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 University [2]Michigan State University
[3]University of Washington [4] Yahoo Research
[5] Huawei Research

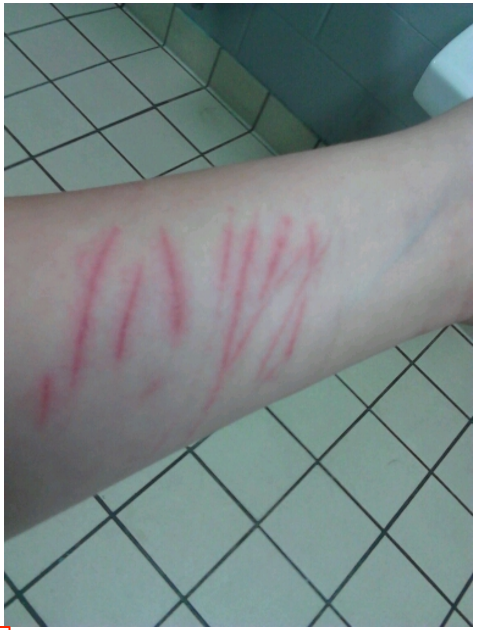
WWW 2017

Self harm: What is it about ?

#self injure

#blithe

#olive



#svv

#secret society 123

Self harm

Self mutilation

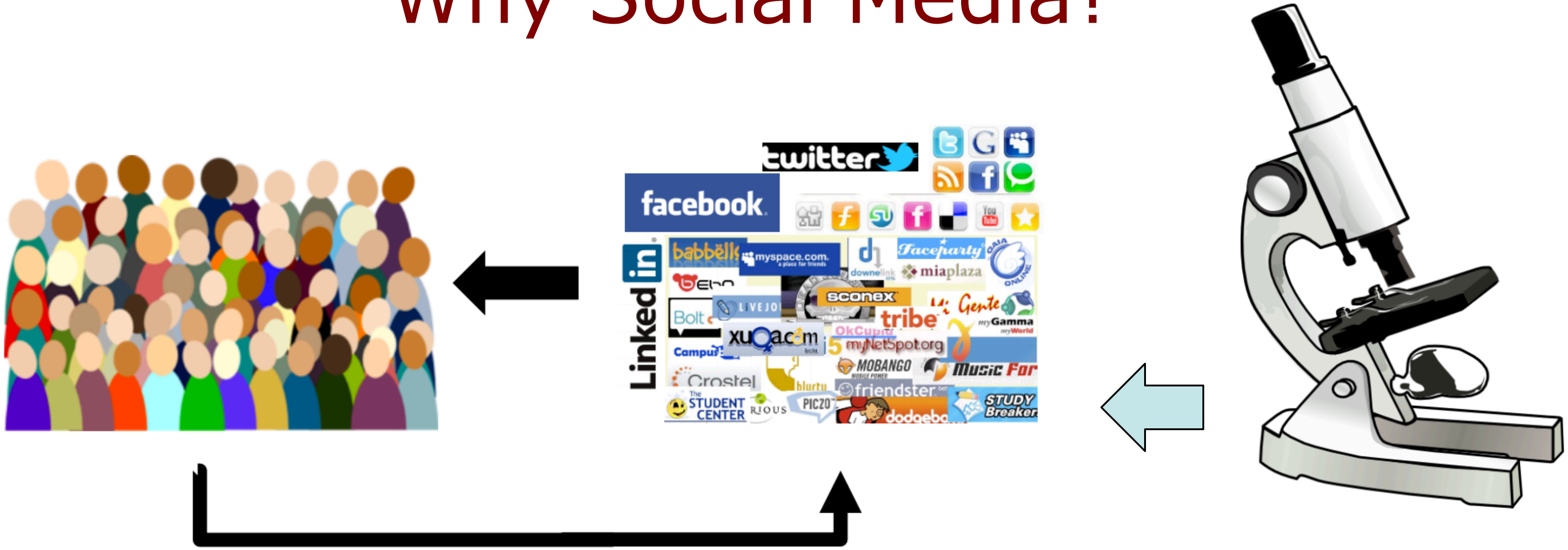
#ana mia

Self harm: How common?

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

- Existing efforts only relied on self and friends/families reports, but most of self harm symptom is very difficult to discover.
- The relatively rare occurrence of completed self-harm treatment and the rare population made the studies expensive to conduct.

Why Social Media?

