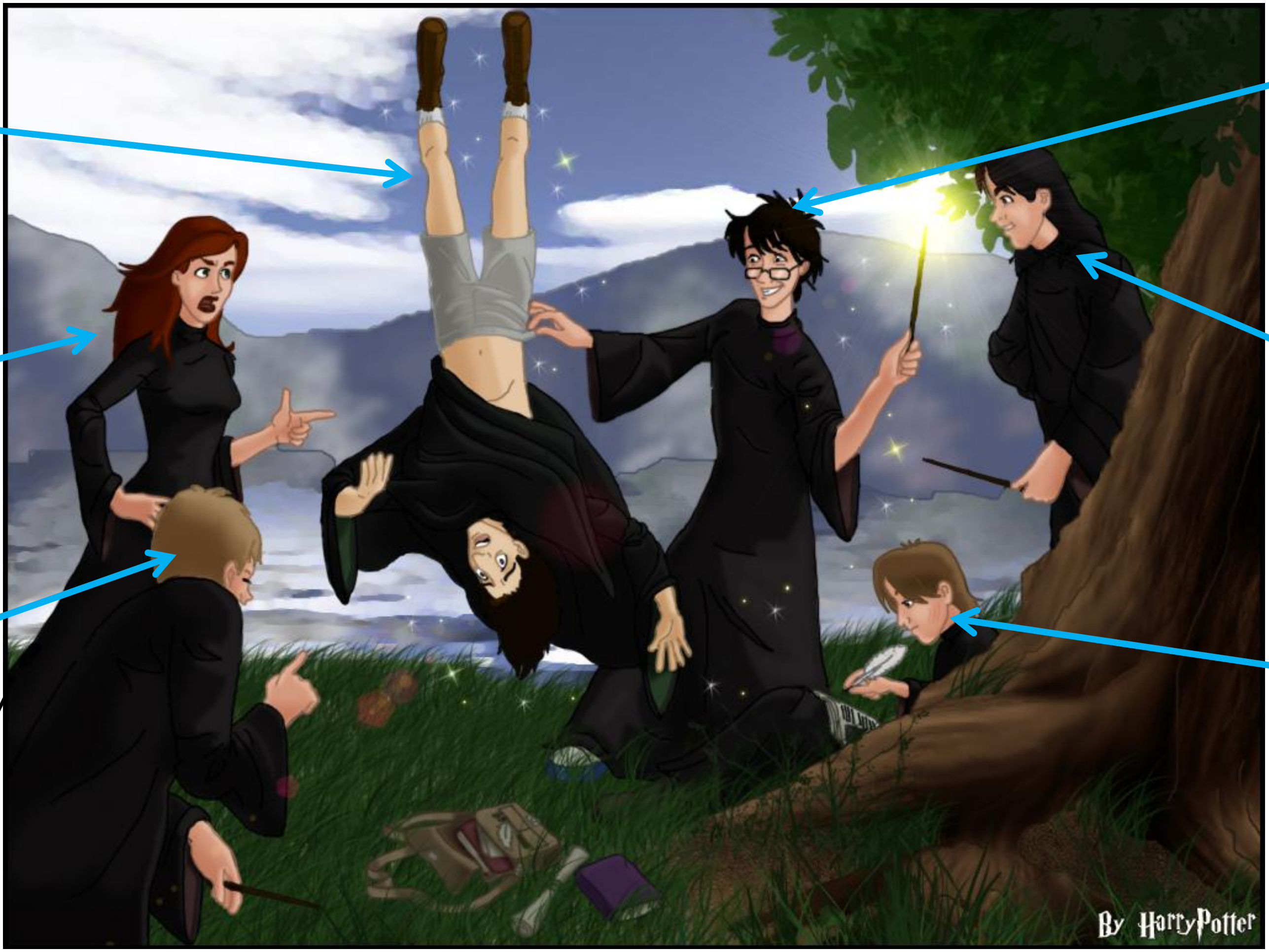
2012 NAACL HLT MONTREAL

WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

LEARNING FROM BULLYING TRACES IN SOCIAL MEDIA

Jun-Ming Xu, Kwang-Sung Jun, Xiaojin Zhu, Amy Bellmore*
Dept. of Computer Sciences, Educational Psychology*
University of Wisconsin-Madison

