SUNY The State University of New York

UNIVERSITY AT ALBANY State University of New York

Characterization, Detection, and Mitigation of Cyberbullying

Tutorial @ ICWSM 2018

Charalampos "Harris" Chelmis

Daphney–Stavroula Zois

{cchelmis, dzois}@albany.edu

@Cchelmis

College of Engineering and Applied Sciences
UNIVERSITY AT ALBANY State University of New York

Organizers

Charalampos Chelmis
Assistant Professor in CS
University at Albany

– Network Science
– Big Data Analytics

http://www.cs.albany.edu/~cchelmis/

IDIAS

Daphney–Stavroula Zois
Assistant Professor in ECE
University at Albany

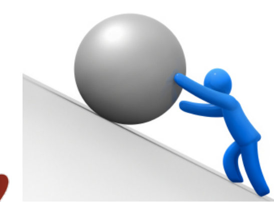
– Statistical Signal Processing
– Machine Learning

https://www.albany.edu/~dz973423/

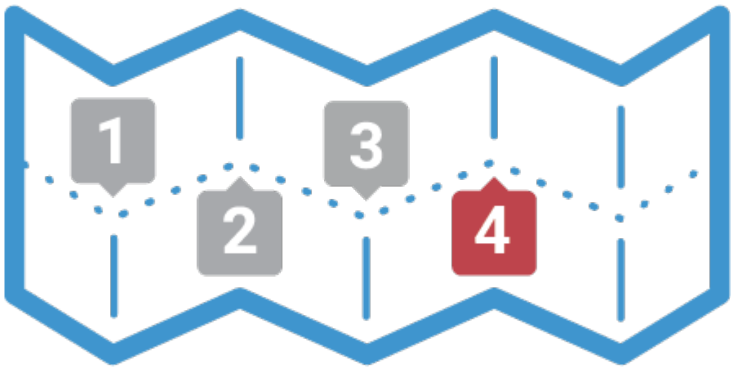
IMAgINE Lab

Tutorial Objectives

- Overview the **state of the art**
  - Provide audience an interesting emerging area to work in
  - Discuss how advances across domains can be useful in advancing the field
- Describe some of the **open problems** and **challenges**
  - Provide audience with a thought provoking description of heterogeneous **factors** that may drive cyberbullying behavior
  - Recognize the broad variety of challenges and pitfalls that prevent existing approaches from being deployed in the real–world
  - Discuss some major limitations around the use of commonly used **evaluation** criteria and some of their consequences
- Give us Food for Thought
  - Look critically at our work as a community



Tutorial Outline



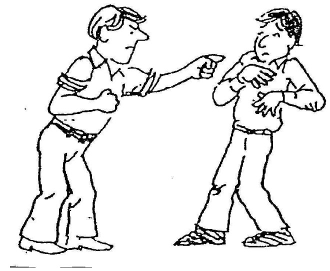
- Introduction to the problem of cyberbullying characterization, detection, and mitigation
  - Definition
  - Challenges
- Publicaly available datasets
- Characterization
- Detection (and prediction) methods
  - Data Mining and Machine Learning approaches
- Mitigation strategies
- Interactive session
  - Hands-on with a real-world dataset
- Summary & concluding remarks

Section

Cyberbullying

What does Bullying Refer to?

- **Bullying** was originally used as a term of endearment applied to either sex
  - Mid 16th century: used as a form of address to a male friend
- The current sense dates from the late 17th century
  - "The use of **force**, **threat** or **coercion** to **abuse**, **intimidate**, or aggressively **dominate** others" [Wikipedia: https://en.wikipedia.org/wiki/Bullying]
  - **Aggression** that is intentionally carried out by **one or more individuals** and **repeatedly** targeted toward a person who **cannot easily defend** herself [Olweus1978, Olweus1994]
  - **Aggressive** behavior (**repeated** or with the potential to be repeated over time) involving real or perceived power imbalance [**stop**bullying.gov]



An inherently **social phenomenon** which can only be understood in the context of social interactions

Types of Bullying

- **Physical**: hurting a person's body or possessions
  – Pushing, hitting/kicking, spitting, breaking things, making rude hand gestures, …
- **Verbal**: intimidating a victim by saying/writing mean things
  – Teasing, name-calling, inappropriate sexual comments
- **Indirect:** hurting someone's reputation or relationships
  – Backbiting and spreading of false rumors
- **Social alienation**: leaving someone out on purpose
  – Not letting someone hangout with a group or be part of a conversation
  – Telling others not to be friends with someone

Bullying on the Web

- *"**Cyberbullying** is bullying that takes place using electronic technology and communication tools"* [stopbullying.gov]
  - Cell phones, computers, …
  - Social media sites, websites, …
- *"Examples of cyberbullying include mean text messages or emails, rumors sent by email or posted on social networking sites, and embarrassing pictures, videos, websites, or fake profiles."*

Bullying on (as opposed to off) the Web (2)

- **Bullying** was once limited to physical spaces (e.g., schools or sports fields) and particular times of the day (e.g., school hours)
- **Cyberbullying** (as opposed to regular bullying):

Online:
- Relies on digital media (e.g., hurtful comments, videos and images)
- The Web offers immediate and continuous communication

Frequency:
- Cyberbullying can occur anytime, anywhere
- It can be difficult for victims to find relief

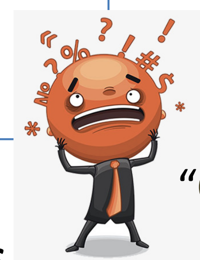
Permanency:
- Content remains (publically) accessible online unless reported and removed

Audience & Spread:
- Online social media provide a large audience, and quick (potentially wide) spread

Bullying on the Web (3)

*"**Cyberbullying** is **bullying** that takes place using electronic technology and communication tools"* [Campbell2005, Slonje2008, Vandebosch2008, Dooley2009, Erdur-Baker2010, Kowalski2012]

*"**Cyberharassment** refers to **repetitive**, invasive and anxiety provoking online interpersonal attacks"* [Li2005]



*"**Cyberstalking** is the use of electronic means to stalk or **harass** an individual, group, or organization"* [Bocij2004, Pittaro2007, Sheridan2007, Reyns2011]

Instilling fear
Emptying bank accounts

"**Cyber-aggression** refers to **one-off** (or occasional) occurrence of offensive, derogatory, harmful, or unwanted behavior using electronic means to **harm** a person or a group of people [Grigg2010, Smith2012, Corcoran2015]

Fundamental Aspects of Cyberbullying

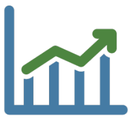
- **Repetition:** often used in the definition to exclude occasional acts of aggression directed at different people at different times
  - Ongoing feelings of stress about an incident may be considered repetitive even though the act occurred only once
  - 50% of victims do not consider the frequency of occurrence to be important
  - Can be "easily" quantified by measuring the number of text messages, e-mails, tweets, Instagram posts …
  - A single aggressive act (e.g., uploading an embarrassing picture on the Web) can result in **continued** ridicule and humiliation for victims
  - Not all actions have equal **effects** in inflicting harm
    - e.g., threatening comment vs an embarrassing picture
  - Information posted online can be widely **disseminated** (repetition may not be as important)

Fundamental Aspects of Cyberbullying (2)

- **Power Imbalance:** Refers to observed or perceived personal or situational characteristics to exert control over a victim or to limit the victim's ability to respond or stop the aggressive behavior
  - Can be social, psychological, or physical
  - One of the distinguishing features of cyberbullying is the inability of victims to get away from it
    - May result in feelings of powerlessness for the victim
    - Not knowing the identity of the bully may increase feelings of frustration and powerlessness
  - Anonymity appears to be an important feature of cyberbullying for perpetrators who would not engage in offline bullying
  - Difficult to conceptualize and assess in online interactions
    - Only few have explicitly measured it [Dooley2009]

Why Cyberbullying Matters

- **Early** detection of **cyberbullying content** becomes of utmost importance



Growing Number of Incidents

- The time users spend in online social media is growing rapidly [Benevenuto2009, Tokunaga2010]
- & so is the number of users abusing the Internet to harass, threat, and frighten others [Tokunaga2010, Jones2013, Al-garadi2016, Anderson2017]

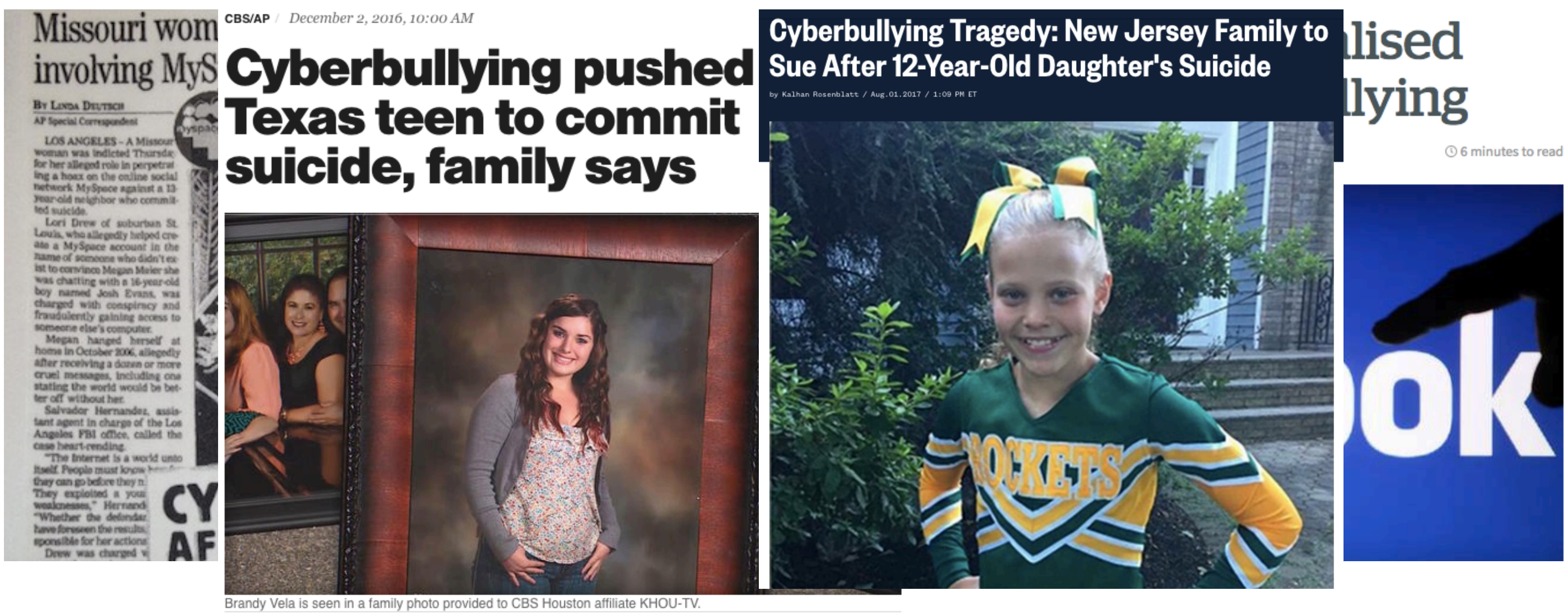
Potentially Detrimental Effects

- Learning difficulties
- Psychological suffering and isolation
- Escalated physical confrontations
- Suicide

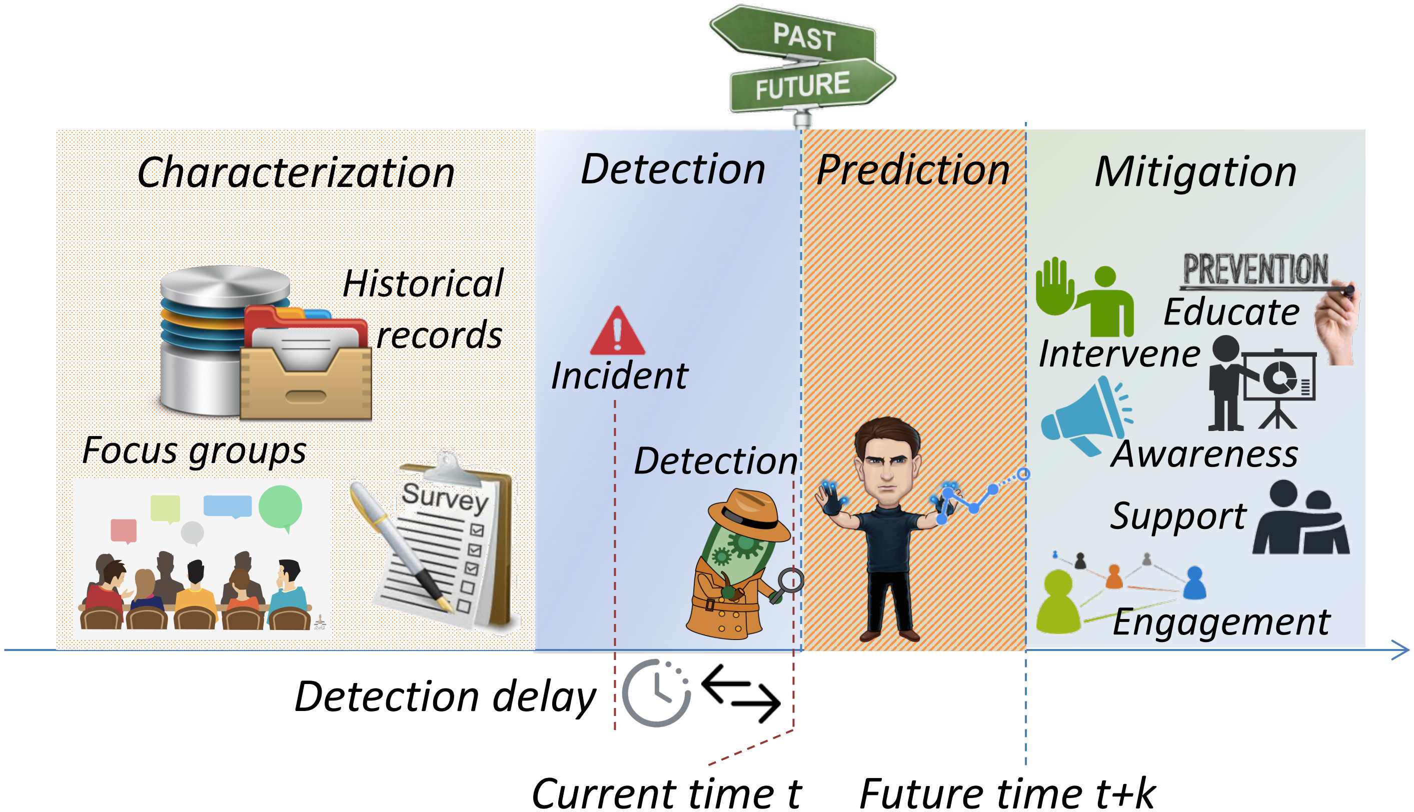
Why Cyberbullying Matters

- Over ½ of adolescents and teens have been **bullied** online
    - About the same number have engaged in cyber bullying!
- > **1 in 3** young people have experienced cyberthreats
- >**25%** of adolescents and teens have been cyberbullied **repeatedly**
- Only **1 in 10** teens tells a parent that they have been a victim!