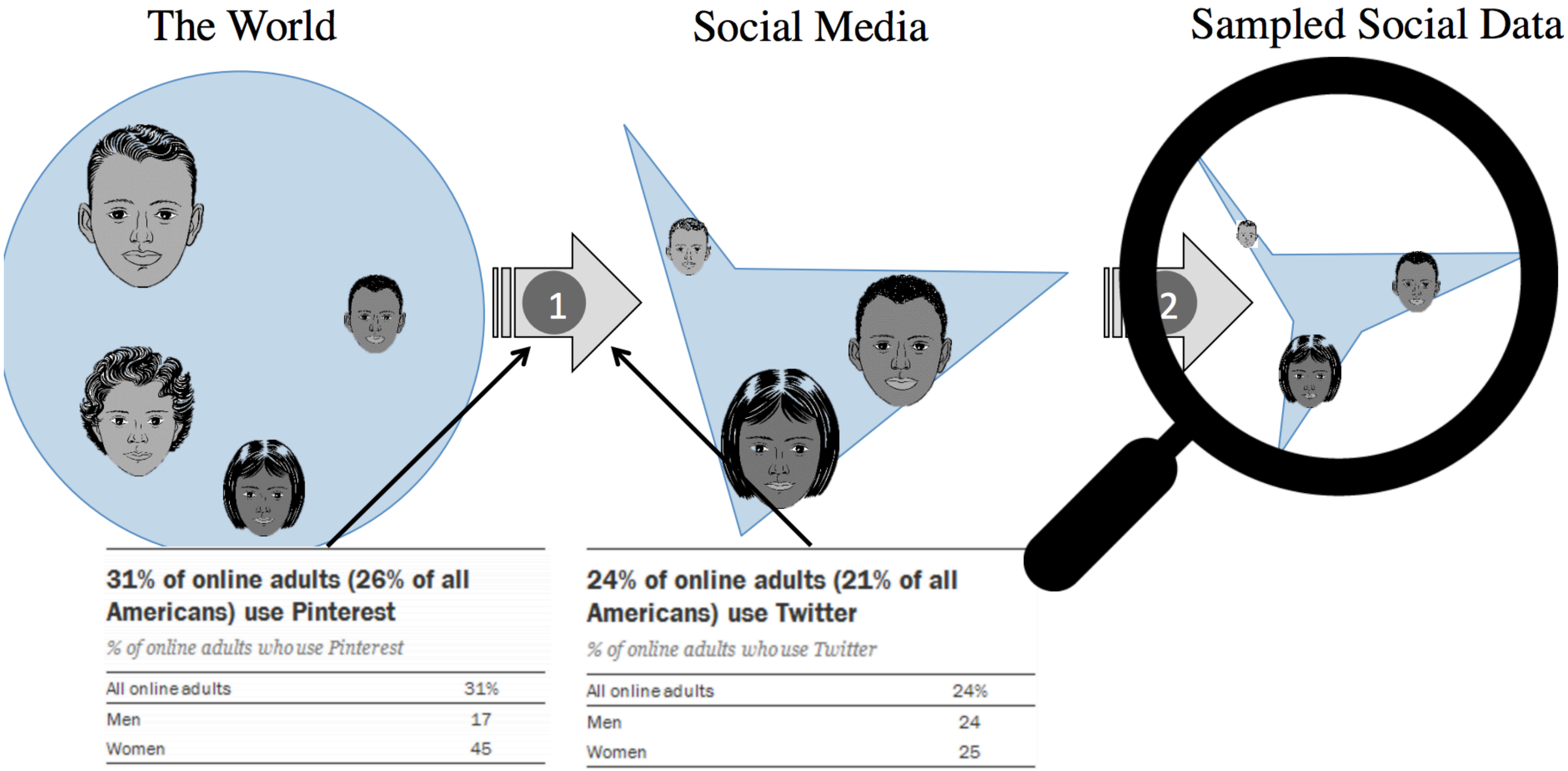


- Monitoring human behavior
- A better place for young adolescents

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"