Snape's Worst Memory



Victim:
Severus Snape

Defender:
Lily Evans

Reinforcer:
Peter Pettigrew

Bully:
James Potter

Assistant:
Sirius Black

Bystander:
Remus Lupin

http://harry-potter-spain.deviantart.com/art/Snape-s-Worst-Memory-27310861

Bullying (Peer Victimization)

Across a national sample of students in grades 4-12 in the U.S.,

38% reported being bullied by others

32% reported bullying others [Vaillancourt et al., 2010]

More students involved as assistants, reinforcers , bystanders…

Multiple forms:

physical

relational

verbal

Venues: physical world, online (cyber-bullying)

Bullying Hurts

- Symptoms of Victims
  - Interpersonal problems
  - Depression, anxiety, loneliness, low self-worth
  - Absent from school more often and lower grade
    - Every day, about 160,000 kids stay home from school because of the fear of being bullied [The U.S. CDC]
  - Lethal school violence and suicide
    - Bullying victims are between 2 to 9 times more likely to consider suicide than non-victims [Kim et al., 2009]

Limitations of the State-of-the-Art

- Traditional social science studies are handicapped by data scarcity
  - Small sample size
  - Low/no temporal resolution
  - Time consuming

- Computational study is largely unexplored
  - Only a few studies on cyber-bullying, overlooked other bullying episode

Bullying Traces in Social Media

Bullying trace: social media post talking about actual bullying episode (in physical world or online)

- Reporting a bullying episode: *"some tweens got violent on the n train, the one boy got off after blows 2 the chest... Saw him cryin as he walkd away :(bullying not cool"*

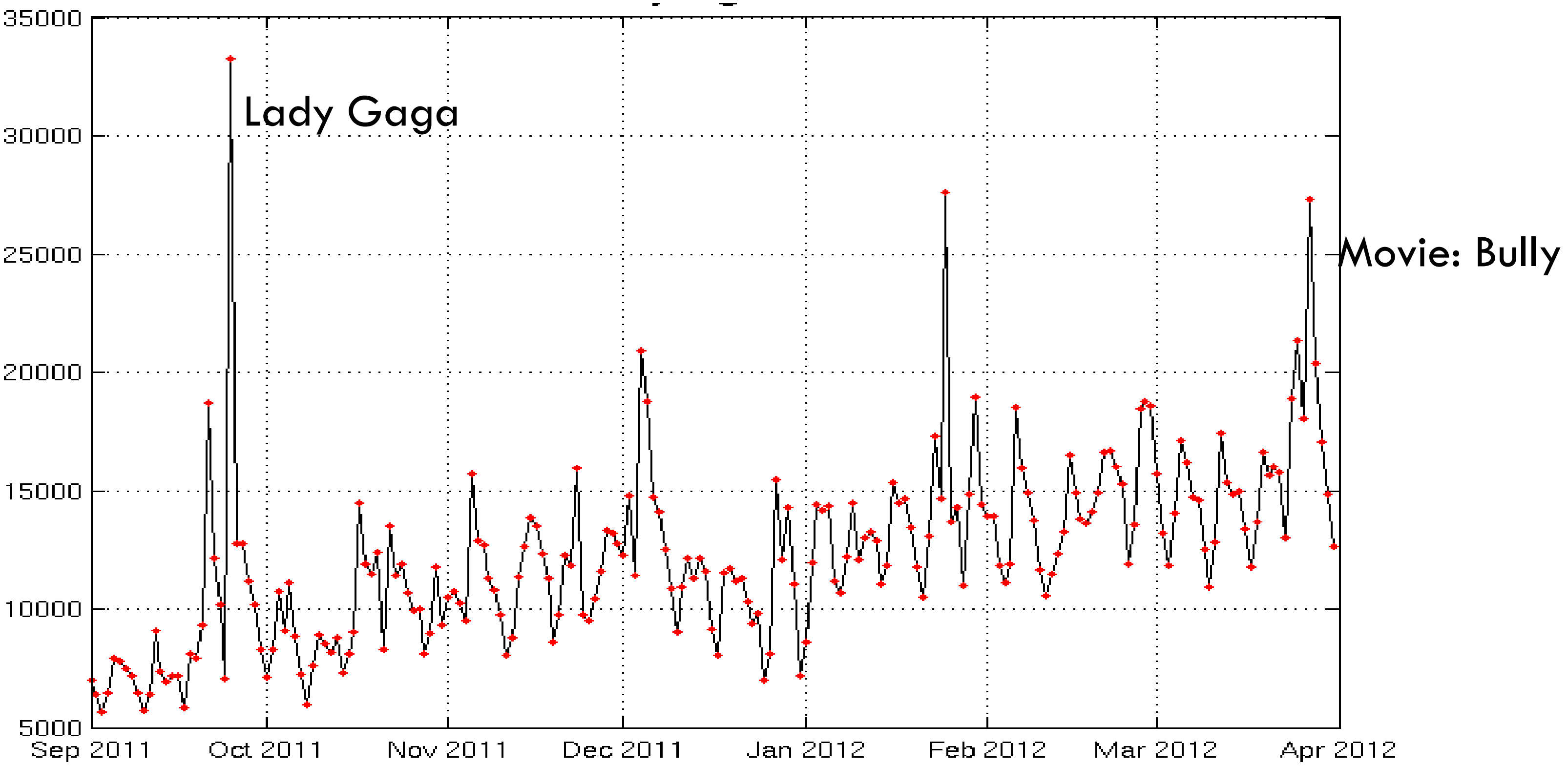
- Accusing someone as a bully: *"@USERNAME i didnt jump around and act like a monkey T T which of your eye saw that i acted like a monkey :(you're a bully"*