Source: https://www.ditchthelabel.org/research-papers/the-annual-bullying-survey-2017/



Broad Themes of Cyberbullying Research



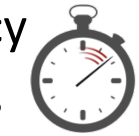
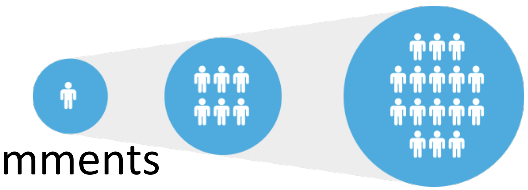
Cyberbullying Research Pipeline

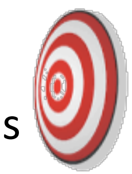
- Problem definition
  - Is the goal to characterize, detect, predict or mitigate?
- Data acquisition
  - Are there existing datasets? If so, what is the data source?
  - How is the data collected (e.g., using streaming 1% vs. Twitter firehose)
  - Is the data representative?
  - Is the dataset balanced or skewed?
  - Are labels available / Do we need to annotate the data?
    - How are these produced (manually by experts vs. automatically)
- Feature selection
  - Are there multiple classes of (heterogeneous) features? If so, what are these?
  - What kind of information do features capture?
  - What is the information gain from each feature?
  - Would dimensionality reduction be preferable?

Cyberbullying Research Pipeline (2)

- Method selection:
  - Is the data used for exploratory analysis/characterization?
  - Is a specific hypothesis being tested?
  - What are the main metrics to be improved (e.g., Precision/Recall)?
  - Which metric is more important (e.g., is recall more desirable)?
  - Is the method suitable for the task?
- Validation & evaluation:
  - Evaluation on training set: does the model accurately model training data?
  - Evaluation on testing set: does the model generalize well to new data?
  - What type of errors does the model make?
  - Does accuracy hold across folds/datasets/platforms?
- Interpretation
  - Which features best explain model performance?
  - What are the data &/or model limitations?
  - Are findings consistent with the literature? If not, why?

Ideal Cyberbullying Detection System

- High detection accuracy 
  - Precision vs. Recall vs. …
- Small detection latency 
  - Every second counts
- High scalability 
  - Millions of users, Billions of comments
- Adaptability 

  - Hate speech/profane keywords may change as language evolves 
  - Technology progresses fast
  - Notion of cyberbullying may change over time
  - Bullying follows evolutionary principles [Rigby2004, Espelage2012, Volk2012]
- Early prediction 
  - Detection tries to determine whether cyberbullying has occurred after the fact
  - Prediction tries to determine if an event is likely before it even happens

Challenges With Cyberbullying Research

- Data collection and sampling bias
  - APIs limitations
    - e.g. Twitter's streaming API limits access to a small number of tweets as compared to Twitter's Firehose [Morstatter2013, González-Bailón2014]
    - Geo-code filtering returns a nearly complete set of geo-tagged tweets

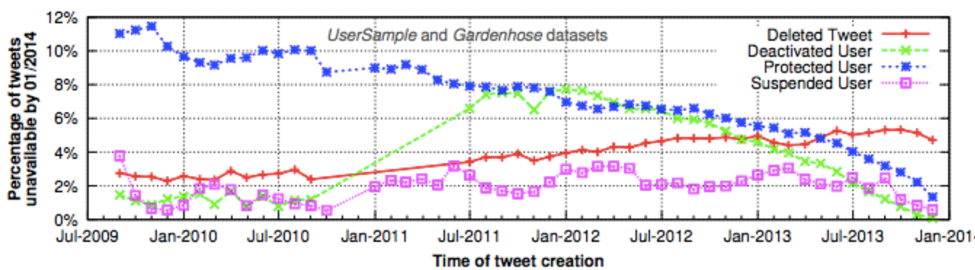
      Not all content is geo-tagged
  - Keyword– & lexicon–based sampling [González-Bailón2014]
    - The choice of keywords/hashtags specifies the boundaries of data collection
    - May cause relevant data to be missed
    - May lead to overrepresentation of one class
    - Use machine learning approaches such as [Raisi2017] to identify new lexical indicators
  - Sampling method [Granovetter1976, Ahmed2012, Morstatter2013, Ahmed2014]
    - Often snowball sampling [Biernacki1981, Atkinson2001]
  - Over–emphasis of a single platform (e.g., Twitter) [Tufekci2014]
    - Findings may be biased to a certain population using the platform
    - User demographics may differ across platforms

Challenges With Cyberbullying Research (2)

- Data cleansing and annotation
  - Outliers (e.g., non– or highly–active users) may hurt the ability of a classification model to discriminate between bulling vs. normal
  - Filtering outliers can introduce biases
  - Label errors can cripple the accuracy of machine learning models [Frénay2014]

- Data (un)availability with time [McCreadie2012, Liu2014]

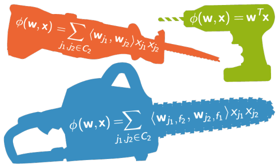
  - Due to terms of use, deleted content by users, suspended accounts), …

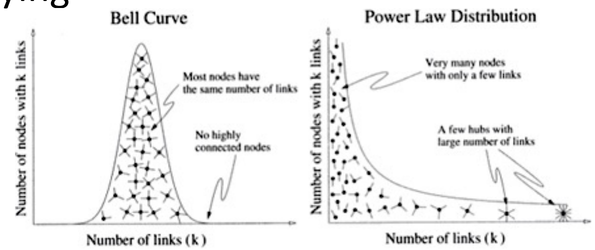


  - More data don't necessarily improve performance [Boivin2006, Dalessandro2014]
    - If data is biased adding more of it won't likely help
    - In general, more complex models are likely to benefit more from larger datasets

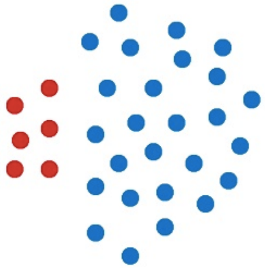
Challenges With Cyberbullying Research (3)

- Feature engineering
  - Many feature selection methods rely on machine learning classifiers
    - May not be robust across datasets
  - Bullying is well studied; good indicators of bullying can be reused
    - Identify new features likely to be indicative of cyberbullying
  - Often features follow a power–law



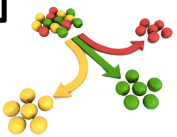


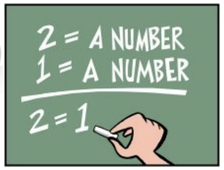
- Severe class imbalance
  - Cyberbullying content is quite rare
  - Even large–scale datasets might contain just a few samples
  - Use crowdsourcing towards developing labeled datasets
    - Often difficult, even for a human, to consistently distinguish between different types of abuse
    - Optimizing the number of annotators employed, their payment, and time for the annotation process to complete is nontrivial
    - Use sampling approaches (e.g., [Chawla2002] or [Founta2018])



Challenges With Cyberbullying Detection

- Objective
  - **Prioritization**: promote certain content at the expense of others
    - The ranking and weighting criteria should be scrutinized
  - **Classification**: derive the class of content/user based on attributes
    - One-off classification vs. tracking
  - **"Guilt by association"**: determine which user is similar to others based on content/activity/interactions
    - Is the association interpretable?



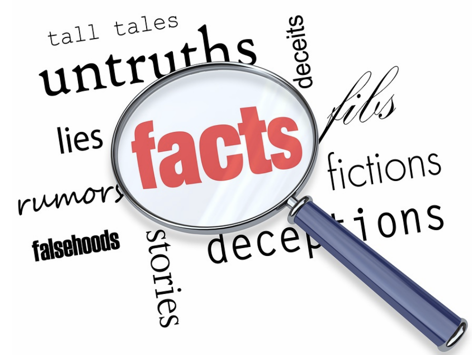


- Evaluation
  - Which metrics are appropriate?
  - What are the costs of different errors (e.g., false positives vs. false negatives)?

- Mitigation may become a strong form of social influence
  - Create a feedback loop to adjust models based on mitigation strategies

Challenges with Cyberbullying Mitigation

- Loss of privacy due to monitoring, forwarding to third parties (e.g., parents/admins), or removal of messages
- Conformance of bullies to education
- Willingness of victims to report cyberbullying incidents
- Willingness of bystanders to intervene
- False reporting of cyberbullying instances
- Accuracy of cyberbullying detection tools
- Timeliness of detection and reporting (mitigation will be obsolete)



Section

Datasets

Datasets

- Publically available datasets can:
  - Significantly accelerate the field
  - Enable direct comparison between state-of-the-art methods
  - Ease the interpretation of results as their properties are studied more
  - Be scarce (c.f. data unavailability with time challenge)
  - Result in a hyper-focus on popular datasets (just because they exist)
  - Be bad proxies of society (c.f. Data collection & sampling challenges)
- Giving back!
  - We are developing a website to assemble & provide a comprehensive index of:
    - Annotated real-world cyberbullying data sets
    - Lexicons for cyberbullying research
  - Share the word: **#CBDatasetsProject**

Datasets (2)

- Formspring
  - Q&A based online social network
  - The ability of users to post questions anonymously opened the doors for harassment/cyberbullying
  - Populated mostly by teens and college students
  - High percentage of bullying content
- Dataset
  - 18,554 Formspring users were randomly selected
  - Profile information for each user was collected
  - Questions and answers from users' profiles were crawled
  - Annotations were acquired from Amazon's Mechanical Turk
  - Both labeled and unlabelled datasets
  - Available at: http://www.chatcoder.com/DataDownload

Datasets (3)

- Myspace
  - The largest online social networking site in the world, from 2004 to 2010
  - Thread-style forum conversations
    - Posts can be lengthy (unlike other online social networks)
- Dataset:
  - Focuses on direct bully–to–victim cyberbullying instances
  - Unlabeled dataset of ~128K users and associated posts
  - Smaller labeled dataset also available
    - Ground truth provided by undergraduate research assistants
    - Labeled cyberbullying if at least two humans flagged content as such
    - Labelers also identified the type of cyberbullying & the exact lines involved
  - Available at: http://www.chatcoder.com/DataDownload

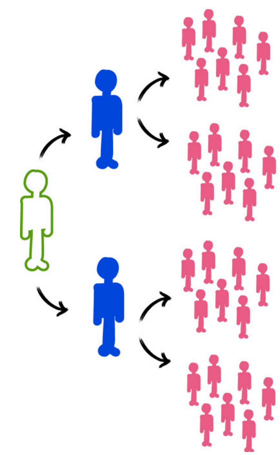
Datasets (4)

- Ask.fm
  - Based on Formspring's interaction model
  - Quite popular among young users
  - Allows for semi-anonymous communication
    - Users can anonymously communicate with known recipients
  - Questions are directed to a particular individual
- Data collection method:
  - Queried ASKfm through Google for variations of terms "go kill yourself" and "go die"
  - Performed snowball sampling:
    - Crawled users who interacted with the original Google search result users

- Unlabeled dataset
  - 261K users and ~ 3M question–answer pairs
- Available at: https://sites.google.com/site/cucybersafety/home/cyberbullying-detection-project/dataset

Datasets (5)

- Instagram
  - Media-based mobile social network that allows users to post and comment on images/videos
  - Platform with the highest reported cases of cyberbullying
- Dataset
  - ~25K public user profiles crawled using snowball sampling
  - For each public profile the following data was collected
    - Media objects/images that the user has posted
    - Their last 150 associated comments
    - Followers/followees
    - User id of each user who commented on or liked the media objects shared by the user.
  - Media sessions are scored for cyberaggression/cyberbullying
  - Labeled and unlabeled dataset
- Available at: https://sites.google.com/site/cucybersafety/home/cyberbullying-detection-project/dataset



Datasets (6)

- Vine
  - Mobile based video–sharing online social network
  - Allows users to record and edit videos, which they can share on their profiles for others to see, like and comment upon
  - Offers the opportunity to explore cyberbullying in the context of video-based communication
- Dataset
  - Collected profile information and activity data for 60K users using snowball sampling
  - ~ 652K media sessions with ≥ 15 comments
  - CrowdFlower was used to label media sessions for cyberaggression/cyberbullying
- Available at: https://sites.google.com/site/cucybersafety/home/cyberbullying-detection-project/dataset

Datasets (7)

- Twitter
  - Online news and social networking service
  - Users post and interact with short messages
- Dataset:
  - 7,321 Bullying Traces
    - Tweets collected using the Twitter streaming API
    - Each tweet contains at least one of the keywords: “bully, bullied, bullying”
  - Each tweet is labeled, participants’ bullying roles are identified, and emotion labels are provided
- Open source code
  - Code to classify
    - tweets as bullying or not
    - Given a tweet, the author's role
    - The type, form and sentiment of the tweet
- Available at: http://research.cs.wisc.edu/bullying/data.html

Datasets (8)

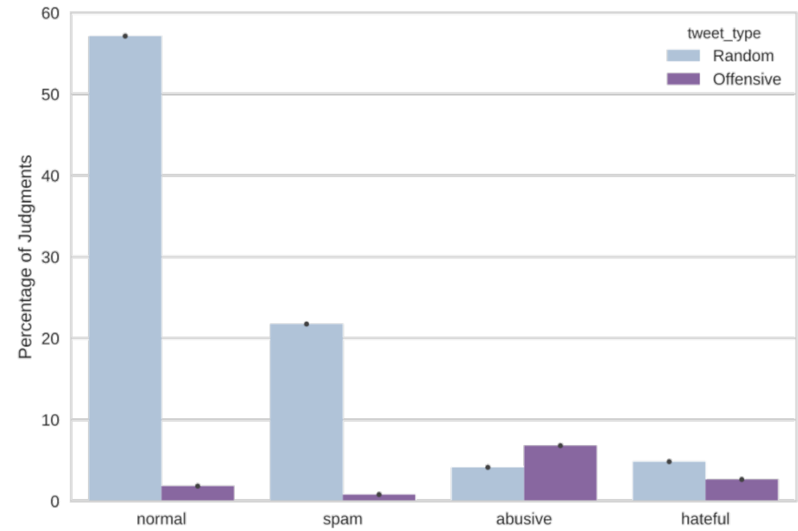
- Twitter [Rezvan2018]
- Lexicon of 737 offensive words
- Corpus of 50K tweets
  - Collected from 12/18/16 – 01/10/17
  - 10K tweets for each type with at least one lexicon item
  - ~25K tweets manually annotated
- Five types of harassment content captured:
  - Sexual
  - Racial
  - Appearance–related
  - Intellectual
  - Political
- Dataset (and lexicon) available at: https://github.com/Mrezvan94/Harassment-Corpus

Datasets (9)

- Twitter [Chatzakou2017]
- Collected 1M random tweets and a set of 650K hate-related tweets using the Twitter Streaming API
  - Hate-related tweets: posts mentioning at least one of 309 hashtags related to bullying and hateful speech
  - List hashtags was created by obtaining a 1% sample of all public tweets in a given time window and selecting all tweets containing #GamerGate
    - #GamerGate is a known large-scale instance of bullying/aggressive behavior
- Tweets from the same user were grouped based on time into sessions
- Ground truth was obtained from human annotators on CrowdFlower
- Users (not single tweets) are labeled
  - Normal, aggressive, bullying, or spammer
- Available upon request

Datasets (10)

- Annotated Twitter Dataset [Founta2018]
  - ~100k tweets
  - Each tweet is labeled as abusive/hateful/spam/normal by 5 CrowdFlower workers
    - Majority vote used for final annotation
  - Format: <
    - e.g.,
      848306464892604416,abusive
      850010509969465344,normal
      850433664890544128,hateful
      847529600108421121,abusive



  - To get the tweet text using the Twitter API
    - e.g., twitter.com/anyuser/status/850660404770590720
      https://api.twitter.com/1/statuses/show/850660404770590720.json

"THE FORCE AWAKENS: A Bad Lip Reading" (Featuring Mark Hamill as Han Solo) youtube.com/watch?v=Sv_hGI…