Motivation



31% of online adults (26% of all Americans) use Pinterest

% of online adults who use Pinterest

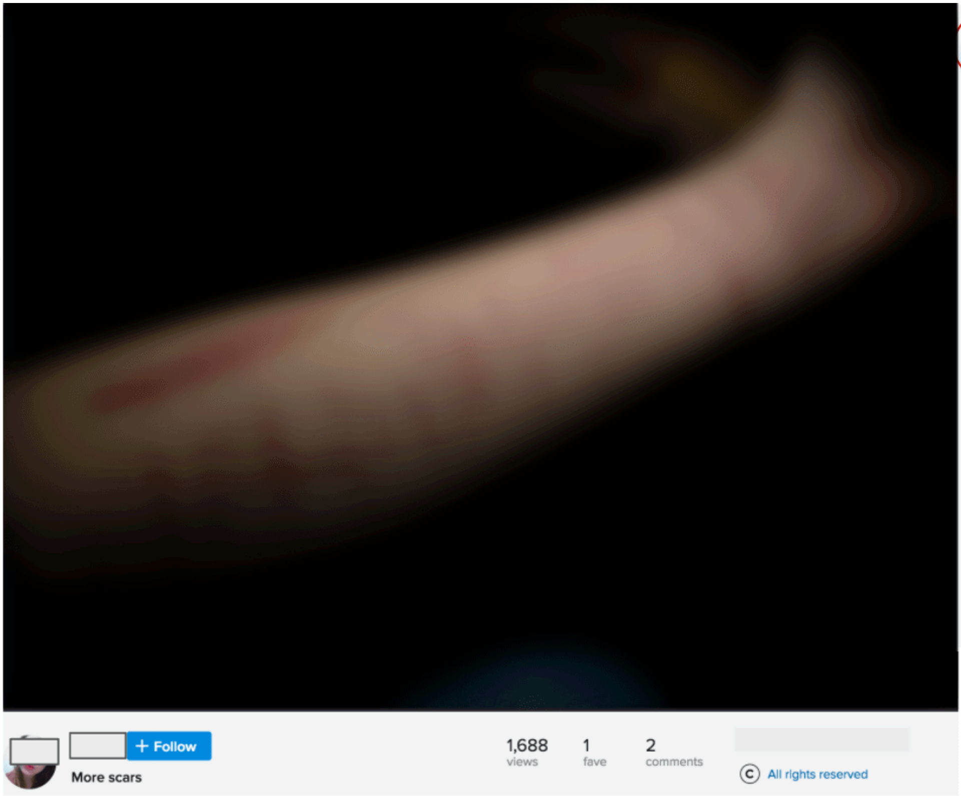
All online adults	31%
Men	17
Women	45

24% of online adults (21% of all Americans) use Twitter

% of online adults who use Twitter

All online adults	24%
Men	24
Women	25

Research Questions

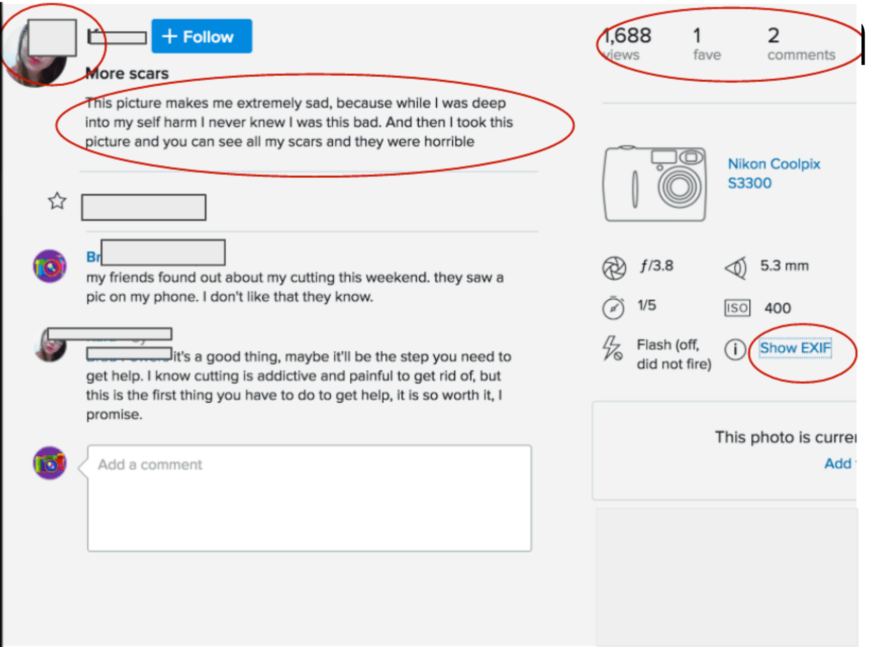


[redacted] + Follow

More scars

1,688 views 1 fave 2 comments

Ⓒ All rights reserved



[redacted] + Follow

More scars

This picture makes me extremely sad, because while I was deep into my self harm I never knew I was this bad. And then I took this picture and you can see all my scars and they were horrible

1,688 views 1 fave 2 comments

☆ [redacted]

Br[redacted]
my friends found out about my cutting this weekend. they saw a pic on my phone. I don't like that they know.

[redacted]
[redacted] it's a good thing, maybe it'll be the step you need to get help. I know cutting is addictive and painful to get rid of, but this is the first thing you have to do to get help, it is so worth it, I promise.

Add a comment

Nikon Coolpix S3300

ƒ/3.8 5.3 mm

1/5 ISO 400

Flash (off, did not fire)

Show EXIF

This photo is curre

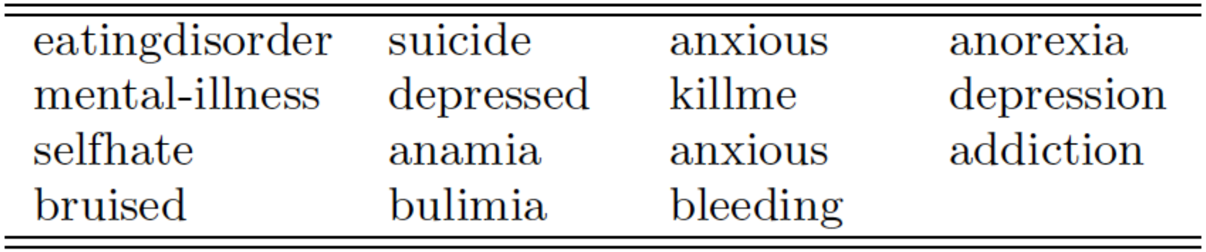
Add

Data Collection

Flickr: 10 Billion posts with 50 Million users

- Self-harm Content: “#selfharm” “#selfinjury”

