YAHOO!

Discovering and Understanding Self-harm Images in Social Media

Neil O'Hare, MFSec 2017. Bucharest, Romania,
June 6th, 2017

Who am I?

PhD from Dublin City University, 2007. Multimedia Search

Currently Senior Research Scientist with Yahoo (Since 2011)

Research Interests: on ***Image Search,*** *Video Search*, Multimedia data mining, social media data mining, computational aesthetics

Past work related to Multimedia Forensics

> *"Combining social network analysis and sentiment analysis to explore the potential for online radicalisation", ASONAM 2009.*

Collaborators

This work was a colloboration with Yilin Wang, Jiliang Tang, Yi Chang

"Understanding and Discovering Deliberate Self-harm Content in Social Media", Yilin Wang, Jiliang Tang, Jundong Li, Baoxin Li, Yali Wan, Clayton Mellina, **Neil O'Hare,** Yi Chang. ***WWW 2017, Perth, Australia***

Roadmap

Motivation

Collecting self-harm data from Social Media

Analysis of Self Harm Behavior

Detecting Self Harm Behavior

Computational Forensics for Personal Health

YAHOO!

Computational Forensics for Personal Health

Mining large-scale data to identify people or groups exhibiting potentially behavior that could lead to dangerous or life-threatening outcomes

Computational Forensics for Personal Health

Mining large-scale data to identify people or groups exhibiting potentially behavior that could lead to dangerous or life-threatening outcomes

But they are a danger mainly to themselves, not to society at large

Computational Forensics for Personal Health

Mining large-scale data to identify people or groups exhibiting potentially behavior that could lead to dangerous or life-threatening outcomes

But they are a danger mainly to themselves, not to society at large

The ultimate goal is to lead to an intervention that helps them, rather than intervening for public security

Understanding Self-Harm Behavior

The field of *computational health* aims to apply computational methods to the field of health and medicine

- can potential lead to better treatment

Recent work, for example, in understanding social media content related to eating disorders [1]

In this work, we want to use social data to increase our understanding of self harm behavior

[1] “this post will just get taken down”: Characterizing removed pro-eating disorder social media content

What do we mean by self-harm?

YAHOO!

What do we mean by self-harm?

Self–harm, also known as self-injury and self-mutilation, is the intentional, direct injuring of body tissue, but not intended to be lethal.

What do we mean by self-harm?

#self injure

#blithe

#olive

Self harm

#secret society 123

#svv

Self mutilation

#ana mia

How common is self-harm?

- 2 Million cases reported annually (US)
- 2nd leading cause of teenage deaths (world wide)

YAHOO!

How common is self-harm?

- 2 Million cases reported annually (US)
- 2nd leading cause of teenage deaths (world wide)

Existing efforts and understanding self-harm mostly rely on self and friends/families reports, but most self harm symptoms are very difficult to discover

The relatively rare occurrence of completed self-harm treatment and the stigma associated with self-harm reports make the studies expensive to conduct

Why analyze self-harm on social media?

Social stigma exists for people who engaged in self harm

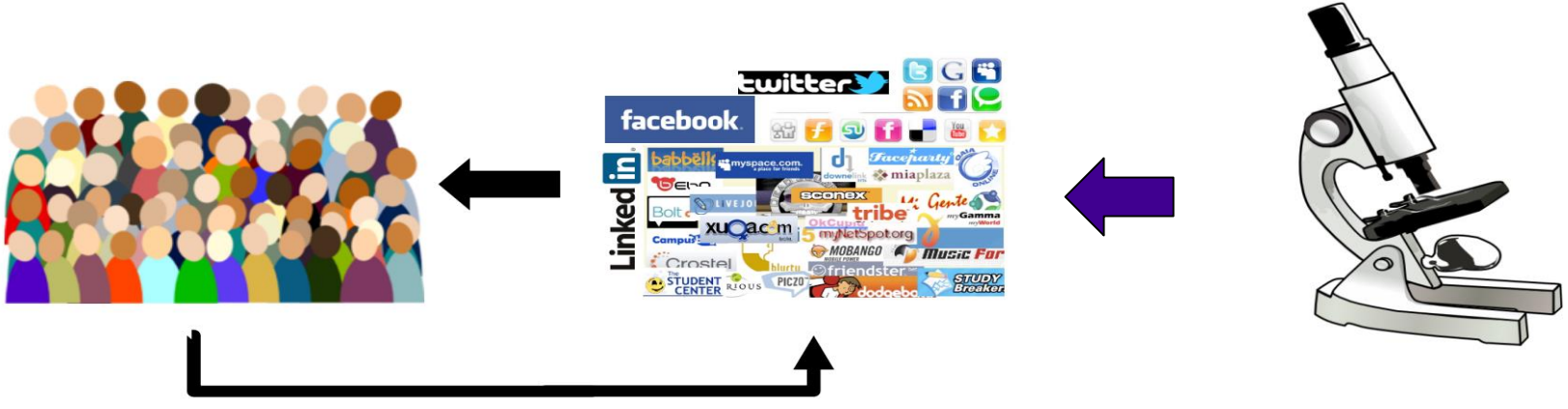
“*I swear to god, I got worse panic attack ever when adults talk about cutting and force you to show the wrist* ”