- Revealing self as a victim: *"People bullied me for being fat. 7 years later, I was diagnosed with bulimia. Are you happy now?"*

- Cyber-bullying direct attack: *"Lauren is a fat cow MOO BITCH"*

Bullying Traces in Social Media

Daily number of bullying traces collected



Pros: Large sample size, High temporal resolution, Easy to collect

Cons: Unknown biases

Questions NLP Can Help with

Is the post a bullying trace or not?

Text Categorization

Who are the participants? What are their roles?

Role Labeling

How do they feel during the episode?

Sentiment Analysis

What are people saying about bullying?

Latent Topic Modeling

……

Manually Labeled Training Data

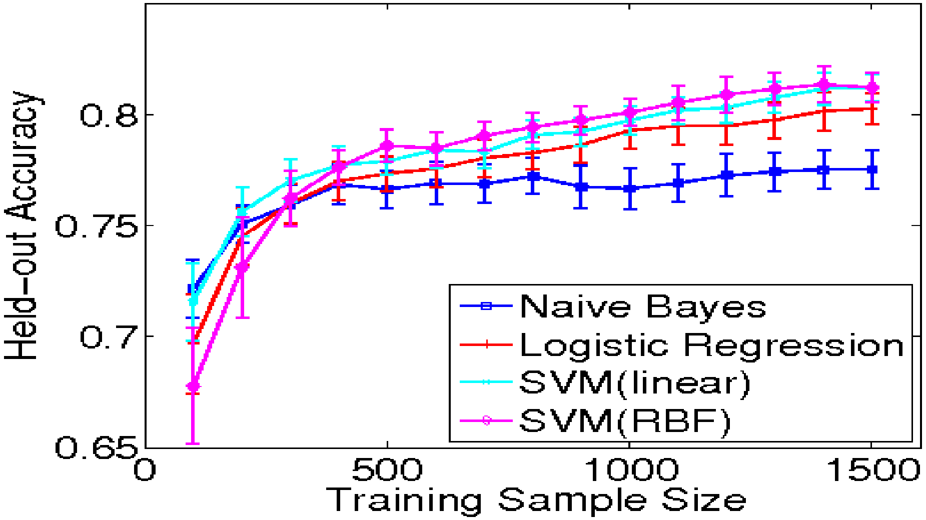
- Collected from Twitter Streaming API
  - Each has keyword such as “bully”, “bullied”, or “bullying”
  - Re-tweets are removed

- Annotated by experienced experts
  - Is it a bullying trace or not $(\kappa = .72)$
  - Bullying roles of author and person mentions $(\kappa = .61)$
  - Is the bullying trace written jokingly $(\kappa = .81)$

Task A: Text Categorization

- Classify bullying or not
  - 684 out of1762 posts in enriched dataset were bullying traces
  - *"Is Google + the new BULLY PLAYGROUND?"* (no episode)
  - *"You know what? I've never seen a Bully beat a kid up for their lunch money lol."* (the episode did not happen)
- Standard procedure
  - Tokenization (emoticon, hash-tag, user mention, url)
  - Bag-of-words representation (unigrams and bigrams)
  - SVM (RBF kernel)
  - Tune parameters with 10-fold cross-validation

Task A: Text Categorization (cont.)