  1B-> 15,792 posts



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

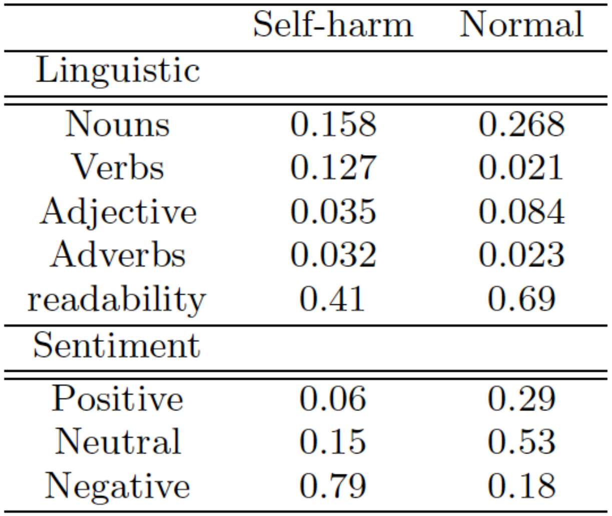
  Refine 383,614 posts and 63,949 users

- Normal User drawn from YFCC 19720 users and 93286 posts for each group

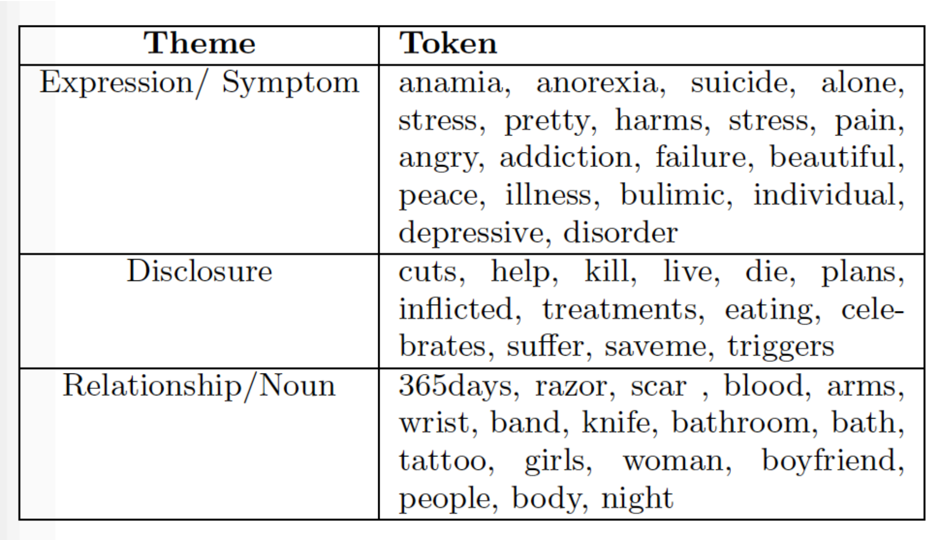
Data Analysis

- Textual Analysis
- User Analysis
- Temporal Analysis
- Visual Content Analysis

Textual Analysis



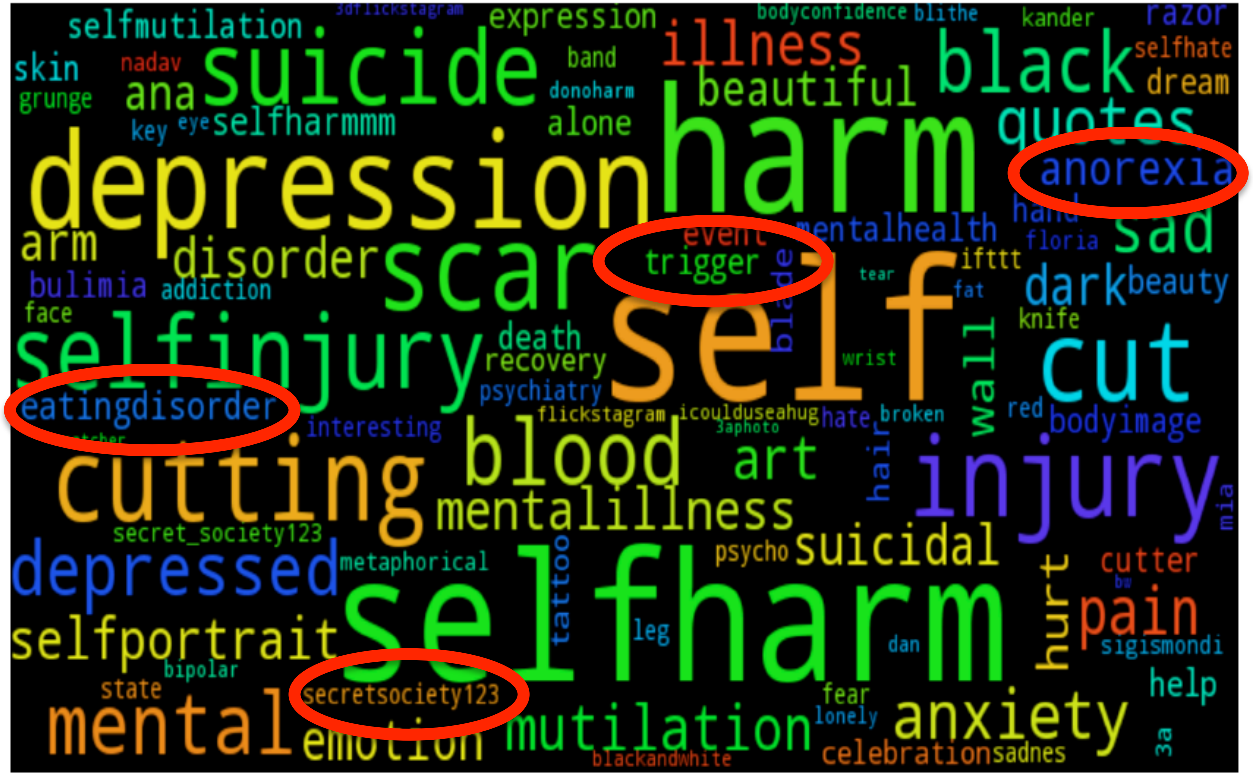
	Self-harm	Normal
Linguistic		
Nouns	0.158	0.268
Verbs	0.127	0.021
Adjective	0.035	0.084
Adverbs	0.032	0.023
readability	0.41	0.69
Sentiment		
Positive	0.06	0.29
Neutral	0.15	0.53
Negative	0.79	0.18



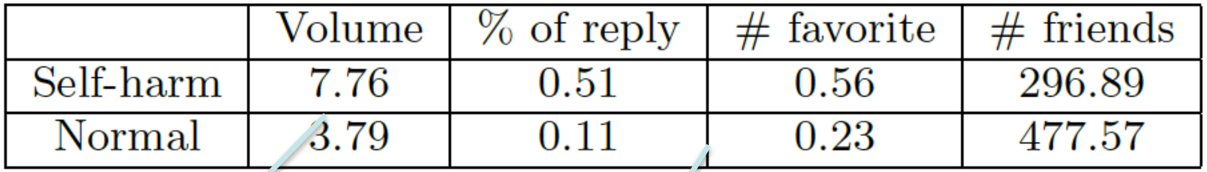
Theme	**Token**
Expression/ Symptom	anamia, anorexia, suicide, alone, stress, pretty, harms, stress, pain, angry, addiction, failure, beautiful, peace, illness, bulimic, individual, depressive, disorder