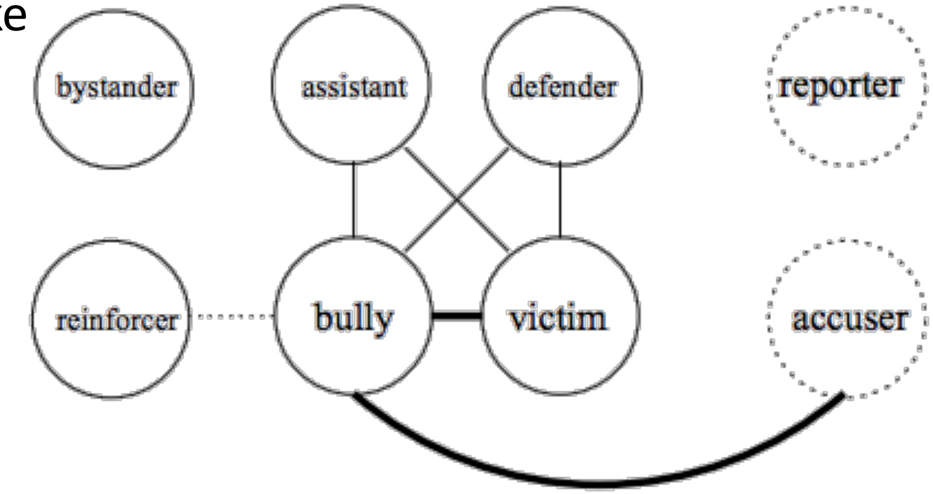
- Available at: https://github.com/ENCASEH2020/hatespeech-twitter

Section

Characterization of Cyberbullying Behavior

The Structure of a Bullying Episode

- Participants in a bullying episode take well–defined roles [Salmivalli1999, Xu2012]
  - **Bully** (or bullies)
  - **Victim** (or victims)
  - **Bystanders** (who saw the event but did not intervene)
  - **Defenders** of the victim
  - **Assistants** to the bully (who did not initiate but went along with the bully)
  - **Reinforcers** (who did not directly join in with the bully but encouraged the bully by e.g., laughing)



Note 1: More than one person can have the same role in a bullying episode

Note 2: One person can assume multiple roles in different bullying episodes

Bullying Traces in Social Media

[Xu2012]

- **Bulling traces**: content (i.e., text, images, videos) participants of a bullying episode post in online social media about the experience
  - Either in physical or cyber venues
  - Food for Thought: How does the physical world (i.e., offline interactions) impact online behavior?
  - Important Note most bullying traces are responses to a bullying experience, i.e., the actual attack is hidden from view
- Forms of bullying traces:

Reporting

Accusing

Revealing

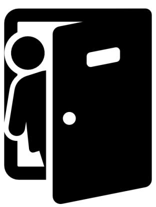
Attacking

Bullying Traces in Social Media

[Xu2012]

Reporting

Accusing



Revealing

Attacking

"@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"

"People bullied me for being fat. 7 years later, I was diagnosed with bulimia. Are you happy now?"

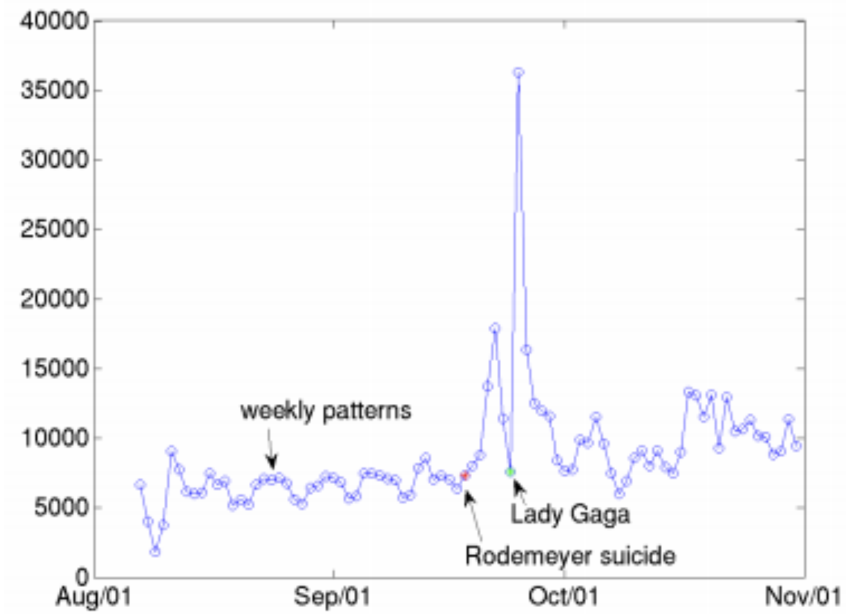
"Lauren is a fat cow MOO BITCH"

"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"

Bullying Traces in Social Media

[Xu2012]

- Bullying traces are abundant
  - By some estimates (circa 2011) ~50,000 English bullying traces per day are to be expected in Twitter
- Recall, however, the class imbalance problem
  - Frequency of bullying traces is tiny in comparison (~0.002)
- Figure shows daily pattern of bulling traces identified by classifier
- Note the weekly pattern in late August
- The small peak was caused by 14-year-old bullying victim suicide on Sept. 18
- The large peak was caused by Lady Gaga's song dedication to the victim on Sept. 24.



Using Social Media for the Study of Bullying

[Xu2012]

- Major NLP Task 1: Text Categorization
  - Need to distinguish bullying traces from other “irrelevant” social media posts
  - Often formulated as a binary text classification problem
  - The short text nature of social media posts becomes a challenge
  - **Note**: multi-class classification for fine–granularity recognition of bullying traces forms is still open
- Major NLP Task 2: Role labeling
  - A prerequisite of studying how a person’s role evolves over time
  - Goal is to classify the role of the **author** and any person **mentioned** in post
    - Labeling author’s role can be formulated as a multi-class text classification task
    - Labeling mentioned user(s)’ roles can be formulated as a sequential tagging task

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

Key TAKE AWAY: “In general, bullying role labeling may be improved by **jointly considering multiple tweets** at the episode level.”

The Five W's of "Bullying" on Twitter: Who, What, Why, Where, and When

[Bellmore2015]

- <u>Goal</u>: explore the utility of supervised Machine Learning methods for understanding bullying
  - Q1: **Who** posts/participates about/in bullying on Twitter?
  - Q2: **What** form of bullying is mentioned/used on Twitter?
  - Q3: **Why** are people posting about bullying on Twitter?
  - Q4: **Where** are people posting about bullying on Twitter?
  - Q5: **When** are people posting about bullying on Twitter?
- **Dataset**:
  - Tweets collected using the Tweeter Streaming API between September 1, 2011 - August 31, 2013
  - Used a small keyword list (bullied, bully, bullyed, bullying, bullyer, bulling, ignored, pushed, rumors, locker, spread, shoved, rumor, teased, kicked, crying)
  - Human coders labeled 7321 randomly selected tweets
- **Definition of bullying**: Any mention of bullying

The Five W's of "Bullying" on Twitter: Who, What, Why, Where, and When

[Bellmore2015]

- **Bullying tweets identification:**
  - A dictionary including all words (and all pairs of any two consecutive words) in the corpus was constructed
  - Each tweet was represented as a frequency vector
    - Number of times each word and word pair in dictionary occurred in the tweet
  - A text classifier was trained based on 7,321 human-coded tweets
    - Achieved 86% accuracy on the training set
  - Text classifier was applied on the remaining 32,477,558 tweets
    - Classified 30.07% (i.e., 9,764,583) as bullying
- **Analysis**:
  - The role of the author of every tweet classified as bullying in the training set was manually annotated as (bully, victim, bystander, defender, assistant, reinforcer, reporter, or accuser)
  - Each tweet classified as bullying was evaluated according to the five categories

The Five W's of "Bullying" on Twitter: Who, What, Why, Where, and When

[Bellmore2015]

- **Who:**
  - Trained an author role support vector machine (SVM) classifier
    - Classifier achieved 70% cross validation accuracy
  - The classifier agreed with human annotators on victims (36.01%) and reporters (32.52%)
- **What**:
  - Manually annotated the training set into:
    - General, cyberbullying, physical, and verbal
  - Classifier achieved 70% cross validation accuracy
  - Cyberbullying tweets are frequent (4.14%)
  - General tweets are the most common (95.21%)

The Five W's of "Bullying" on Twitter: Who, What, Why, Where, and When

[Bellmore2015]

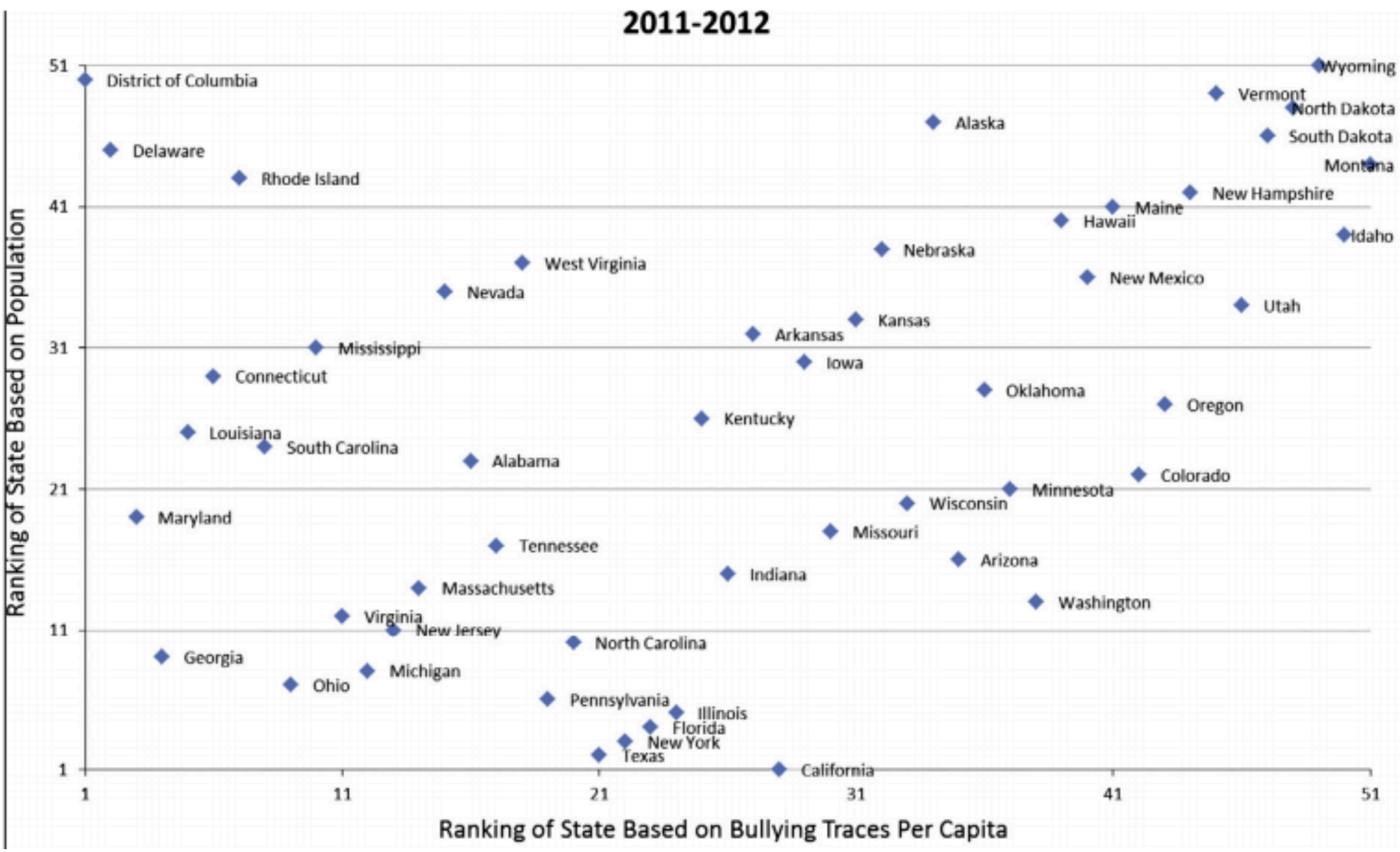
- **Why**:
  - Trained an author role support vector machine (SVM) classifier
    - Classifier achieved 72% cross validation accuracy
  - Found self-disclosure posts (54.34%) to be the most common followed by reports (28.57%), accusations (15.19%) and denials (1.90%)

Reports: Posts that described a bullying episode someone knows about, "*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.*"
Accusations: Posts that accused someone as the bully in an episode, "*@USER 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.*"
Self-Disclosures: Posts that revealed the author himself/herself as the bully, victim, defender, bystander, assistant, or reinforcer, "*People bullied me for being fat. 7 years later, I was diagnosed with bulimia.*"
Denials: Posts where the author denied a bullying role, "*@USER lol I'm not a bully man*"
Cyberbullying: Posts that were direct attacks from a bully to a victim. "*@USER really I am just cyberbullying you right now*").

The Five W's of "Bullying" on Twitter: Who, What, Why, Where, and When

[Bellmore2015]

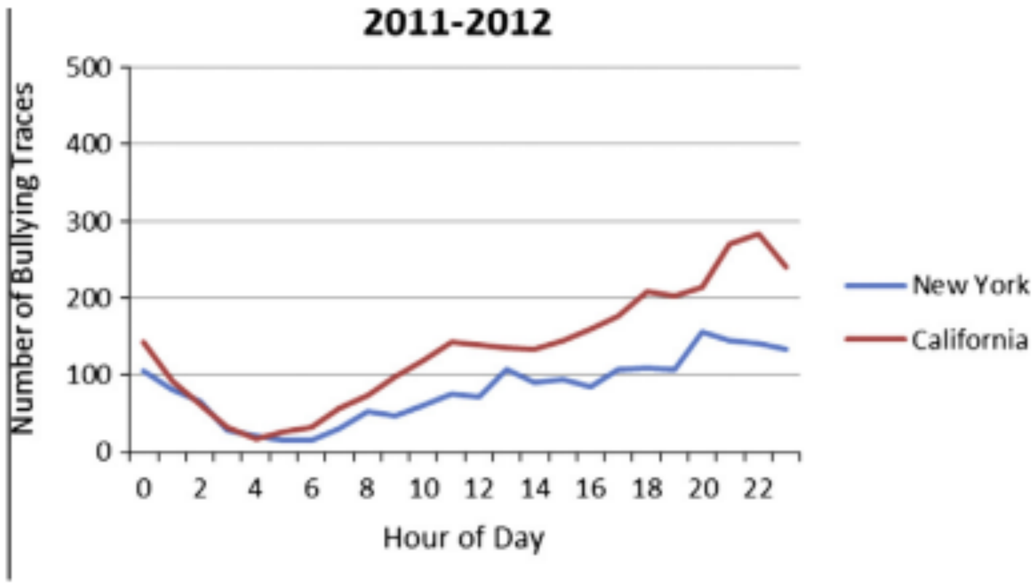
- **Where**:



The Five W's of "Bullying" on Twitter: Who, What, Why, Where, and When

[Bellmore2015]

- **When**:
  - Studied the distribution of bullying tweets across time
  - Focused on New York and California as the states with the largest number of geo-tagged bullying tweets



Analyzing Negative User Behavior in a Semi-Anonymous Social Network

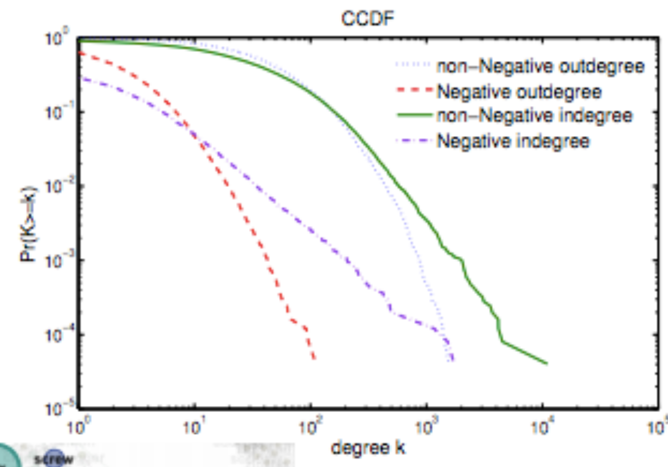
[Hosseinmardi2014corr]