Majority class baseline
Accuracy 61%

Future Work:

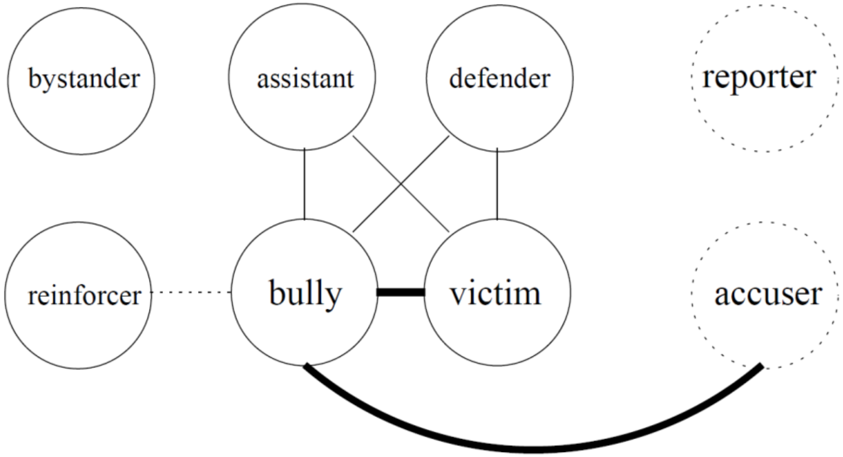
Apply the trained model to other social media stream

Keyword filtering -> Other forms

Twitter -> Facebook, google+,…

English -> Other language, Weibo

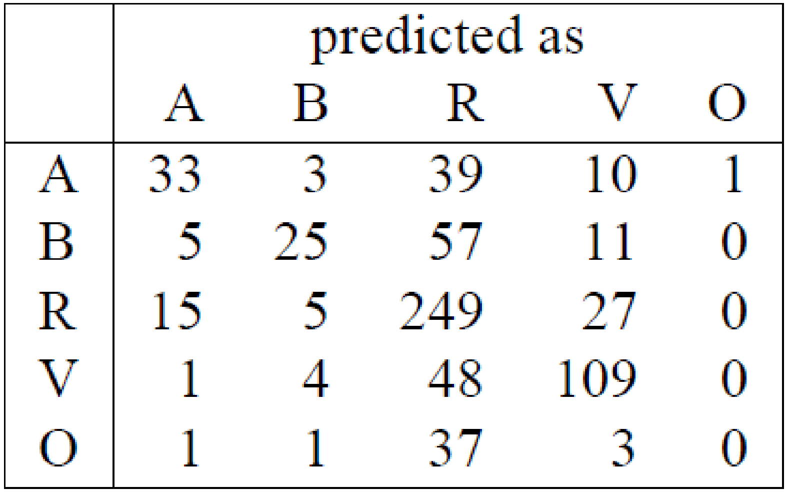
Task B: Role Labeling



AUTHOR(R): "*We(R) visited my cousin(V) today & #Itreallymakesmemad that he(V) barely eats bec he(V) was bullied . :(I(R) wanna kick the crap out of those mean kids(B).*"

Task B-1: Author's Role Labeling

Multi-class text categorization task



	predicted as				
	A	B	R	V	O
A	33	3	39	10	1
B	5	25	57	11	0
R	15	5	249	27	0
V	1	4	48	109	0
O	1	1	37	3	0

A: Accuser

B: Bully

R: Reporter

V: Victim

O: Other

Accuracy: 61%

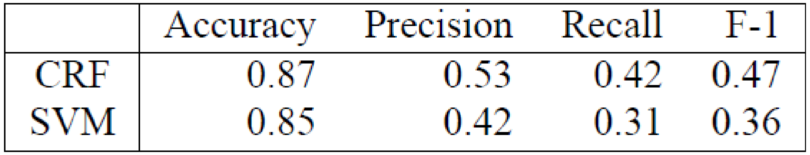
Baseline acc: 43%

Future work:

- Take advantage of self-mention
- Jointly classify many tweets authored by the same person