Disclosure	cuts, help, kill, live, die, plans, inflicted, treatments, eating, celebrates, suffer,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

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

Beyond Text


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

User Analysis



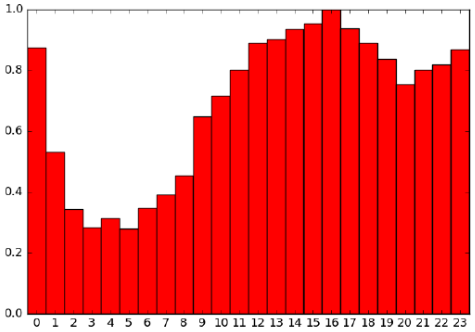
	Volume	% of reply	# favorite	# friends
Self-harm	7.76	0.51	0.56	296.89
Normal	3.79	0.11	0.23	477.57

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

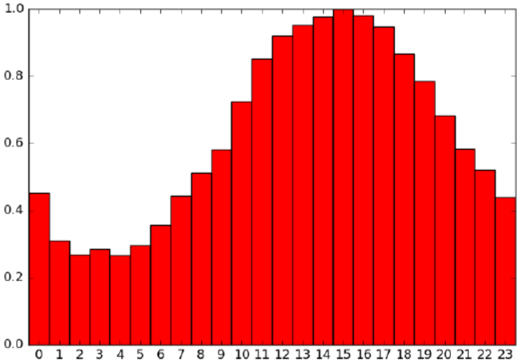
High proportion of reply and number of favorites indicate that self harm content receives more social response.

Self harm uses has less friends shows that it could be berried by the large portion of normal users.