- Goal: Analyze negative behavior on the semi-anonymous question+answer (QA) online social network Ask.fm
  - **Challenge**: Constructing a social graph based on friendships is impossible
  - **Focus** on the interaction graph extracted from the "likes" of comments
    - A directed edge connects user i to j if i has liked a QA in j's profile
  - **Core assumption:** repetitive negative words represent the core of abusive text posted on Ask.fm profiles
  - **Observation**: users vulnerable to negative questions were often isolated, with few "likes" and also rarely liking others' comments
- **Approach**:
  - Constructed a bipartite network such that if user i likes a QA in j's profile
    - Link from i to words on that question
    - Link from words to node j
  - Projected the bipartite network with adjacency matrix B, to the network of words $W = BB^T$ (similarly for the network of users)

Analyzing Negative User Behavior in a Semi-Anonymous Social Network

[Hosseinmardi2014corr]

- **Findings**:
  - Interaction network exhibits similar properties to other online social networks and the Web



  - Analyzed 150 profiles expressing users' experience with "cutting" (slicing one's wrists)
  - Among the words connected to "cutting", "depress", "stressful", "sad", and "suicide" are identified as prominent

Analyzing Labeled Cyberbullying Incidents on the Instagram Social Network

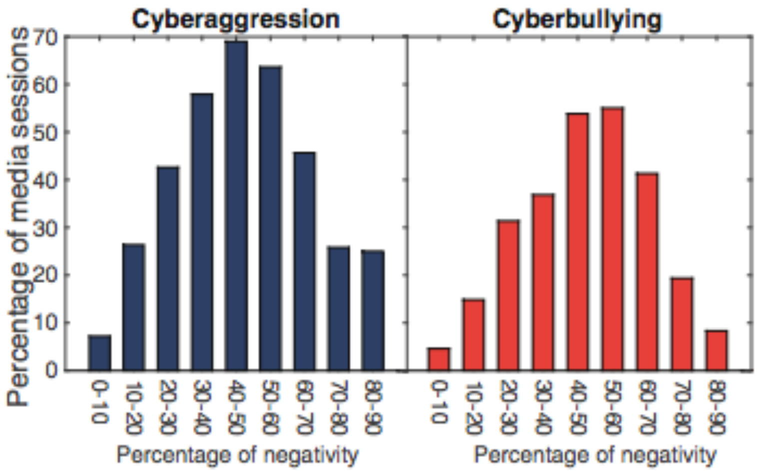
[Hosseinmardi2014corr]

- **Goal**: understand how cyberbullying occurs on Instagram
  - Makes distinction between cyberaggression and cyberbullying

aggressive online behavior

A repetitive act of aggression online with an imbalance of power between the individuals involved

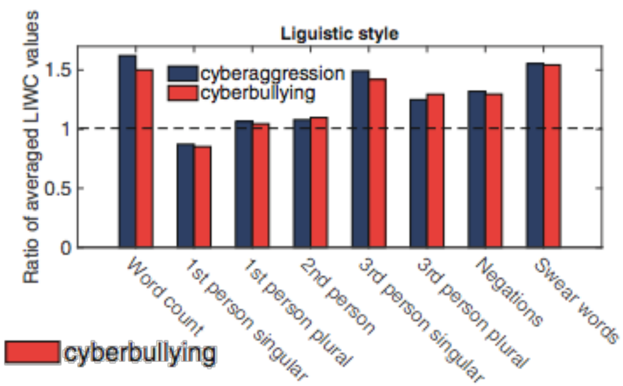
- **Findings**:
  - High agreement between human labelers on which behavior constitutes cyberaggression vs cyberbullying
  - High correlation between cyberbullying/cyberaggression and the percentage of negativity in the comments

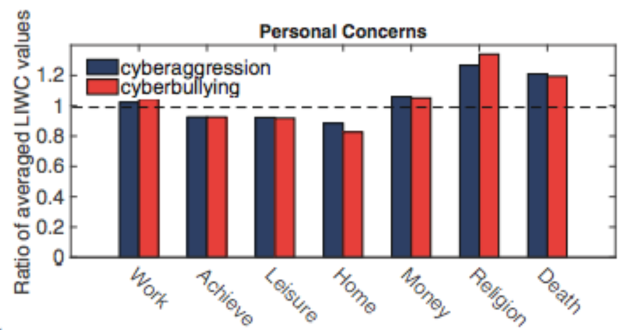


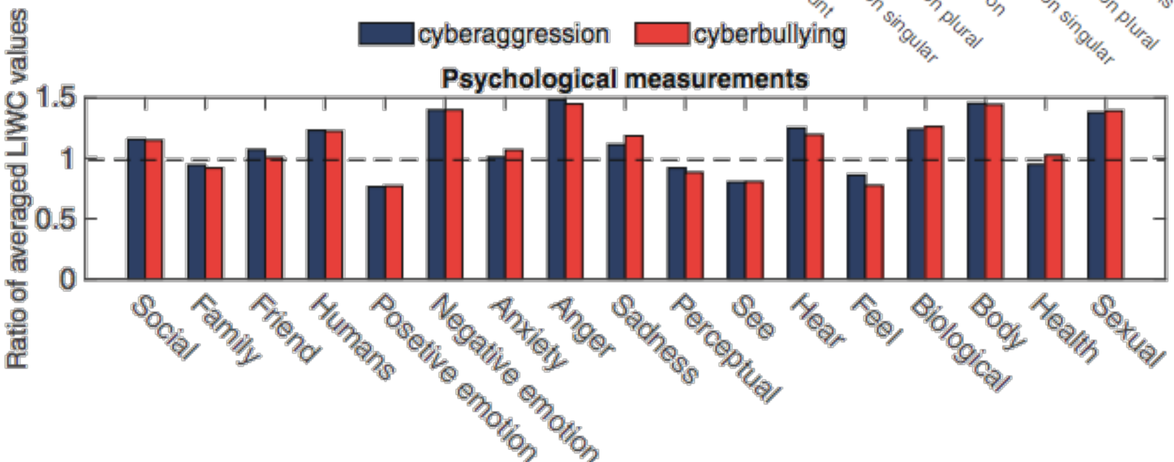
Analyzing Labeled Cyberbullying Incidents on the Instagram Social Network

[Hosseinmardi2014corr]

- **Findings**:
  - Applied Linguistic Inquiry and Word Count (LIWC) to find which categories of words have been used for cyberbullying/cyberaggression labeled media sessions



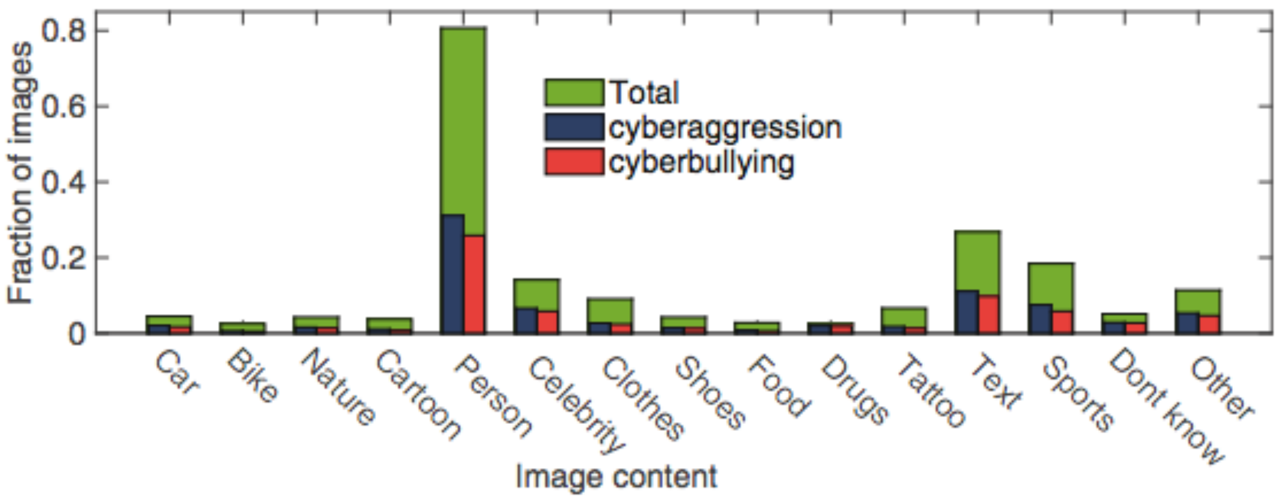




Analyzing Labeled Cyberbullying Incidents on the Instagram Social Network

[Hosseinmardi2014corr]

- **Findings**:
  – Certain image contents (e.g., Drug) are strongly related with cyberbulllying



Prominent Indicators of Cyberbullying

- Four broad categories of features have been used in the literature to study and detect cyberbullying [Al-garadi2016, Salawu2017]
- Mainly derived from user profiles, contents and activity
  - User profile
    - Personality
  - User activity
    - Measure the online communication activity of a user (e.g., number of tweets)
  - Demographics (i.e., gender, age)
  - Content
    - Based on profane and vulgar words/expressions
  - Network
    - Measure the sociability of users online (e.g., number of followers)

Prominent Indicators of Cyberbullying (2)

- Personality [Biel2011, Mishna2012, Liu2016, Edwards2016 , Gosling2017]
  - Hostility significantly predicts cyberbullying
  - Both bullying and cyberbullying have been found to be strongly related to neuroticism (i.e., anxiety, anger, and moodiness)
- Demographics [Edwards2016, Al-garadi2016]
  - Gender and age have been shown to be indicative of cyberbulling in some cases but not in others
  - Nevertheless, most users don't disclose their age and gender in their profiles
- User activity
  - Considerably active users are likely to engage in cyberbullying behavior [Balakrishnan2015]

Race/ ethnicity	Offline bullying[a]	Cyber Bullying	Offline victimi-zation	Cyber victimi-zation
	%	%	%	%
Black	18–46	7–11	7–30	4–17
Hispanic	18–37	16–18	10–17	6–13
Asian	-	-	20-24	15-57[b]
White	11–23	4–42[b]	10–22	18–30

Prominent Indicators of Cyberbullying (3)

- Content
  - Often measured as the number of offensive terms [Dinakar2012, Dadvar2013, Kontostathis2013, Al-garadi2016, Teh2018]
    - Effective in detecting offensive and cursing behavior

  - Popular dictionaries include
    - HateBase: https://www.hatebase.org
    - Noswearing: https://www.noswearing.com/dictionary
    - Offensive/profane word list from https://www.cs.cmu.edu/~biglou/resources/bad-words.txt
    - Slang list: http://www.dailymail.co.uk/news/article-2673678/Why-guide-cyber-bullying-slang-save-childs-life-From-IHML-I-hate-life-Mos-mum-shoulder.html

Words and acronyms used in cyberbullying change [Raisi2017, Raisi2017b]

  - First and second person pronouns
    - A text containing cyberbullying–related features and a second person pronoun is most likely to be meant for harassing others

Prominent Indicators of Cyberbullying (4)

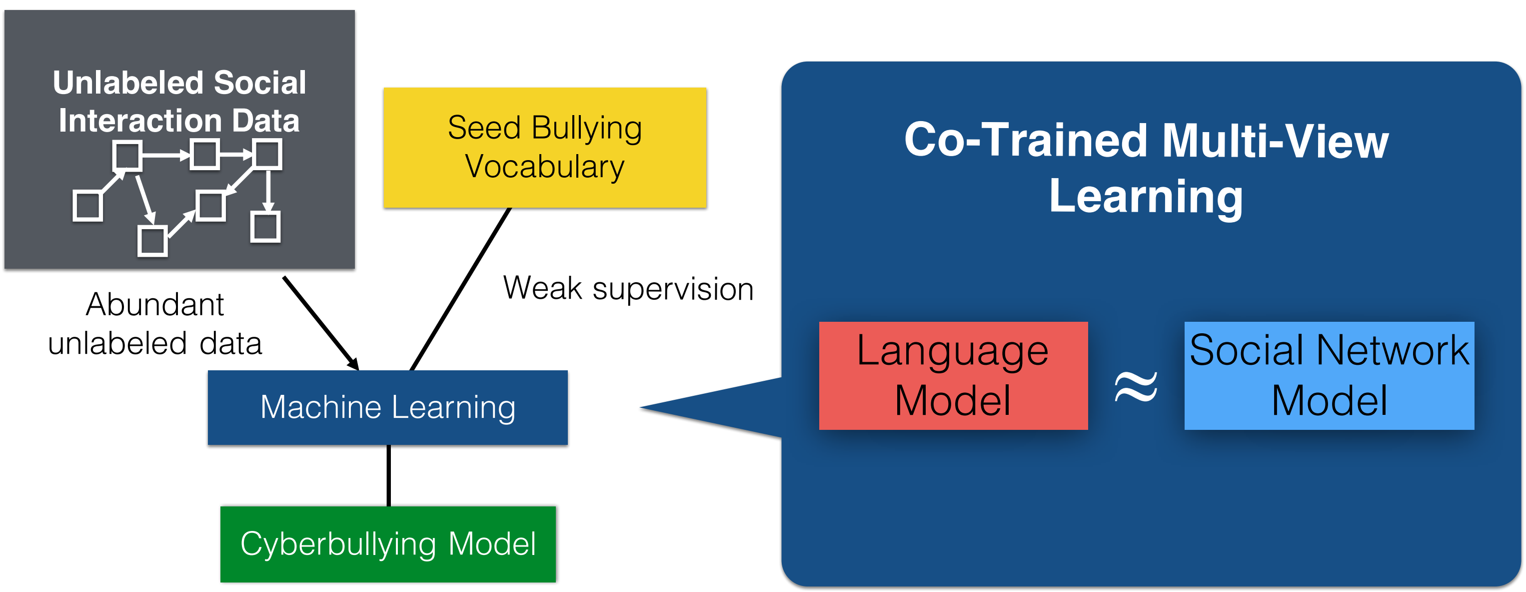
- Visual cues (i.e., features extracted from images and videos) [Zhong2016]
  - Standard image-specific features such as color histogram
  - Features extracted with deep learning
    - **Challenge**: Deep neural networks require a large number of images for training
    - Used a pre-trained neural network & clustered available images
- Photo captions
  - Latent Dirichlet Allocation [Blei2013] to extract latent topics from captions

(a) Cyberbullying (b) Cyberbullying (c) No cyberbullying (d) No cyberbullying

Weakly Supervised Machine Learning

[Raisi2017, Raisi2017b]