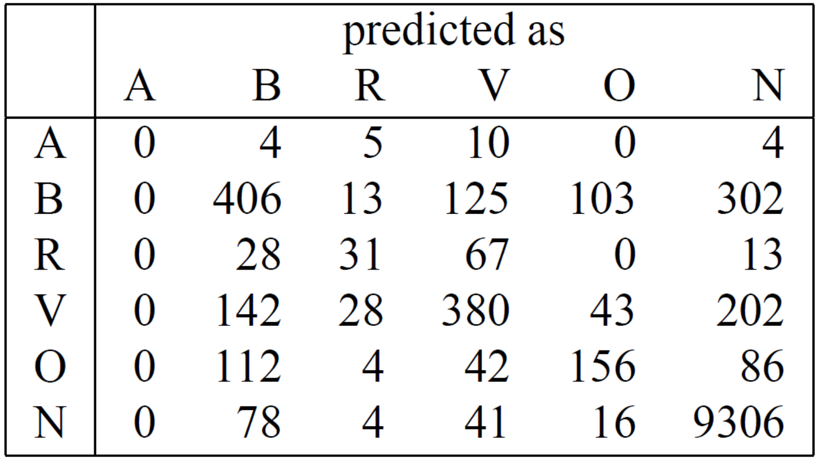
Task B-2: Person-Mention's Role Labeling

Sequential tagging



	Accuracy	Precision	Recall	F-1
CRF	0.87	0.53	0.42	0.47
SVM	0.85	0.42	0.31	0.36

A: Accuser	B: Bully
R: Reporter	V: Victim
O: Other	N: Not a person



	predicted as					
	A	B	R	V	O	N
A	0	4	5	10	0	4
B	0	406	13	125	103	302
R	0	28	31	67	0	13
V	0	142	28	380	43	202
O	0	112	4	42	156	86
N	0	78	4	41	16	9306

CRF

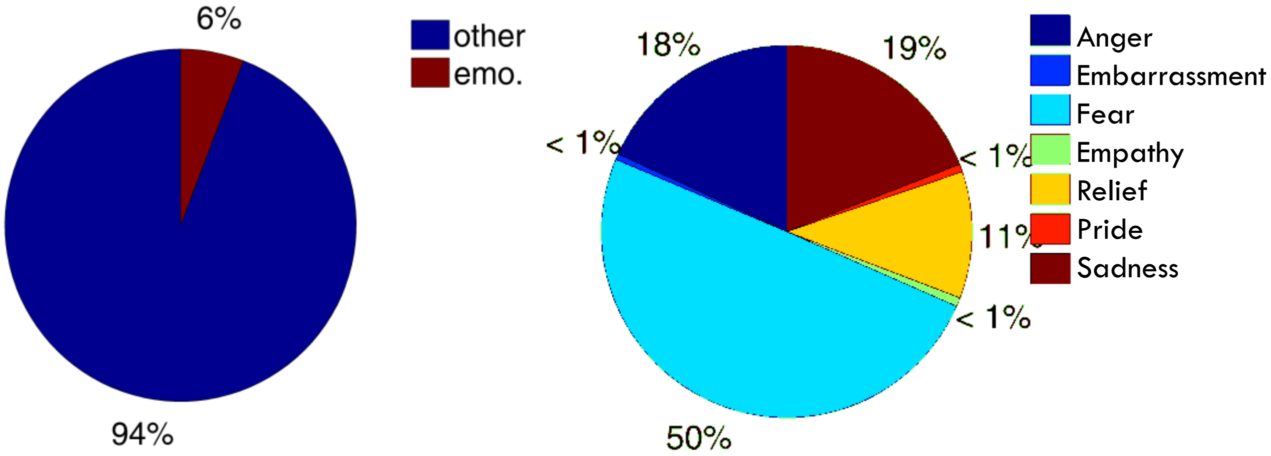
Future work:

Recognize person mentions: "sister", "teacher", "girls"…

Train NER on informal tweets

Task C: Sentiment Analysis

Emotions: a wide range, some in extremes



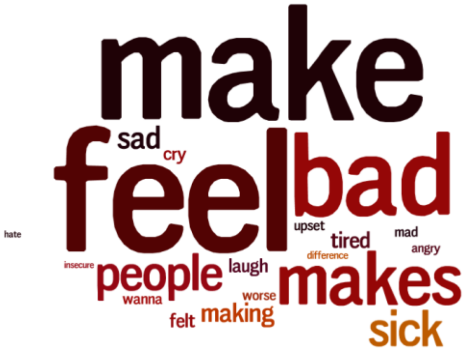
Teasing: lacking of severity, orthogonal to other emotions

"I may bully you but I love you lots. Just like jelly tots!"

