Fast Learning for Sentiment Analysis on Bullying

Jun-Ming Xu, Xiaojin Zhu*, Amy Bellmore

Dept. of Computer Sciences
Dept. of 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

Bullying (Peer Victimization)

 Venues: physical world, online (cyber-bullying)

 Students in grades 4-12 in the US [Vaillancourt et al., 2010]:
  - 38% reported being bullied by others
  - 32% reported bullying others
  - More involved as assistants, reinforcers, bystanders
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 Traditional Study

Traditional social science studies of bullying: relying on student self-reports
- Small sample size
- Low/no temporal resolution
- Time consuming

Computational approach is largely unexplored
- Only a few studies on cyber-bullying, overlooked other bullying episode
Bullying Traces in Twitter

**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 (~4%):** “Lauren is a fat cow MOO BITCH”
Finding Bullying Traces

• Collected from Twitter Streaming API
  – keywords “bully”, “bullied”, “bullying”...
  – remove re-tweets

• Training set annotated by domain experts

• Machine learning
  – Is the post a bullying trace or not? Text Categorization
  – Who are the participants? Role Labeling
  – How do they feel during the episode? Sentiment Analysis
15,000 Bullying Traces a Day in Twitter

[Graph showing the number of bullying traces on Twitter over the months of 2011 and 2012, with spikes aligned with the release of the Lady Gaga movie 'Bully']
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).”
Latent Dirichlet Allocation on Bullying Traces

“feelings”

“school”

“verbal bullying”

“suicide”

“family”

“physical bullying”
Bullying Sentiment Analysis

• Someone affected by a bullying episode
• feeling extreme emotions (anger, sadness, …)
• they could be a danger to themselves or others (suicidal, violent, …)
• (Eventual goal) Can machine learning recognize such extreme cases?
• (Initial study) Can it recognize any emotion?
7 Emotions Suggested by Domain Experts

1. **Anger**: “He is always laughing at me because he is a bully damnit! #Ashley”

2. **Embarrassment**: “@USER everyone is bullying me because I couldn’t find the word peach in a crossword. It’s 1am”

3. **Empathy**: “@USER I’m sorry you get bullied. I’m really surprised at how many people this has happened to. #bulliesSuck”

4. **Fear**: “I was being bullied and I didn’t want to go to school really I would throw fits every morning and I hope that Michel sees this”

5. **Pride**: “Everyone on this earth is a bully, except me. Because I’m perfect. #jillism”

6. **Relief**: “@USER I was rambling and then... I cried. Like, CRIED. He was touched! APC helped me thru the teasing and bullying man...”

7. **Sadness**: “things were bad when I was younger I got bullied so much because of my disabilities I don’t want the same thing happening to my brother.”
For each emotion, collect 5 bag-of-words:
1. Synonyms
2. Wordnet subtree
3. Twitter search results of each word above
4. Union of all above

These 7*5 BOWs \( v_1, \ldots, v_{35} \) are feature extractors. Document \( X \) represented by 35-dim \( (X'v_1, \ldots, X'v_{35}) \)

For each synonym of an emotion \( Y \), download its Wikipedia webpage \( X \rightarrow 964 (X,Y) \) training items.

Train 7-class SVM (RBF)

Apply to 3 million (8 months) bullying trace tweets \( X \)
Fraction of Emotional Bullying Traces

94% Not emotional

Emotional
6%

18% Anger
19% Embarrassment
< 1% Fear
11% Empathy
< 1% Relief
< 1% Pride
< 1% Sadness

("Fear" over represented. Can be improved by more labeled data)
Ethical Discussions

• Say our community’s algorithms predict individual’s emotional risks via social media.
• Give population statistics to policy makers, hope they are better informed?
• Alert schools / parents / police? How?
• Liability of prediction (in)accuracy?
• Legality of monitoring?
SOCIAL MEDIA
Scouring Twitter for signs of bullying

[University of Wisconsin-Madison News]