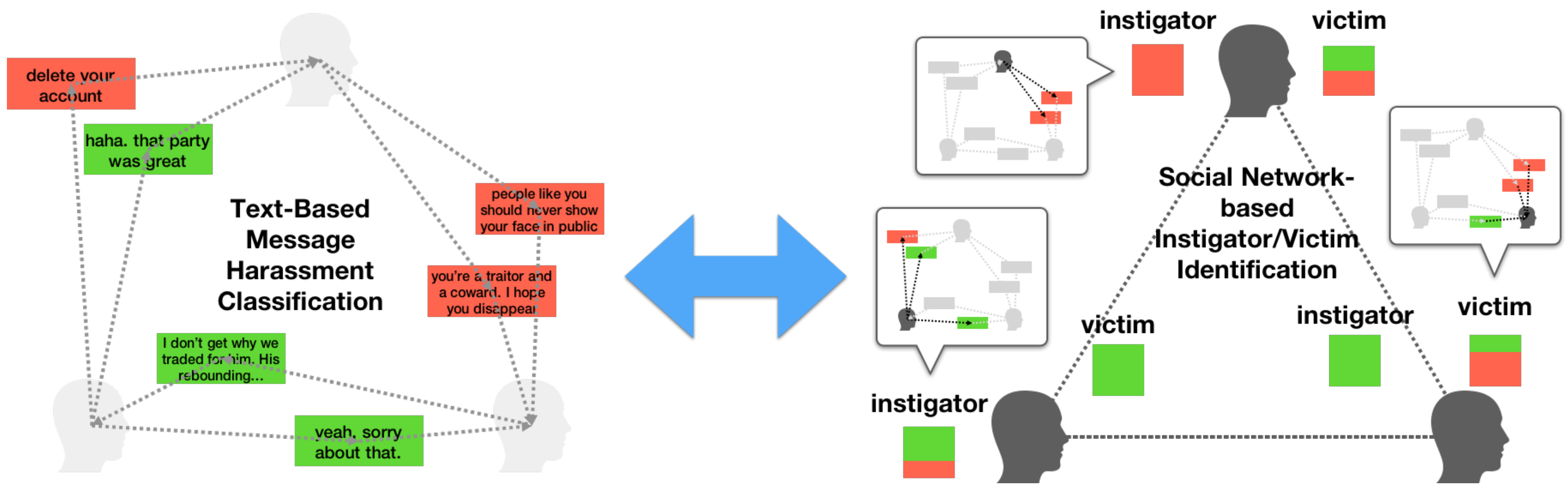
- Methods to characterize (and detect) cyberbullying require labeled data
  - Rely heavily on dictionaries of profane/vulgar words to identify offensive terms in bullying traces
  - Require human annotators to annotators to provide large amounts of labeled examples (tedious, laborious, and often costly, process)
- **Main idea**:



Participant Vocabulary Consistency

[Raisi2017]

- **Goal**: find a consistent parameter setting for all users and key phrases in the data that:
  - Characterizes the tendency of each user to harass or to be harassed, and
  - Characterizes the tendency of a key phrase to be indicative of harassment
  - Parameters are optimized to minimize disagreement with training data
- After convergence, <u>previously unknown terms</u> used by bullies/victims are "learned"



Participant Vocabulary Consistency

[Raisi2017]

- Each user is attributed a bully score and a victim score
  - Bully score encodes how much the model believes a user has a tendency to harass others
  - Likewise, the victim score encodes how much the model believes a user has a tendency to be harassed
- Each n-gram has a harassment–vocabulary score
  - Encodes how much the presence of the feature indicates harassment
- Expert provides seed set of n-grams (i.e., harassment score set to 1.0)

regularizer

for all messages

vocabulary score of word

$$\min_{\mathbf{b},\mathbf{v},\mathbf{w}} \; \frac{\lambda}{2}\left(\|\mathbf{b}\|^2 + \|\mathbf{v}\|^2 + \|\mathbf{w}\|^2\right) + \boxed{\frac{1}{2}\sum_{m \in M}\left(\sum_{k : w_k \in f(m)} \left(b_{s(m)} + v_{r(m)} - w_k\right)^2\right)}$$

$$\boxed{\text{s.t.} \quad \mathbf{w}_k = 1.0 \text{ for } k \in S}$$

bully score of sender

victim score of receiver

expert-provided seed set

for words in message

Participant Vocabulary Consistency

[Raisi2017]

- Once the model is trained, the harassment score of each message can be computed by combining the vocabulary score and the participant score
- The more the model believes user b_s is a bully and v_r is a victim, the more it should believe a given message is an instance of harassment
- For directed pair of users, bullying score sums

bully score of sender

victim score of receiver

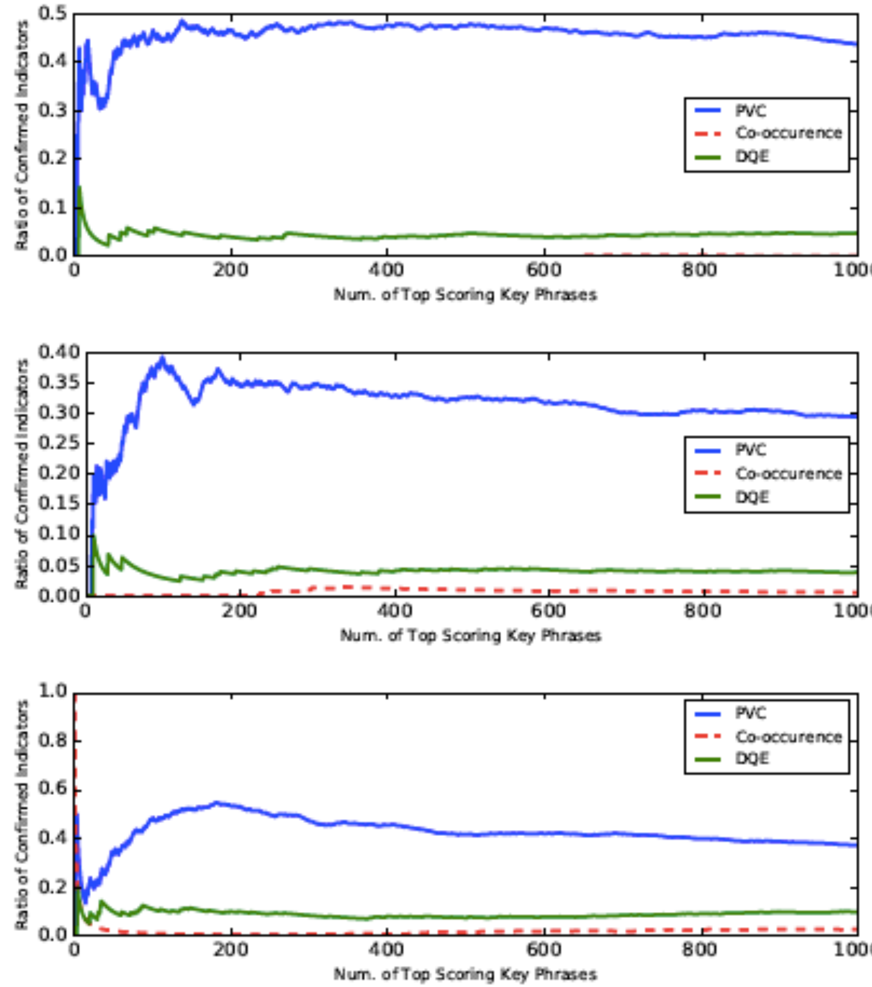
$$\underbrace{\left(b_{s(m)} + v_{r(m)}\right)}_{\text{participant score}} + \underbrace{\frac{1}{|f(m)|} \sum_{k \in f(m)} w_k}_{\text{vocabulary score}}$$

Average word score of n-grams in messages

Participant Vocabulary Consistency

[Raisi2017]

- How good are newly discovered vocabulary terms?
- Human annotators were asked to rate 1,000 highest scoring terms identified by the method (excluding seed words)
- Comparison against
  - Co-occurrence
  - Dynamic query expansion
    - Co-occurrence variation
    - Iteratively grows a query dictionary by co-occurrence and frequency



Key TAKE AWAY: previously unknown indicators of harassment can be identified in a cost–effective way

Prominent Indicators of Cyberbullying (5)

- Social network features
  - A strong correlation between cyberbullying behavior and online sociability has been established [Navarro2012, Hosseinmardi2015, Al-garadi2016, Singh2016, Squicciarini2016, Chatzakou2017, Chelmis2017]
  - Node-level

Metric	Definition	Description
k_u	$\lvert\Gamma(u)\rvert$	Total number of u's neighbors, i.e., degree of u
k_u^+	$\lvert\Gamma^+(u)\rvert$	Total number of outgoing neighbors, i.e. out–degree of node u. In–degree, k_u^-, of a node is defined similarly
$k_u^{(n)}$	$\frac{1}{k_u}\sum_{m\in\Gamma(u)} k_m$	Mean degree over all immediate contacts of node u
ε_u	$\frac{1}{k_u}\sum_{m\in\Gamma(u)} \frac{\lvert\Gamma(u)\cap\Gamma(m)\rvert}{\lvert\Gamma(u)\cup\Gamma(m)\rvert}$	Mean of the ratio between the set of common neighbors and the set of all neighbors of node u and each of its contacts, i.e., embededness of node u [17]
C_u	$\frac{c_u}{k_u/(\lvert V\rvert-1)}$	The ratio between the clustering coefficient of u, $c_u = \frac{2\lvert(u,m)\rvert}{k_u(k_u-1)}$ and the graph density, $k_u/(\lvert V\rvert-1)$ [18]
T_u	$\frac{1}{2}\sum_{m\in V}\sum_{n\in V} 1\{(u,m)\&(u,n)\&(m,n)\}$	Number of triangles containing node u

Prominent Indicators of Cyberbullying (6)

[Chelmis2017]

- Contextual relationship features (i.e., from the combined 1.5 ego–network between sender and receiver)

Metric	Definition	Description
V_{um}	$\lvert V_u^{1.5} \cup V_m^{1.5} \rvert$	Number of nodes in the combined ego–network of u and m
E_{um}	$\lvert E_u^{1.5} \cup E_m^{1.5} \rvert$	Number of edges in the combined ego–network of u and m
ω_{um}	$\lvert V_{um} \rvert (\lvert V_{um} \rvert - 1)$	Maximum number of edges that can be drawn among nodes $m \in \Gamma^+(u)$
CN	$\lvert \Gamma(u) \cap \Gamma(m) \rvert$	Number of nodes linked to both u and m, i.e., common neighbors [19]
JC	$\frac{\lvert \Gamma^+(u) \cap \Gamma^-(m) \rvert}{\lvert \Gamma^+(u) \cup \Gamma^-(m) \rvert}$	Number of common neighbors of u and m divided by the number of neighbors of either node, i.e., Jaccard's coefficient [20]
AA	$\sum_{z \in \Gamma(u) \cap \Gamma(m)} \frac{1}{\log k_z}$	The number of neighbors shared by u and m, divided by the log of the frequency of the neighbors, i.e., Adamic Adar similarity [21]
PA	$k_u^+ \cdot k_m^-$	The product of degrees of the two nodes, i.e., Preferential Attachment [22]
k–core		Obtained by recursively removing all nodes $m \in V_e^{1.5}$ such that $k_m < k$, until all nodes in the remaining graph have at least degree k [8].

Prominent Indicators of Cyberbullying (7)

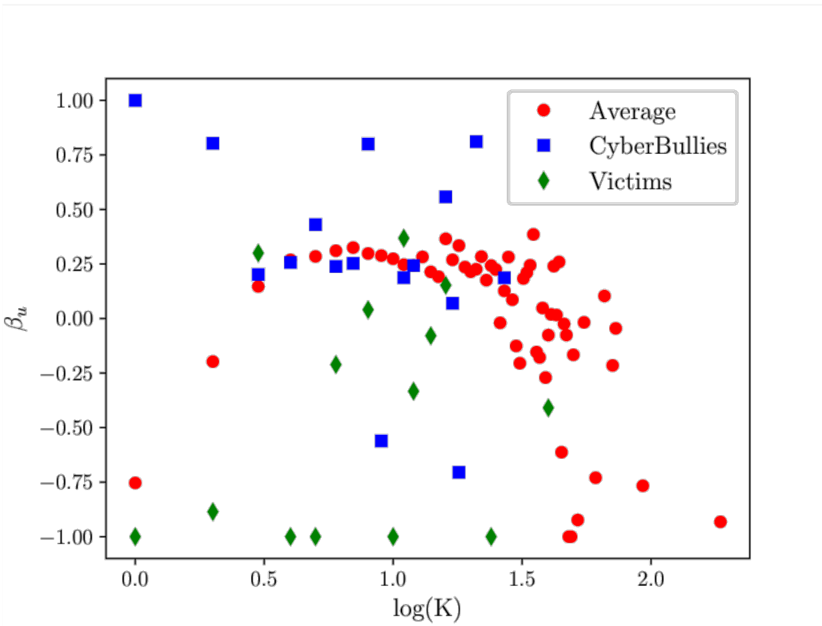
[Chelmis2017]

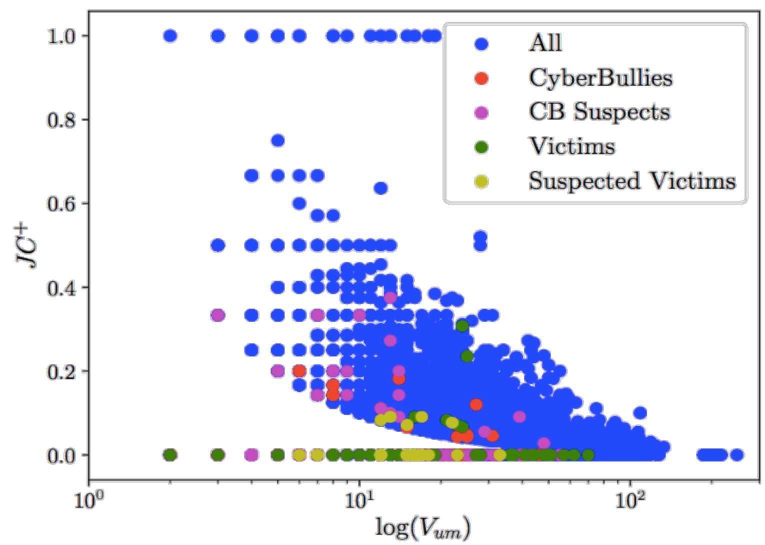
– Activity measures

Metric	Definition	Description
M_u^s	$\sum_{m \in \Gamma^+(u)} w_{(u,m)}$	Total number of messages sent by u
M_u^r	$\sum_{m \in \Gamma^-(u)} w_{(m,u)}$	Total number of messages received by u
β_u	$\frac{M_u^s - M_u^r}{M_u^s + M_u^r}$	Balance ratio of messages sent and received by u, i.e., contribution index [23]
S_{um}	$w_{(u,m)}$	Number of messages u has sent to m, i.e., tie strength
$S_{um}^{(n)}$	$\sqrt{w_{(u,m)} w_{(m,u)}}$	Geometric mean of the number of messages exchanged between u and m
K_{um}	$\frac{k_u^-}{k_m^-}$	Ratio of in–degrees (similarly for out–degrees) of nodes u and m [9]
I_{um}	$\frac{M_u^r}{M_m^r}$	Ratio of incoming (similarly for outgoing) messages that nodes u and m receive regardless of the nodes that such messages are sent from [9]
Θ_{um}	$\frac{k_u^+}{k_u^-} / \frac{k_m^+}{k_m^-}$	Out–degree to in–degree ratio of nodes u and m [9]
Δ_{um}	$\frac{M_u^r}{k_u^+}$	Incoming messages to in–degree ratio (similarly for M_u^s and k_u^+) [9]
χ_u	$\sum_{m \in \Gamma^+(u)} \left(p_{um} + \sum_{q \in \Gamma^+(u), q \neq m} p_{uq} p_{qm} \right)^2$	We define "attention spanning" of node u as given in [7], where $p_{ij} = \frac{w_{(i,j)}}{\sum_i w_{(i,j)}}$ denotes the amount of direct attention that node i gives to j, and the inner sum represents the total amount of indirect attention that u gives to m through some intermediary q
$B(e)$	$\sum_{m \in V} \sum_{w \in V \setminus \{m\}} \frac{\sigma_{m,w}(e)}{\sigma_{m,w}}$	The betweeness centrality of an edge e [5], where $\sigma_{m,w}$ denotes the number of shortest paths between nodes m and w, and $\sigma_{m,w}(e)$ indicates the number of shortest paths between nodes m and w through edge e.