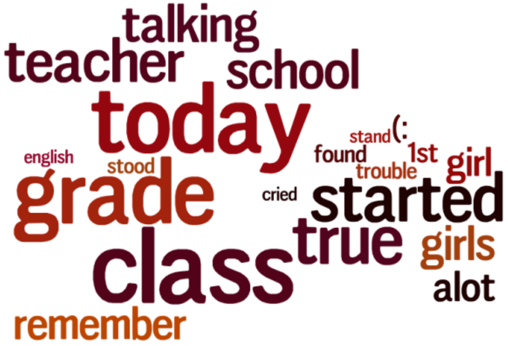
"@USERNAME lol stop being a cyber bully lol :p."

Binary text classification as in Task A (accuracy 0.89, baseline 0.86)

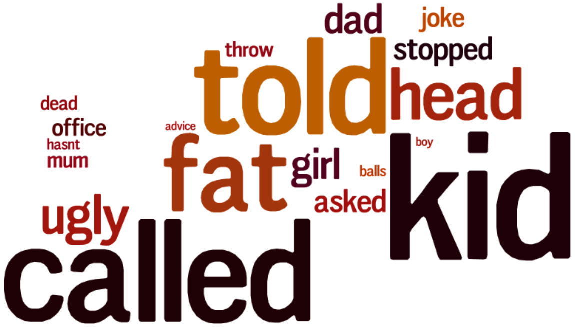
Task D: Latent Topic Modeling



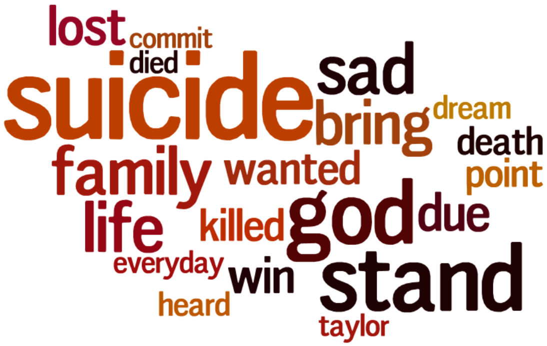
“feelings”



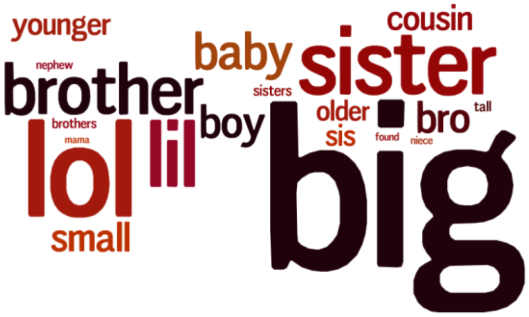
“school”



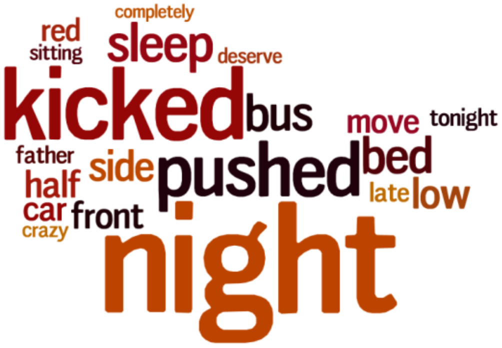
“verbal bullying”



“suicide”



“family”



“physical bullying”