Temporal Analysis



(a) Self-harm related Content



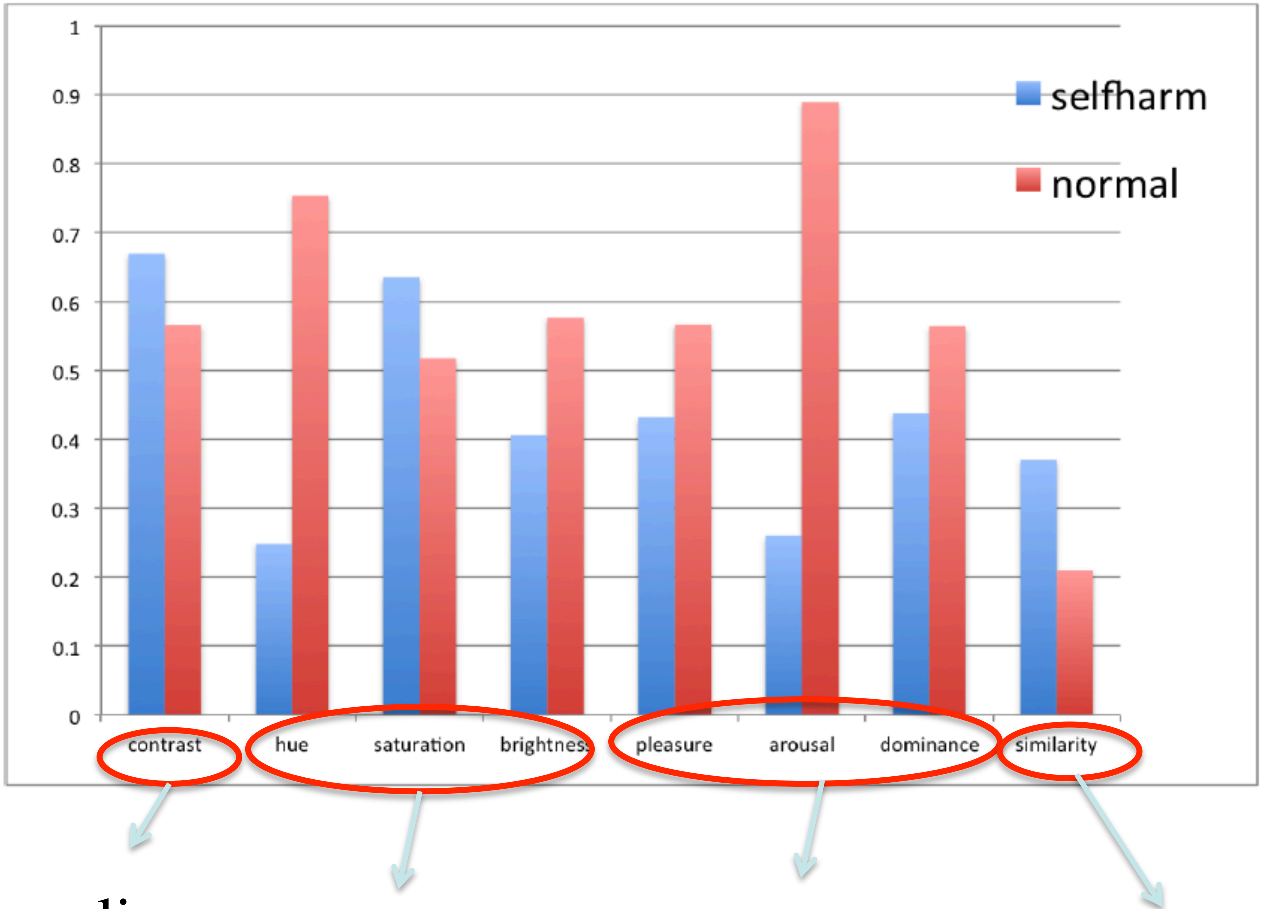
(b) Normal Content.

Normal users:

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

Such reason could be the mental issues related insomnia.