YAHOO!

Why analyze self-harm on social media?

Social stigma exists for people who engaged in self harm

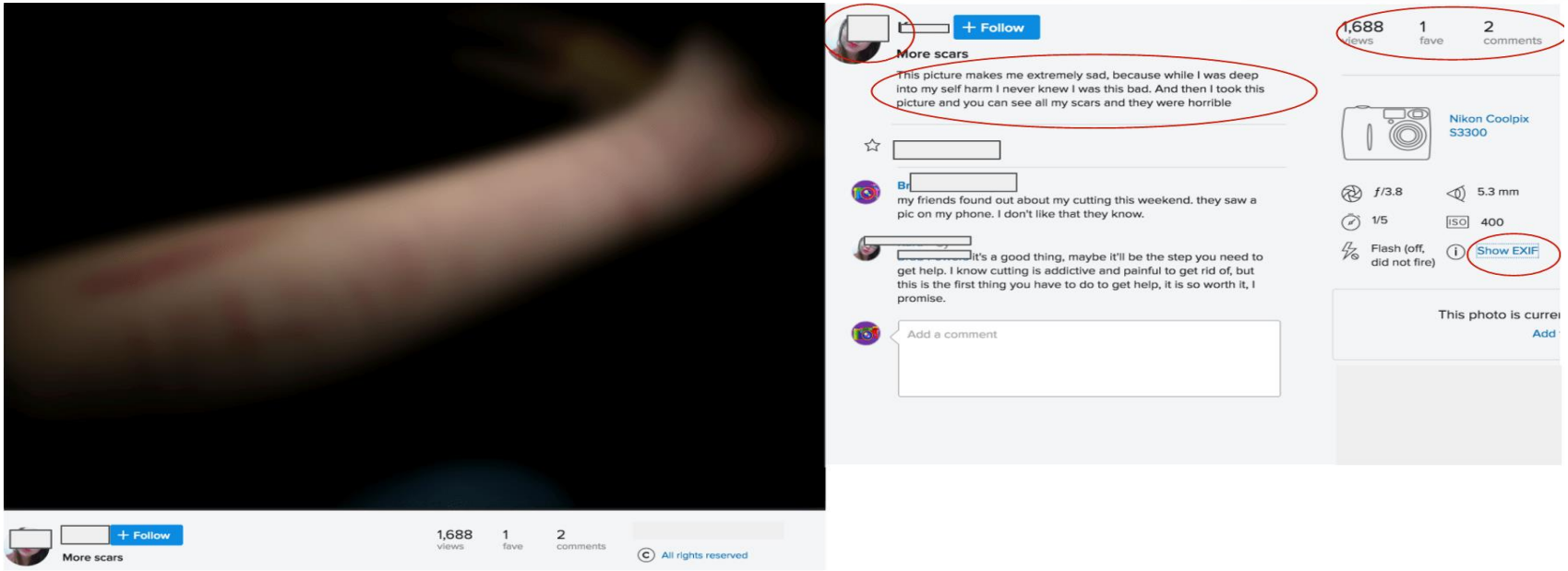
"I swear to god, I got worse panic attack ever when adults talk about cutting and force you to show the wrist "



- Social media as a monitor/sensor for human behavior
- A place where young adolescents often feel comfortable and escape stigma around their behavior, especially given anonymous setting

YAHOO!

Self-harm Content on Social Media: Example



YAHOO!

Research Questions

Does self-harm content have distinct characteristics?

Research Questions

Does self-harm content have distinct characteristics?

Can we use these characteristics to gain new insights into self-harm behavior?

Research Questions

Does self-harm content have distinct characteristics?

Can we use these characteristics to gain new insights into self-harm behavior?

Can we leverage these characteristics to build models to automatically discover self harm content?