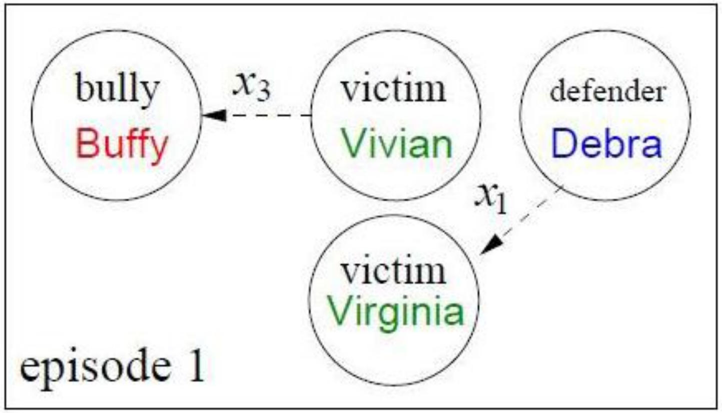
Future Works

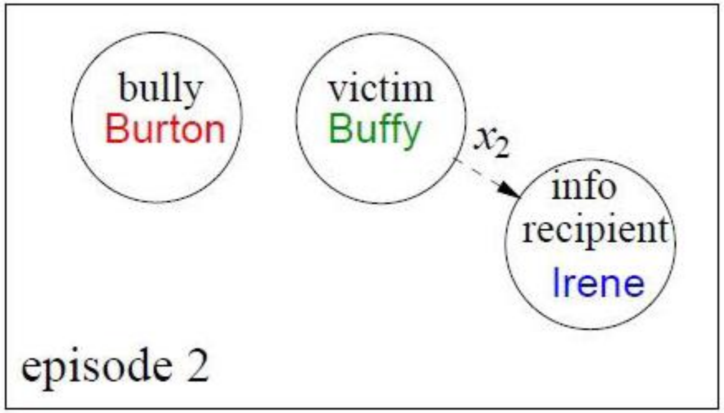
Recovering the whole structure of an episode

Debra: "*Virginia, I heard Buffy call you and Vivian fat—ignore her!*"

Buffy to Irene:"*Burton picked on me again because I'm only 5 feet*"

Vivian: "*Buffy I'm not fat! Stop calling me that.*"

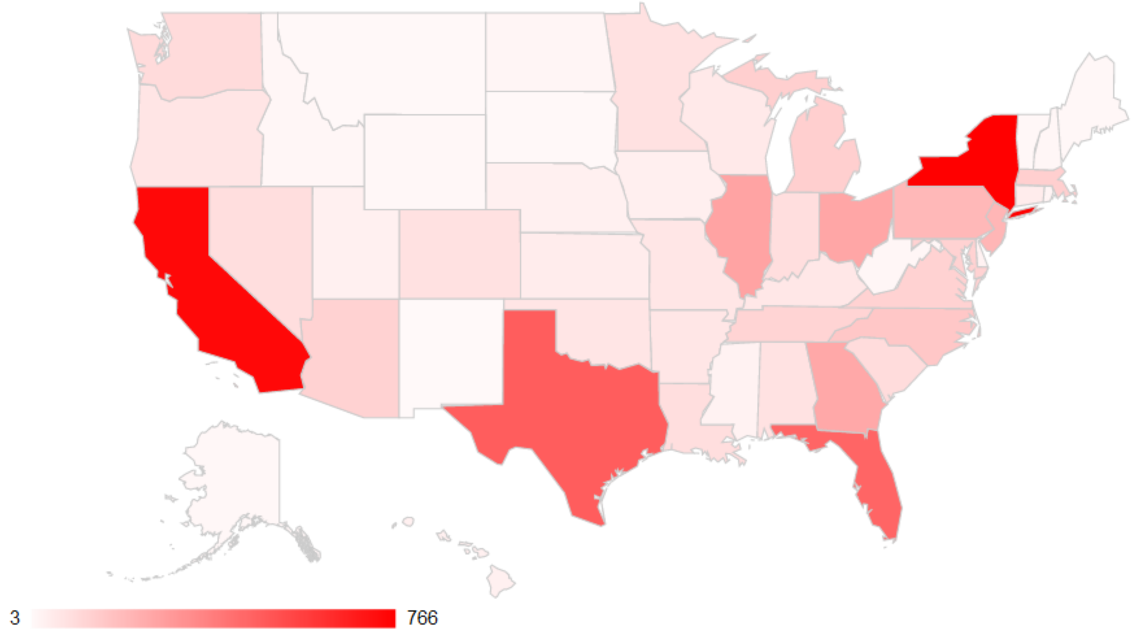




Future Works (Cont.)

