, L. Fernandez-Luque, I. Weber, and S. P. Crain. Pro-anorexia and pro-recovery photo sharing: a tale of two warring tribes. *Journal of medical Internet research*, 14(6):e151, 2012.