Data Collection - 1

Flickr: 1 Billion posts, 50 Million users

Filter by Self-harm tags: “#selfharm” “#selfinjury”

1B-> 15,792 posts, 3,328 users

Manual analysis of 2k of these posts

identify 15 additional tags associated with self-harm content

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

Data Collection - 2

Extended tagset

383,614 posts from 63,949 users

Remove users with less than 5 self-harm posts

93,286 posts from 20,495 users

Manual annotation: 95% precision for ‘self-harm’ and ‘self-injury’. 83% precision for other tags.

Average/normal users from YFCC corpus [1]

All data **anonymised** before processing

[1] “The new data and new challenges for Multimedia Research”, Thomee et al

YAHOO!

Data Analysis

Focus on ‘interpretable’ features:

Data Analysis

Focus on ‘interpretable’ features:

Textual Analysis – text of title, descriptions & comments associated with posts

User Analysis – attributes of the user who owns the post

Temporal Analysis – temporal distribution of self-harm posts

Visual Content Analysis – visual content of self-harm images

Textual Analysis - Features

Distributions of nouns, verbs, adverbs, adjectives in posts, reflecting **language structure** (from CMUTweetTagger [1])

Readability score to measure complexity and readability of texts [2]

Sentiment polarity of text, based on off-the-shelf manual lexicon, i.e. MPQA

We build a ***lexicon*** of terms associated with self-harm posts

We **visualise** frequent tags

[1] Part of Speech Tagging for Twitter: Annotation, Features and Experiments, Gimpel et al, HLT 2011
[2] https://pypi.python.org/pypi/textstat/0.2

Textual Analysis – Language Structure, Readability, Sentiment

	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

Textual Analysis - Results

Self-harm content tends to include more verbs and adjectives/adverbs than nouns which is very consistent with suicidal word usage

Low noun usage – lack of interest in objects and things

The poor linguistic structure usage and language suggest the decreased cognitive functioning and coherence

A large portion of negative sentiment words are used in self-harm content.

Textual Analysis - Lexicon

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

Tags


selfmutilation
expression
bodyconfidence
blithe
kander
razor
skin
nadav
band
illness
black
selfhate
grunge
ana
suicide
donoharm
beautiful
dream
key
eye
selfharmmm
alone
quotes
depression
harm
anorexia
arm
disorder
scar
trigger
mentalhealth
sad
bulimia
addiction
ifttt
tear
fat
dark
beauty
face
death
knife
selfinjury
self
recovery
wall
cut
wrist
psychiatry
eatingdisorder
interesting
flickstagram
icoulduseahug
hate
broken
red
bodyimage
cutting
blood
art
injury
secret_society123
mentalillness
depressed
metaphorical
psycho
suicidal
cutter
selfportrait
selfharm
tattoo
leg
dan
hurt
pain
sigismondi
bipolar
state
secretsociety123
help
mental
emotion
mutilation
fear
lonely
anxiety
blackandwhite
celebrationsadnes
3a

YAHOO!

User Analysis

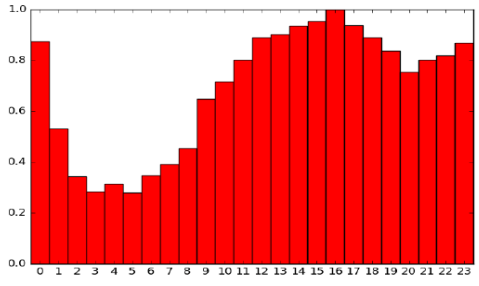
	Volume	% of reply	# favorite	# friends
Self-harm	7.76	0.51	0.56	296.89
Normal	3.79	0.11	0.23	477.57

More Active: average post from create account to last login in.

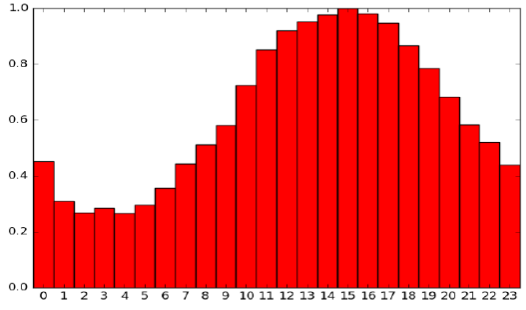
High proportion of reply and number of favorites indicate that self harm content receives more social response.

Self harm users have less friends

Temporal Analysis



(a) Self-harm related Content



(b) Normal Content.

Regular users

- fewer posts published later in the night and early morning.
- the number generally increases through the day (peaks in 3pm)

Such results suggests issues related to insomnia.