Prominent Indicators of Cyberbullying (8)

[Chelmis2017]





Prominent Indicators of Cyberbullying (9)

[Al-garadi2016]

- Feature selection
  - Often used to determine significant features
  - For a review, please refer to [Yang1997, Guyon2003, Peng2005, Saeys2007]
- Top ten significant features
  - By chi-square test [Greenwood1996], information gain, and Pearson correlation [Yang1997]

χ^2 test (chi-square test)	Information gain	Pearson correlation
Vulgarities feature (number of vulgar words in the post).	Vulgarities feature (number of vulgar words in the post).	Vulgarities feature (number of vulgar words in the post).
100 most commonly used words in social media that are positively correlated with neuroticism	100 most commonly used words in social media that are positively correlated with neuroticism	100 most commonly used words in social media that are positively correlated with neuroticism
100 most commonly used words in social media that are used by males	100 most commonly used words in social media that are used by males	100 most commonly used words in social media that are used by males
Average number of followers to following	100 most commonly used words in social media that negatively correlate with age (30 years and above)	100 most commonly used words in social media that positively correlate with age (19–22 years)
100 most commonly used words in social media that positively correlate with age (19–22 years)	100 most commonly used words in social media that positively correlate with age (19–22 years)	100 most commonly used words in social media that negatively correlate with age (30 years and above)
100 most commonly used words in social media that negatively correlate with age (30 years and above)	Number of tweets	Number of tweets
Number of friends following a user	Average number of followers to following	Number of mentions
Number of tweets	Second person pronouns	Second person pronouns
Second person pronouns	Number of friends following a user	Average number of followers to following
Number of mentions	Number of mentions	Slang feature (number of slang words in the post)

Temporal Dynamics of Cyberbullying

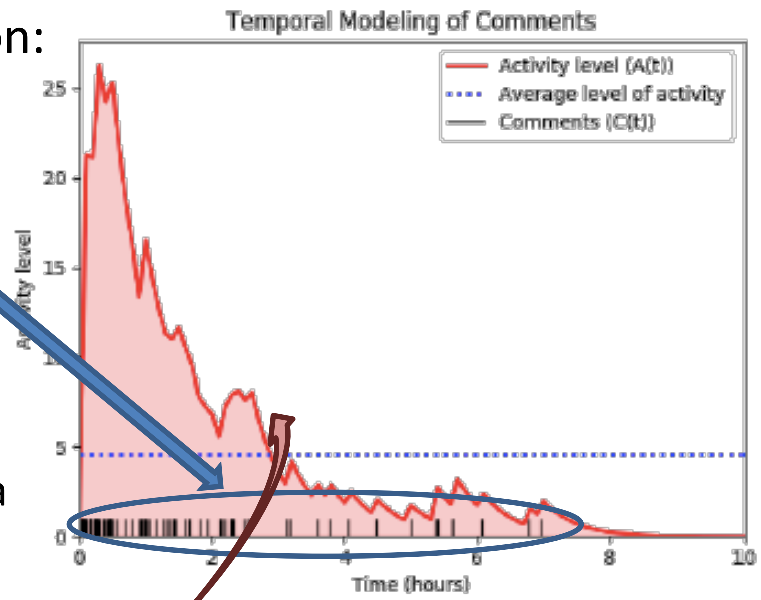
[Soni2018]

- Very little computational work has focused on the **temporal dynamics** and the **repetition** of bullying behavior over time
- **Goals**:
  - Model the temporal aspects of commenting behavior in Instagram media sessions to reveal unique characteristics of cyberbullying (as opposed to regular media sessions)
  - Study the benefit (if any) of augmenting textual features with temporal features to increase cyberbullying detection performance
- **Dataset**: 1,734 Instagram media sessions [Hosseinmardi2015] with labeling confidence of ≥ 0.8
  - 365 media sessions labeled as cyberbullying

Temporal Dynamics of Cyberbullying

[Soni2018]

- Each media session has an initial (logical) submission time (i.e., $t_0 = 0$)
- Each comment i has an associated posting time $t_i \geq t_0$ modeled as a Dirac delta function:
  - Common technique used to model times of interest (e.g., [Hołyst2000, Harabagiu2011, Bourigault2014, Tsytsarau2014, Farajtabar2015])
- Time difference between each chronological pair of comments is measured
  - Comments are assumed to be generated by a homogeneous Poisson point process
- Each comment boosts the activity level of a media session by an exponentially-decaying amount



Temporal Dynamics of Cyberbullying

[Soni2018]

- Duration of a media session (i.e., time difference between submission time and last comment)
- Time to first comment
- Inter-comment interval mean, variance, and coefficient of variation (cv)
  - CV is used to measure how ”Poisson–like” comments are
    - If they were truly generated from a Poisson process, this would equal 1
- Number of bursts
  - Bursts of comments may reflect cyberbullying/abuse in which several people gang up on a victim
  - Measured as the Poisson surprise
- Amount of total activity (measured as the integral of $A(t)$)
- Average level of activity
- Number of mean crosses

Temporal Dynamics of Cyberbullying

[Soni2018]

- Several features found to have statistically significant differences ($p < 0.001$ by t-test) between bullying and non-bullying media sessions

Feature	Difference
Time to first	86.7%
ICI mean	-42.1%
ICI variance	-42.1%
ICI coefficient of variation	-21.0%
Number of bursts	10.8%
Amount of total activity	52.8%
Average level of activity	52.0%

- Notes:
  - Cyberbullying sessions tend to receive a less immediate response
  - Cyberbullying sessions receive a more steady stream of comments that are closer together
  - Cyberbullying sessions tend to exhibit higher level of activity throughout
  - Cyberbullying sessions are more likely to contain bursts in comments

Characterizing and Detecting Hateful Users on Twitter

[Ribeiro2018]

- Methodology to collect and annotate hateful users without depending directly on lexicon
- Users are annotated as hateful or normal based on their entire profile
- Data collection
  - A sample of the Twitter retweet graph is obtained
  - A belief score is assigned to each user based on a lexicon
  - A diffusion process is used to sample users
  - Users are divided into 4 classes according to their associated beliefs after diffusion, and a stratified sampling is performed
- Some findings:
  - Hateful users differ from normal in terms of their activity patterns, word usage and network structure
  - Hateful users are densely connected, tweet more, in shorter intervals, favorite more tweets by other people and follow other users more

Hate Speech in Social Media

- ElSherief, Mai, Shirin Nilizadeh, Dana Nguyen, Giovanni Vigna, and Elizabeth Belding. "Peer to Peer Hate: Hate Speech Instigators and Their Targets." *ICWSM2018*
  - Comparative study that reveals key differences between hate instigators, targets and general Twitter users in terms of profile self-presentation, Twitter visibility, and personality traits
  - Twitter hate speech dataset available at https://github.com/mayelsherif/hate_speech_icwsm18
- ElSherief, Mai, Vivek Kulkarni, Dana Nguyen, William Yang Wang, and Elizabeth Belding. "Hate Lingo: A Target-based Linguistic Analysis of Hate Speech in Social Media." *ICWSM2018*
  - Studies the lexical, semantics, and psycholinguistic patterns of directed and generalized hate and reveal key differences in the linguistic styles of the two types of hate

Section

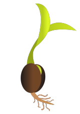
Cyberbullying Detection (& Prediction)

Cyberbullying Detection Methods

[Nadali2010, Salawu2017]

- Supervised learning
  - Typically use naïve classifiers such as SVM and Naïve Bayes

- Weakly-supervised learning
  - Learn previously unknown n-grams from a small seed-vocabulary



- Lexicon based
  - Rely on the presence of words from predetermined dictionaries

- Rule based
  - e.g., match text/user's age/mobile phone usage pattern to predefined rules

- Mixed–initiative
  - Combine human–based reasoning with one or more of the aforementioned approaches

Cyberbullying Detection Methods (2)

- **Detection** methods

  - **Offline** [the majority of methods in the literature: Al-garadi2016, Salawu2017]
    - Emphasis on improving the accuracy of cyberbullying detection classifiers

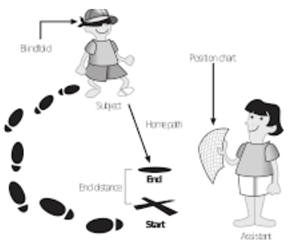
  - **Online** [Rafiq2018, Yao2018, Zois2018]
    - Examining comments as they become available
    - <u>One of the most challenging objectives</u>
    - **Goal** is to reduce the classification time and time to raise alert
- **Apriori prediction** methods [Potha2014, Hosseinmardi2016, Zhong2016, Liu2018]
  - Utilize initial content (e.g., image), metadata (i.e., caption), & user info (i.e., profile and past activity) to predict cyberbullying before it happens
  - <u>One of the most challenging objectives</u>
  - **Goal**:
    - Identification and warning of vulnerable users
    - Targeted (and thus efficient and scalable) detection in large online social networks

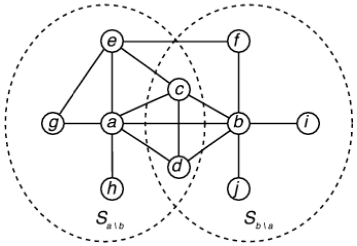
Cyberbullying Detection Methods (3)

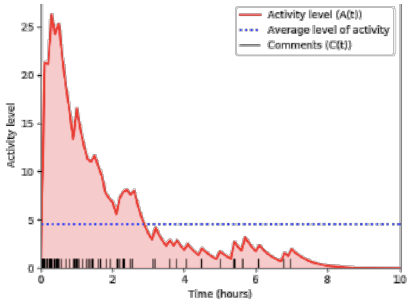
- **Content** and metadata about the content itself (e.g., frequency of profanity)
  - Profane words are overwhelmingly used in the literature
  - Not all cyber aggression constitutes bullying
  - Sentiment & emotion analysis are rarely sufficient on their own to accurately identify bullying
  - The use of content features alone fails to consider other key aspects of cyberbullying such as repetitiveness and power differential
- **Profile** (e.g., # of followers) and demographic information (e.g., age)
  - e.g., age, gender, race, and culture
  - Have been shown to improve performance, however, such user-provided information can be easily falsified
  - A forensic linguistic module could be used (e.g., to assign a “truth score” to age and gender information supplied by a user)

Cyberbullying Detection Methods (4)

- **Visual** cues
  - i.e., features extracted from image and video content
- **Network structure**
  - e.g., features extracted from followership/communication networks
  - Increasingly being used for detection
- **Temporal** (i.e., changing with time) vs. **static**
  - e.g., elapsed time between comments made by two different users to measure the influence of cyberbullies on their peers and map the spread of bullying across a social network
- **Combination** of features leading to multimodal methods



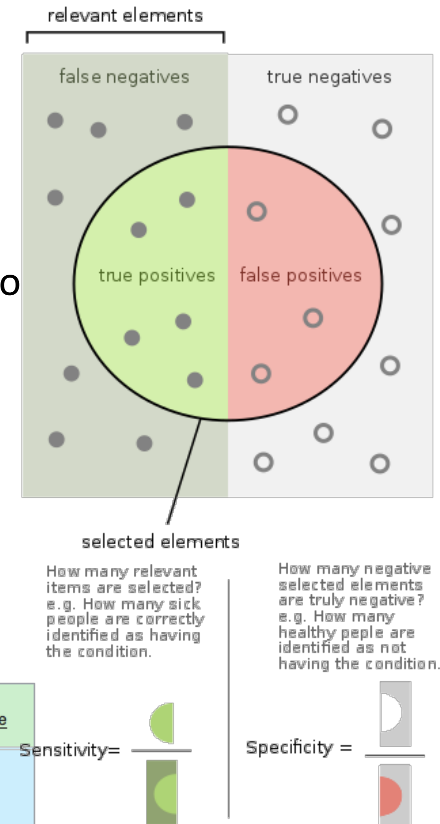




Performance Evaluation & Comparison

- Often used evaluation metrics [Davis2006, Powers2011]
  - Accuracy

    - Inappropriate when dealing with high class imbalance datasets
    - The accuracy of a classifier that labels everything as the "majority" class will be 95% in a dataset with 95% imbalance ratio
  - Precision / Recall / F-score
    - Sensitive to performance for only one class
    - In highly skewed datasets, the recall of the minority class is ~0
    - Better to use average F-score across classes
  - Confusion matrix:

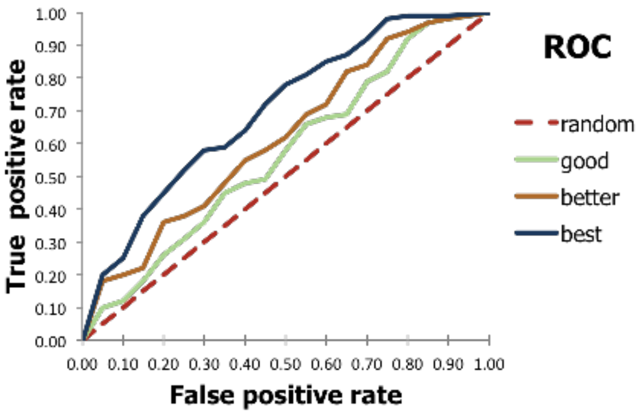


		True condition				
	Total population	Condition positive	Condition negative	Prevalence = $\frac{\Sigma \text{ Condition positive}}{\Sigma \text{ Total population}}$	Accuracy (ACC) = $\frac{\Sigma \text{ True positive} + \Sigma \text{ True negative}}{\Sigma \text{ Total population}}$	
Predicted condition	Predicted condition positive	**True positive**, Power	**False positive**, Type I error	Positive predictive value (PPV), Precision = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Predicted condition positive}}$	False discovery rate (FDR) = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Predicted condition positive}}$	
	Predicted condition negative	**False negative**, Type II error	**True negative**	False omission rate (FOR) = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Predicted condition negative}}$	Negative predictive value (NPV) = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Predicted condition negative}}$	
		True positive rate (TPR), Recall, Sensitivity, probability of detection = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$	False positive rate (FPR), Fall-out, probability of false alarm = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$	Positive likelihood ratio (LR+) = $\frac{\text{TPR}}{\text{FPR}}$	Diagnostic odds ratio (DOR) = $\frac{\text{LR+}}{\text{LR−}}$	F_1 score = $\frac{2}{\frac{1}{\text{Recall}} + \frac{1}{\text{Precision}}}$
		False negative rate (FNR), Miss rate = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$	True negative rate (TNR), Specificity (SPC) = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$	Negative likelihood ratio (LR−) = $\frac{\text{FNR}}{\text{TNR}}$		

Source: https://en.wikipedia.org/wiki/Sensitivity_and_specificity

Performance Evaluation & Comparison (2)

– The weighted area under the ROC curve (i.e., AUC)
  - Created by plotting sensitivity against the probability of false alarm at various threshold settings
  - More robust than Accuracy, Precision, Recall, and F-measure in datasets with high class imbalance [Fawcett2006]
  - High AUC indicates improved classification for both classes regardless of class imbalance [Fawcett2006]