Visual Content Analysis



salienc
y

Image
Style

Image
emotion

patterns

The importance of our finding

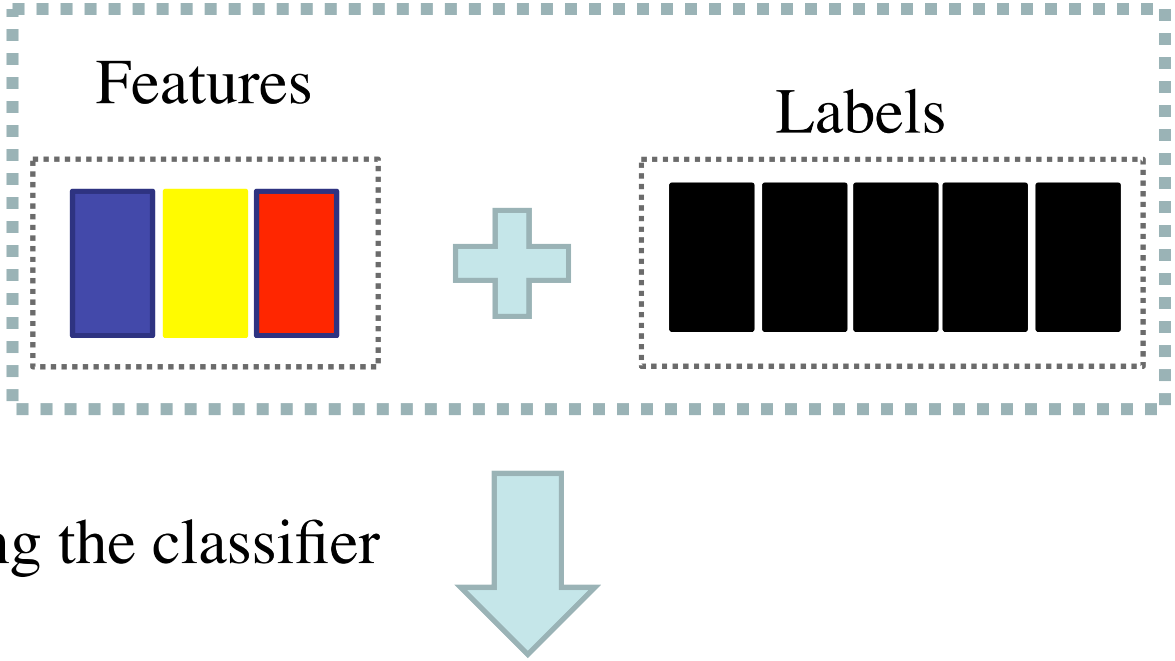
- Let self harm post to be heard.

Common feature: visual feature and textural feature (CNN+WE)

Our findings: language usage, sentiment and lexicon, temporal, user information and visual patterns

How to utilize the findings?

- supervised



Training the classifier

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

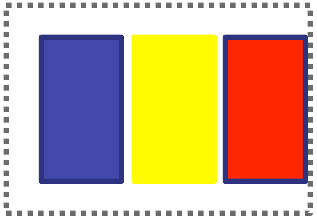
How to utilize the findings?

- unsupervised

visual information

textual information

Our findings



Model Learning

$$\min_{\mathbf{W},\mathbf{Z}} \quad \sum_{i=1}^{m} \lambda_i (Tr(\mathbf{Z}^T \mathbf{L}_i \mathbf{Z})) + \alpha \left\| \mathbf{X}\mathbf{W} - \mathbf{Z} \right\|_F^2 + \beta \left\| \mathbf{W} \right\|_{2,1}$$

$$\text{subject to} \quad \mathbf{Z}^T\mathbf{Z} = \mathbf{I}, \quad \mathbf{Z} \geq 0$$

Experiments

Dataset:

- Balanced dataset: equal size of self harm content and normal content. (150k)
- Imbalanced dataset : 1:10 with self harm to normal content. (150k and 850k)

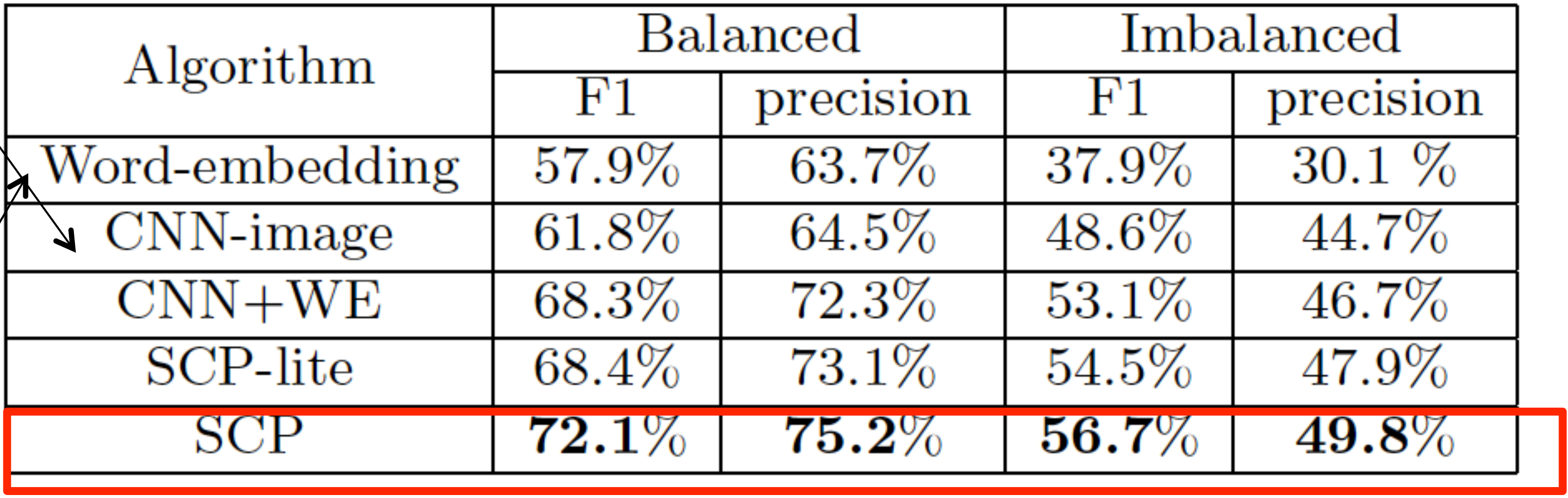
Metric:

- Supervised: F1 and precision
- Unsupervised: accuracy and NMI

Parameter analysis : alpha from 0.0001 to 10

Results of Supervised Method

Visual



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%**

Textual

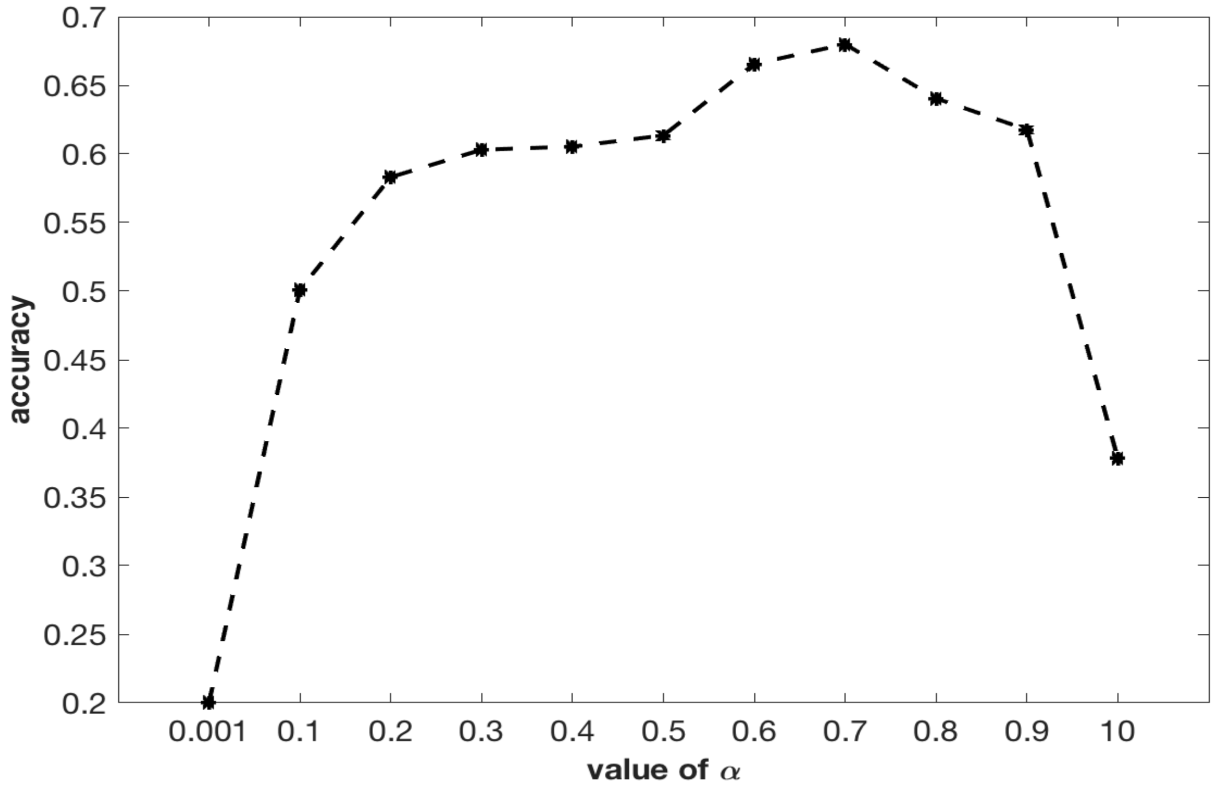
Results of Unsupervised Method

Visual

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%

Textual

Parameter



$$\min_{\mathbf{W},\mathbf{Z}} \quad \sum_{i=1}^{m} \lambda_i (Tr(\mathbf{Z}^T \mathbf{L}_i \mathbf{Z})) + \alpha \left\|\mathbf{XW} - \mathbf{Z}\right\|_F^2 + \beta \left\|\mathbf{W}\right\|_{2,1}$$

$$\text{subject to} \quad \mathbf{Z}^T\mathbf{Z} = \mathbf{I}, \quad \mathbf{Z} \geq 0$$

Conclusion

- Our analysis suggest that the characteristic of self harm content is very different with normal content.
- Features inspired by our findings improve detection of identify 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.