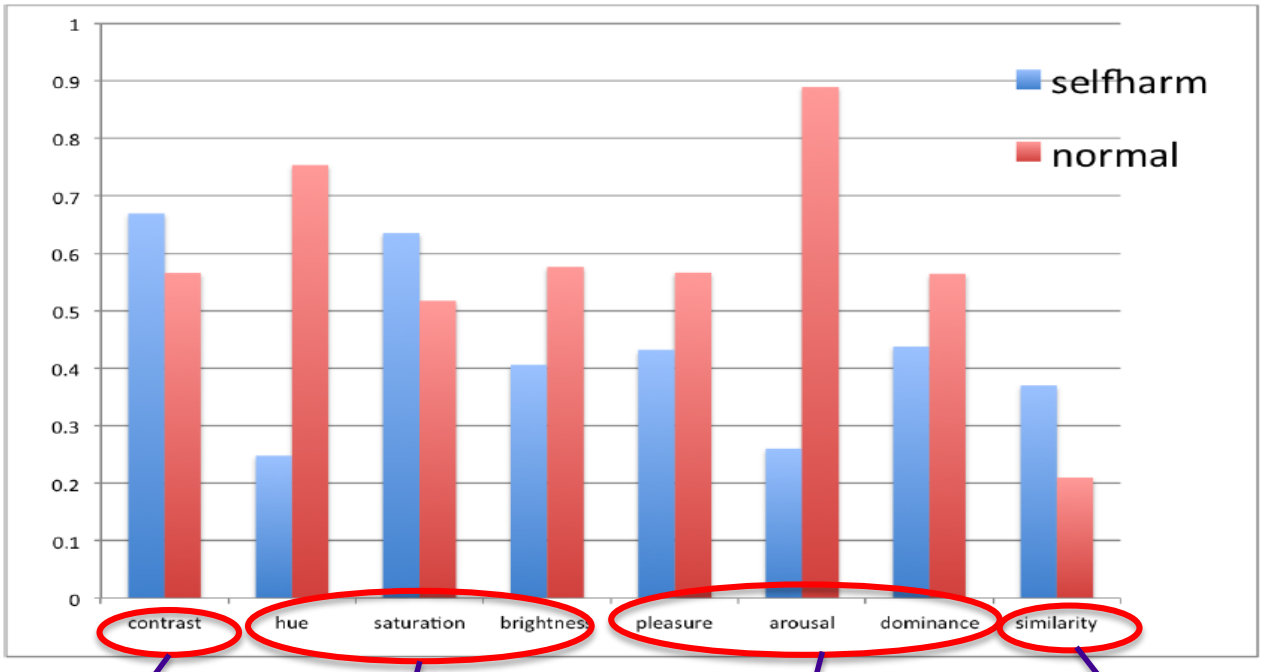
YAHOO!

Visual Analysis - Features

- **Global contrast**, provides saliency information about distinguishability of colors [1]
- Average of the ***hue***, ***brightness*** and ***saturation*** channels
- Measures of ***pleasure***, ***arousal*** and ***dominance*** [2]
- **SIFT**, **LBP** and **GIST** to calculate average similarity with other images

- [1] “Global contrast based salient region detection”, Cheng et al
- [2] “Affective image classification using features inspired by psychology and art theory”, Machajdik & Hanbury

Visual Content Analysis - Results



Importance of Findings

Give insights to help understanding people who engage in self-harm, relating to:

- language usage, sentiment and lexicon,
- temporal, user information and visual patterns

Can also help us to automatically identify users with self-harm issues, so they can be heard (and, potentially, helped)

We can derive features, based on these findings, that can be discriminative for detecting self-harm behavior.

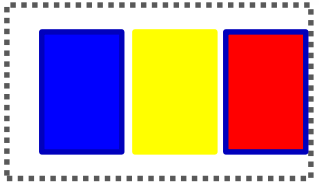
How can we detect self-harm behaviour?

Supervised Self-harm Content Prediction (SCP)

visual information

text information

Our findings



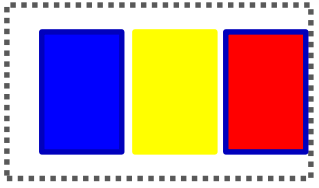
How can we detect self-harm behaviour?