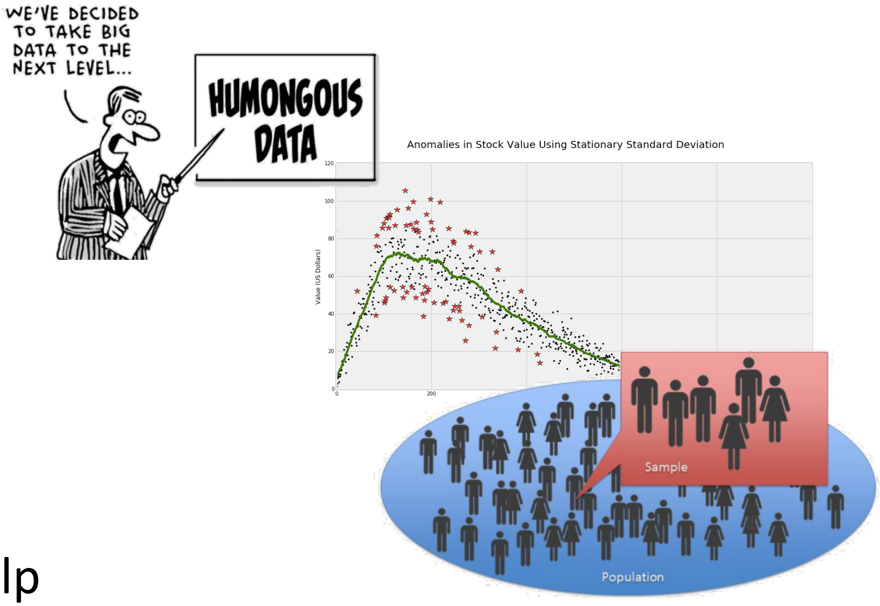
– Matthews Correlation Coefficient (MCC)
  - Less sensitive to data skewness as it considers mutual accuracies of both classes and all four values of the confusion matrix

– G–means: measures the avoidance of overfitting the negative class

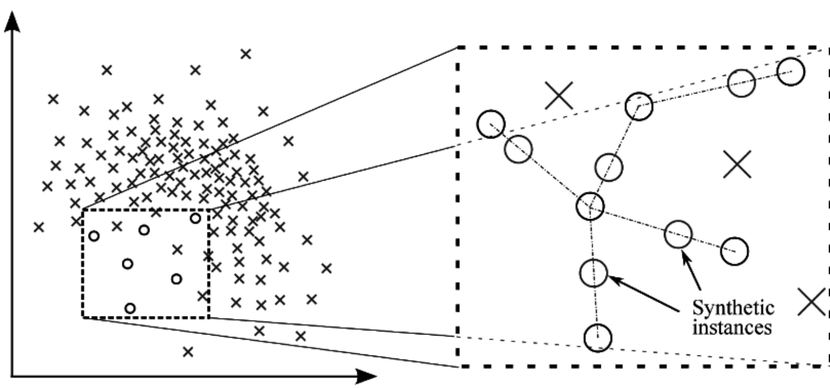
– β – varied F–measure
  - Better captures the trade–off between Precision and Recall

Handling Imbalanced Datasets

- Many ways to handle class imbalance
  - Collect more data
    - May be impossible or costly
  - Try anomaly detection techniques
    - Assumes “abnormal” signal in the data
  - Use over/under sampling techniques
    - Undersampling can lead to loss of important information
  - …

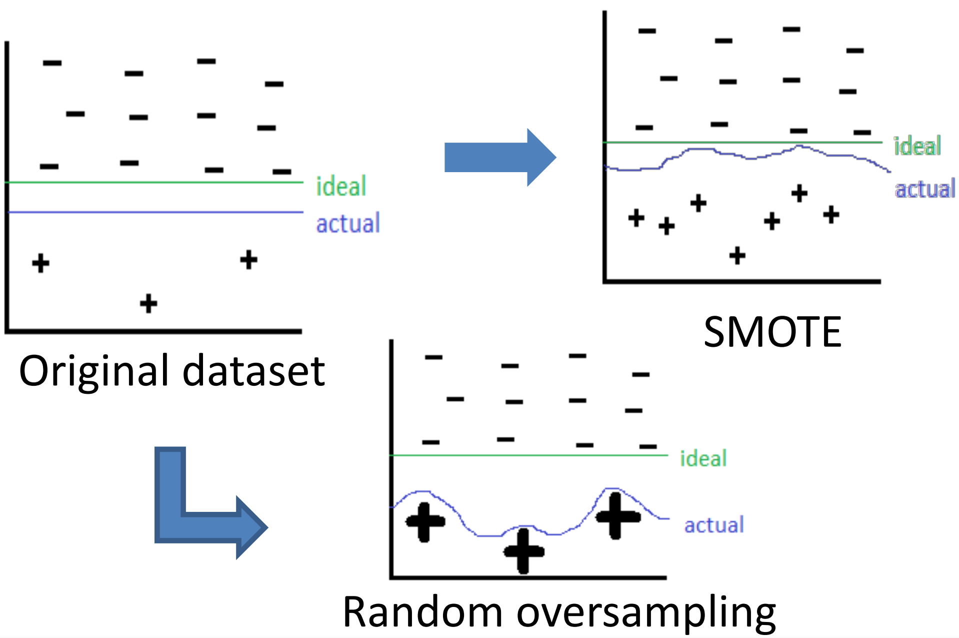


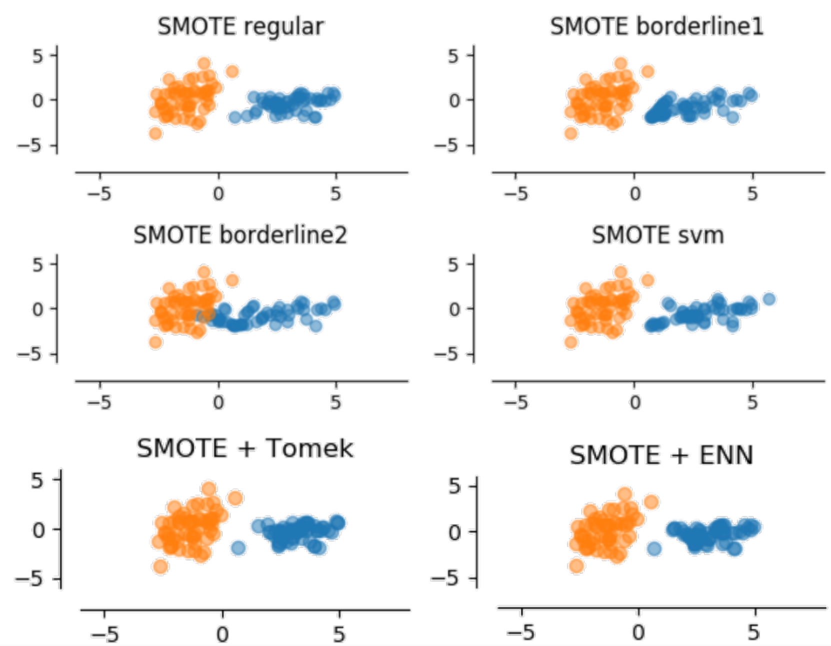
- Oversampling the minority class may help
  - Synthetic Minority Over sampling Technique (SMOTE) [Chawla2002] creates synthetic samples of the minority class around K neighbors of minority samples



Oversampling the Minority Class

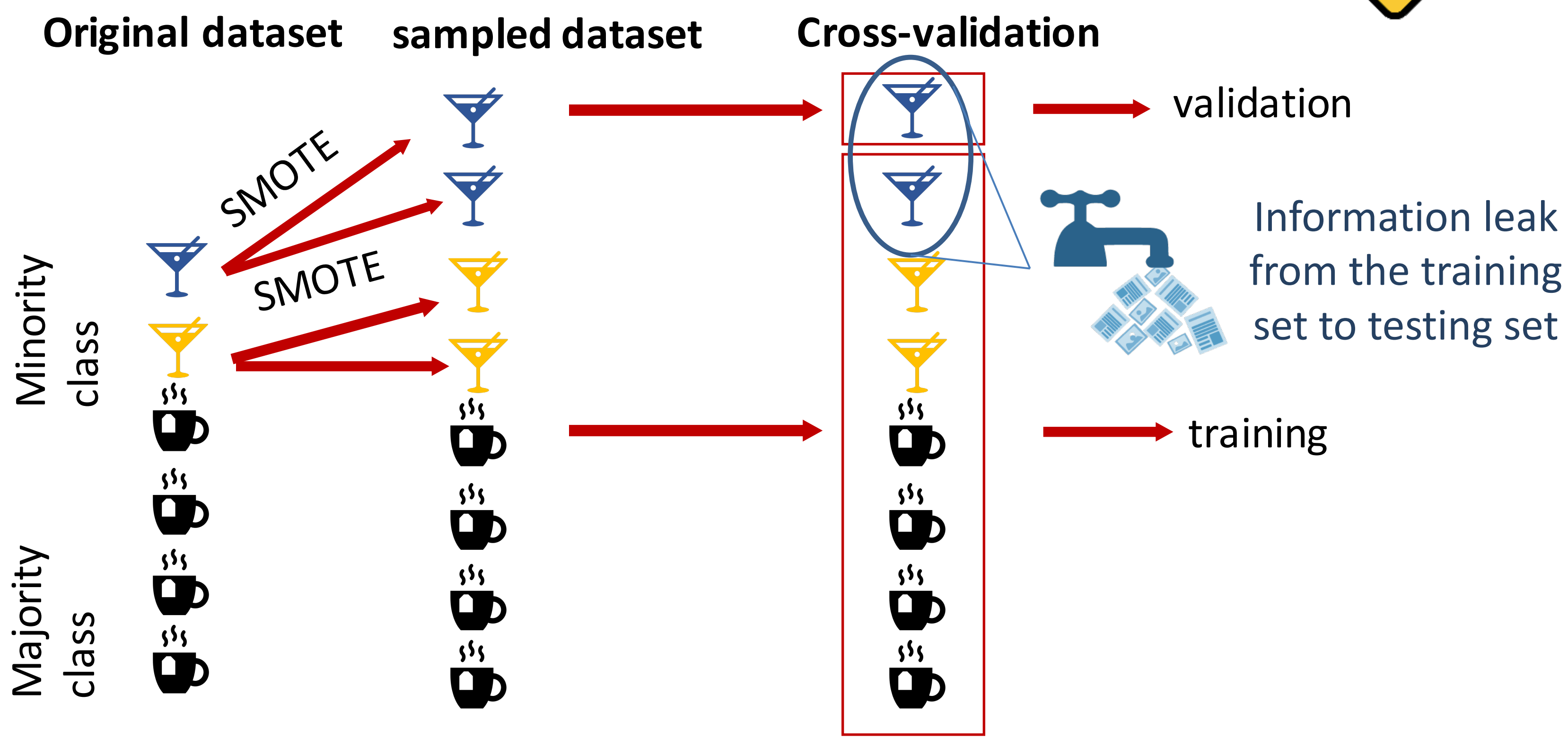
- When duplicating data points (e.g. Random over-sampling), classifiers get “convinced” about data points with small boundaries around it
- SMOTE forces the decision region of the minority class to become more general, partially solving the generalization problem
- Variations of SMOTE (e.g., [Han2005, Bunkhumpornpat2009]) and combinations with cleaning methods [Batista2004]



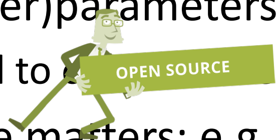


Oversampling the Minority Class

- SMOTE must be applied with care
- Information may leak if oversampling is performed before splitting a dataset into training and testing sets



Performance Evaluation & Comparison (3)

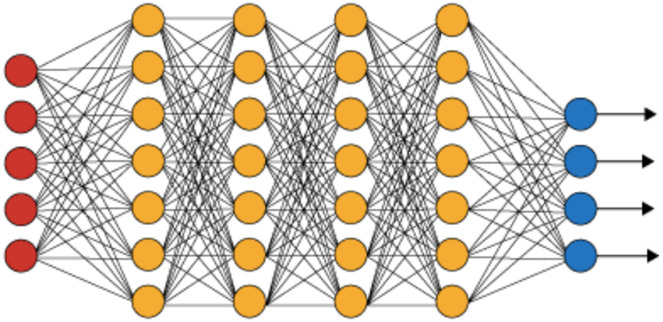
- Direct comparison of state-of-the-art methods is difficult
  - For fair and meaningful comparison, experiments must be conducted on the same exact dataset (c.f. Data Challenges)
  - The (hyper)parameters (if any) of each method must be replicable
    - Need to 

 code for reproducibility (c.f. Giving Back)
  - Objective matters: e.g., binary classification and role identification can result in different accuracy even if performed on the same dataset
- Some of the highest scores reported are on blogs and forum datasets [Salawu2017]

Cyberbullying Detection Based on Semantic-Enhanced Marginalized Denoising Auto-Encoder

[Zhao2017]

- **Goal**: develop a method to learn robust and discriminative numerical representations of text for cyberbullying detection
  - Postulates that textual features are most reliable
  - Automatic extraction of bullying words based on learned word embeddings
- **Challenges**:
  - Messages on online social media are very short
  - Informal language use & misspellings are often
  - Data sparsity (i.e., lack of sufficient high-quality training data)



Cyberbullying Detection Based on Semantic-Enhanced Marginalized Denoising Auto-Encoder

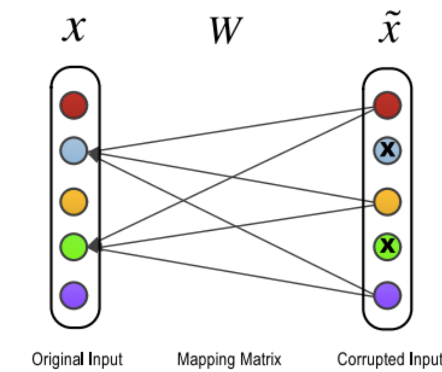
[Zhao2017]

- **Intuition**:
  - Bullying messages may not contain “bullying” words
- **Key idea:**
  - Learn bullying features from normal words by discovering latent structure
  - Enable detection of bullying messages without bullying words
- **Approach**:
  - Deep learning method
  - Each comment is represented using a BoW vector $\boldsymbol{x}$
  - The dataset can be denoted by matrix $\boldsymbol{X} = [\boldsymbol{x}_1, \dots, \boldsymbol{x}_n]$

<u>Stacked structure</u>

The output of the $(k-1)$th layer is fed as input into the kth layer

Denoising Auto-Encoder

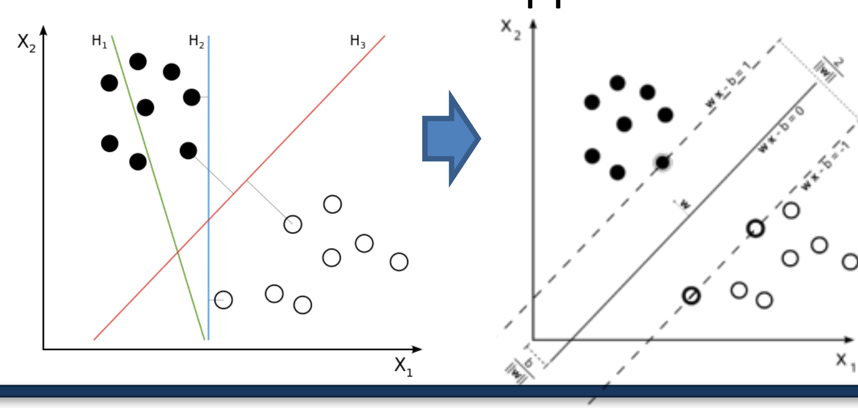


$$\mathbf{W} = \underset{\mathbf{W}}{\operatorname{argmin}} \frac{1}{2n} \sum_{i=1}^{n} \|\mathbf{x}_i - \mathbf{W}\tilde{\mathbf{x}}_i\|^2$$

Cyberbullying Detection Based on Semantic-Enhanced Marginalized Denoising Auto-Encoder

[Zhao2017]