Mild intervention: showing a bullying intensity map

- Help victim to cope
- Raise public awareness



Conclusions

Social media is an important data source for the study of bullying

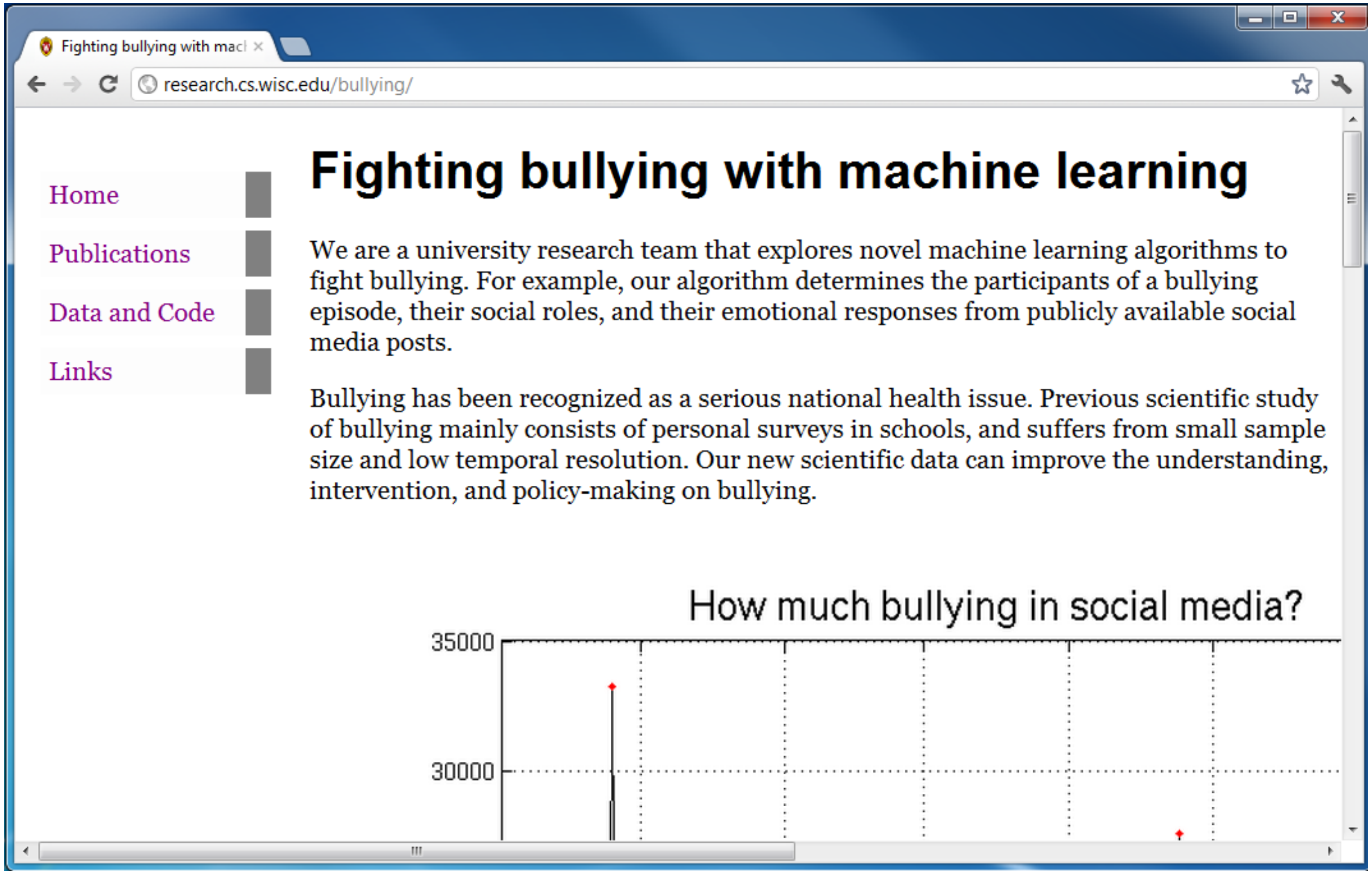
NLP community can contribute more

Data and code available online

http://research.cs.wisc.edu/bullying

Thanks!

http://research.cs.wisc.edu/bullying



Fighting bullying with machine learning

We are a university research team that explores novel machine learning algorithms to fight bullying. For example, our algorithm determines the participants of a bullying episode, their social roles, and their emotional responses from publicly available social media posts.

Bullying has been recognized as a serious national health issue. Previous scientific study of bullying mainly consists of personal surveys in schools, and suffers from small sample size and low temporal resolution. Our new scientific data can improve the understanding, intervention, and policy-making on bullying.