Supervised Self-harm Content Prediction (SCP)

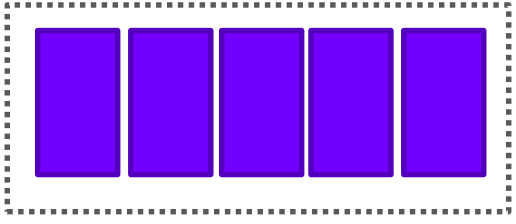
visual information

text information

Our findings



Labels



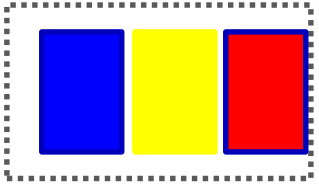
How can we detect self-harm behaviour?

Supervised Self-harm Content Prediction (SCP)

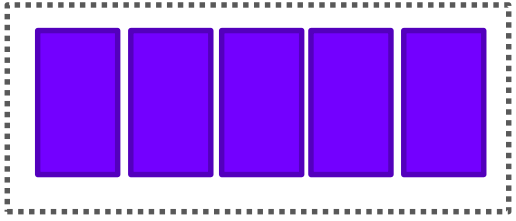
visual information

text information

Our findings



Labels



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

YAHOO!

Features for Self Harm Detection

Traditional Features

Visual Features: AlexNet CNN features: pre-trained on ImageNet, fine-tuned on self-harm corpus.

Text Features: Pre-trained 100-dim Word2Vec embedding, as input into 2 layer CNN with temporal max pooling [1]. Taken from titles, descriptions, comments

Proposed Features

textual: noun/verb/adjective/adverb ratios, readability, sentiment, normalised term frequency from lexicon

user: volume, #replies, #favorites, #friends

temporal: 24-dim one-hot vector

visual: saliency, HSV, pleasure, arousal, dominance, SIFT, LBP, GIST

[1] "Convolutional Neural networks for sentence classification", Y. Kim

Experiments

Dataset

Balanced dataset: equal size of self harm content and normal content (93k x 2)

Imbalanced dataset : 1:10 ratio of self harm to normal content.

60% Training / 40% test. Parameters selected via cross-validation

Metrics

F1 and precision

Model Variants

Word Embeddings (WE): 2 layer CNN with temporal max pooling [1]

CNN Image: AlexNet CNN architecture [2]

SCP-lite: Our SCP model, only considering 'traditional' features (CNN + WE)

SCP: Full model, including proposed features

[1] "Convolutional Neural networks for sentence classification", Y. Kim

[2] "ImageNet Classification with Deep Convolutional Neural Networks", Krizhevsky et al.

Supervised Classification Results (balanced)

Method	F1	Precision
Word Embedding	57.9%	63.7%

Supervised Classification Results (balanced)

Method	F1	Precision
Word Embedding	57.9%	63.7%
CNN-image	61.8%	64.5%

Supervised Classification Results (balanced)

Method	F1	Precision
Word Embedding	57.9%	63.7%
CNN-image	61.8%	64.5%
SCP-lite	68.4%	73.1%

Supervised Classification Results (balanced)

Method	F1	Precision
Word Embedding	57.9%	63.7%
CNN-image	61.8%	64.5%
SCP-lite	68.4%	73.1%
SCP	**72.1%**	**75.2%**

Supervised Classification Results (unbalanced)

Supervised Classification Results (unbalanced)

Method	F1	Precision
Word Embedding	37.9%	30.1%
CNN-image	48.6%	44.7%
SCP-lite	54.5%	47.9%
SCP	**56.7%**	**49.8%**

Conclusions

Our analysis suggests that the characteristics of self harm content is very different fromnormal content

Features inspired by our findings improve detection of self harm content

We can extend our work to a semi supervised learning problem for real-world data

We will explore the network influences to self harm users

Additional analysis of gender and other demographic variations would be interesting

Future Directions

Promising first steps, but not accurate enough for real world applications

Improved accuracy can come from: better training data from multiple platforms, improved features / models, word2vec trained on social media data, etc

Also legal / privacy concerns to be addressed

Future Directions

Promising first steps, but not accurate enough for real world applications

Improved accuracy can come from: better training data from multiple platforms, improved features / models, word2vec trained on social media data, etc

Also legal / privacy concerns to be addressed

This work was originally published in the new *computational health* track in *WWW 2017*

this field should continue to grow over the next number of years

Come see our spotlight presentation tomorrow,
and poster on Thursday!!

“Bridging the Aesthetic Gap: The Wild Beauty of Web Imagery”,

Miriam Redi, Frank Z. Liu and Neil O'Hare

YAHOO!

Questions?

nohare@yahoo-inc.com

YAHOO!