- Bullying words should be chosen properly for the first layer
  - A list of “negative” words (e.g., profane words) must be provided
  - Expand the list of pre-defined words based on word2vec model
    - Pre–trained on a large–scale twitter corpus of 400 million tweets (available at: https://www.fredericgodin.com/software/)
    - For each seed word, “similar” words were extracted using cosine similarity
- Feature selection is performed for subsequent layers
  - Fisher score to select top k discriminative features
- Learned numerical representations are fed into a Support Vector Machine for binary classification

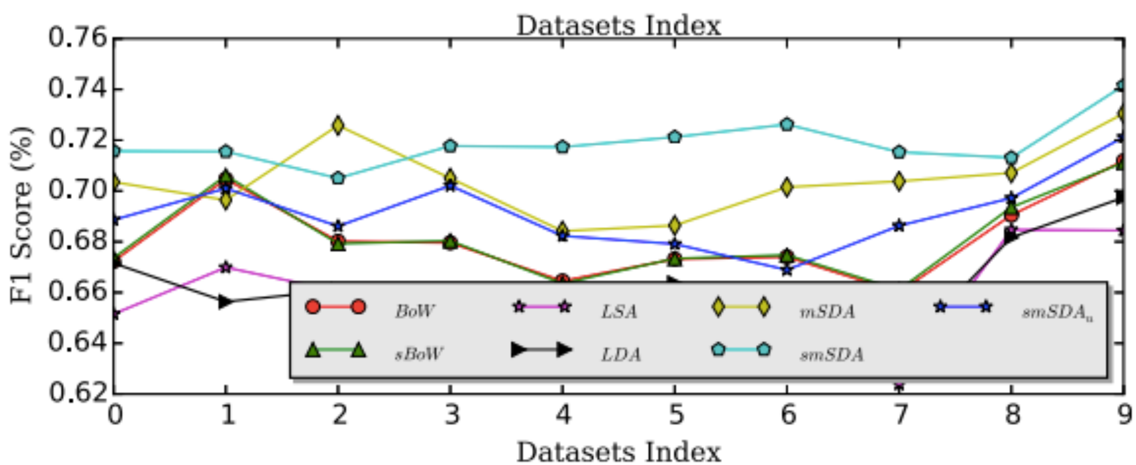


Cyberbullying Detection Based on Semantic-Enhanced Marginalized Denoising Auto-Encoder

[Zhao2017]

- Datasets used for evaluation
  - 7,321 randomly sampled & manually labeled tweets [Xu2012]
  - MySpace (c.f. pointer in the Datasets section of the tutorial)

Bullying Words	Reconstructed Words for	
	mSDA	smSDA
bitch	*@USER shut friend tell*	*@USER HTTPLINK fuck up shut*
fucking	*because friend off gets*	*off pissed shit of*
shit	*some big with lol*	*abuse this shit shit lol big*

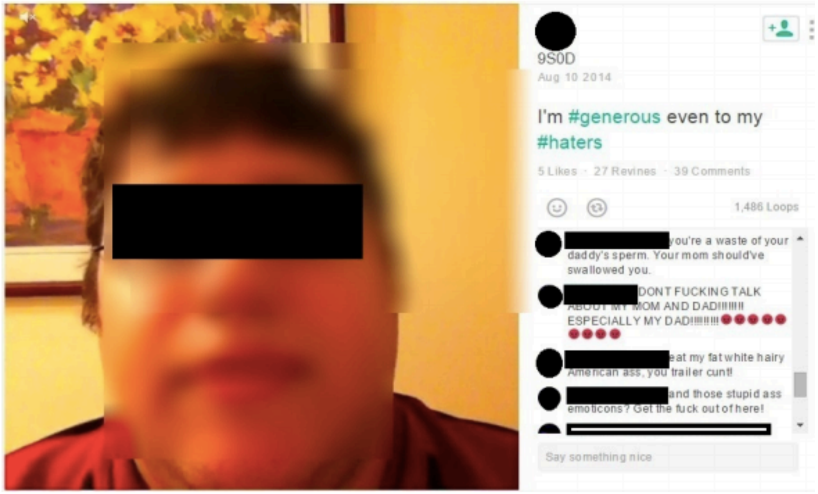


- Observations:
  - Deep learning method outperforms the baselines
  - Correlations between seed words and "normal" words seem to be intuitive

Scalable and Timely Detection of Cyberbullying in Online Social Networks

[Rafiq2018]

- **Goal**: develop a system for **scalable** and **timely** cyberbullying detection
  - **Scalable**: Accommodate the enormous amount of data shared daily on online social media platforms
  - **Responsive**: Be able to monitor a large number of media sessions yet quickly raise an alert (i.e., online approach)

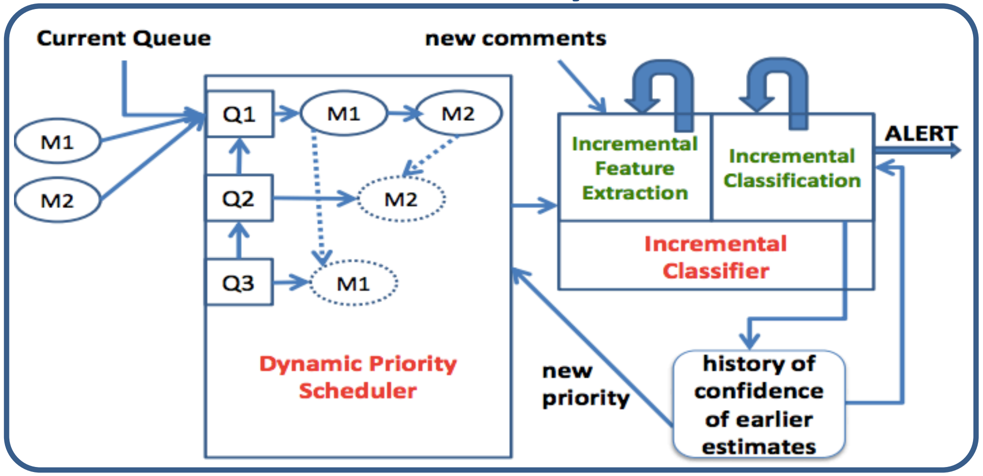


- **Approach:**
  - Multi-stage detection system
  - Incremental feature extraction and classification
  - Reuses previous classification results to reduce overhead with minimal impact on accuracy

Scalable and Timely Detection of Cyberbullying in Online Social Networks

[Rafiq2018]

System Architecture



- Incremental logistic regression classifier
  - Use incrementally linear features
  - Values are computed for first n comments
  - When δn new comments arrive, only the individual feature vector values for the new comments have to be computed
  - Reuse the values for the first n comments to compute the overall feature vector for the $n + \delta n$ comments

- Given features $a_i, i = 0, \ldots n$, LR assigns them weights w_i to compute value $c = \sum_0^n a_i w_i$
- Value c is fed into a sigmoid function to output a value from 0 to 1

Scalable and Timely Detection of Cyberbullying in Online Social Networks

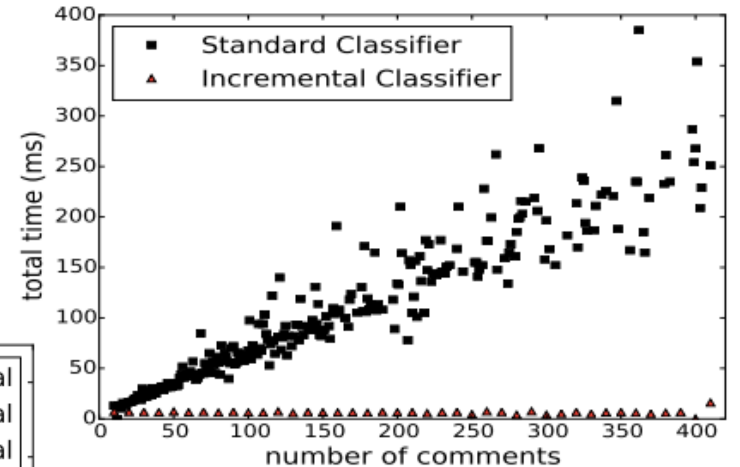
[Rafiq2018]

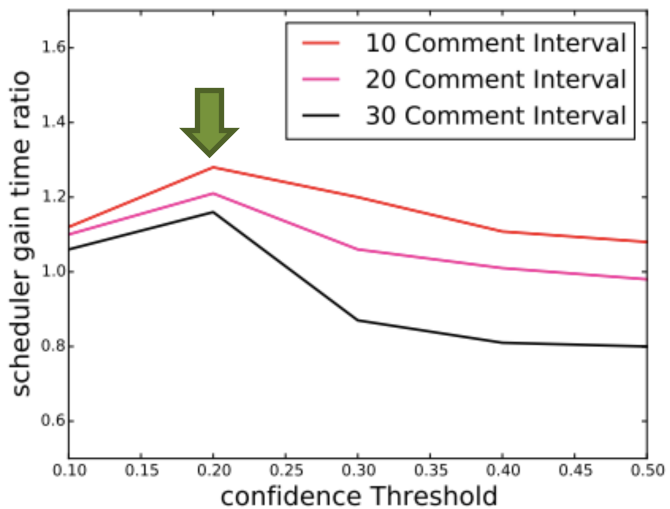
- Observations:
  - Not all media sessions need to be monitored equally
    - Can **prioritize** among media sessions
  - A media session can slowly evolve into a cyberbullying instance (even if it started as a non-bullying session) with the arrival of comments over time
    - Need to <u>eventually examine</u> **all** media sessions (including the low priority)
- Dynamic priority scheduler
  - Two priority levels (high and low)
  - Newly created media sessions are marked high priority
  - Each media session's priority dynamically varies
  - Set priority to high if average of **<u>all</u>** past confidence values (value c) for past classifications is ≥ 0.2
    - Average is used to account for "*repeated aggressive behavior*"

Scalable and Timely Detection of Cyberbullying in Online Social Networks

[Rafiq2018]

- Evaluation
  - 10-fold cross validation on labeled Vine data
  - Incremental Classifier vs AdaBoost
    - Adaboost achieves slightly higher precision
    - LR achieves higher recall and F-1 score
    - LR is 5X faster than Adaboost
  - Dynamic Priority Scheduler threshold value and batch size

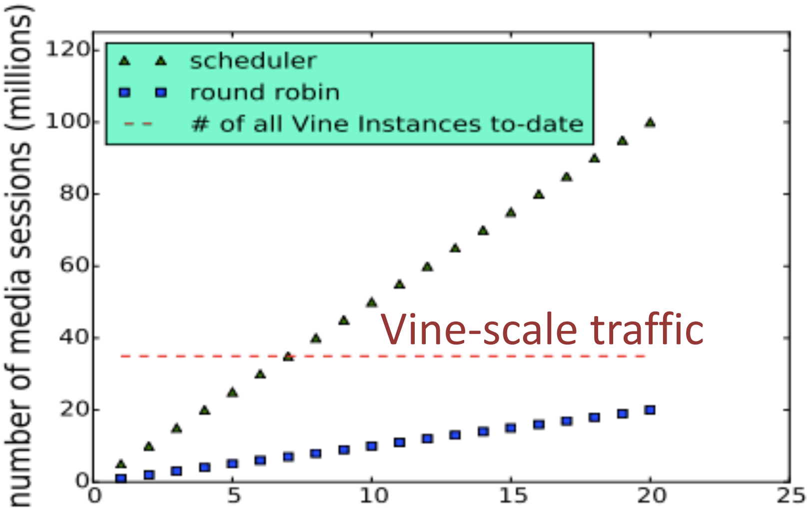




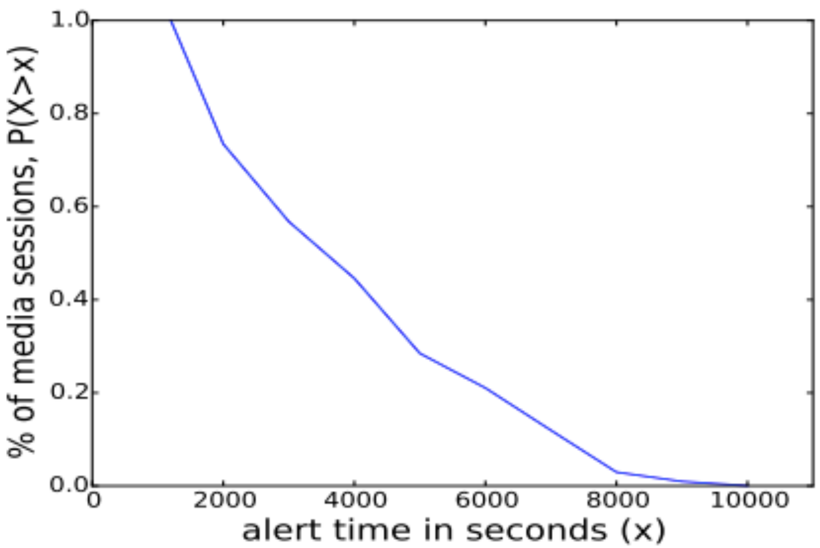
Scalable and Timely Detection of Cyberbullying in Online Social Networks

[Rafiq2018]

- Scalability analysis



Number of Amazon instances

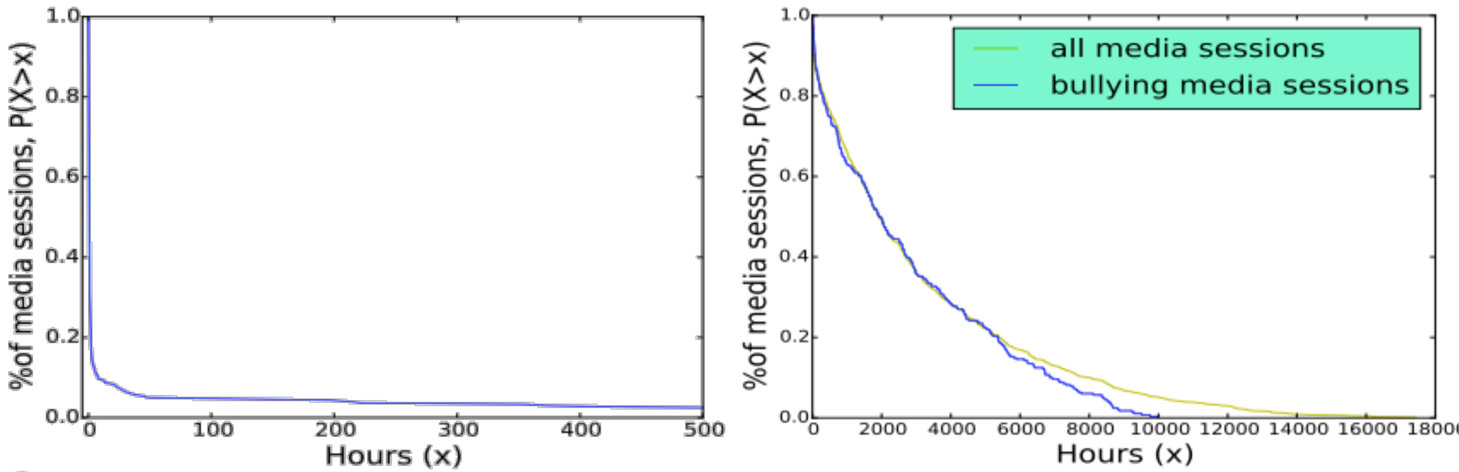


CCDF of alert time for 5 million media sessions in 1GB memory amazon instance

Scalable and Timely Detection of Cyberbullying in Online Social Networks

[Rafiq2018]

- Activity analysis observations
  - Very few bullying media sessions receive their first comment after 500 hours
  - Bullying media sessions receive all their comments within a year of their creation



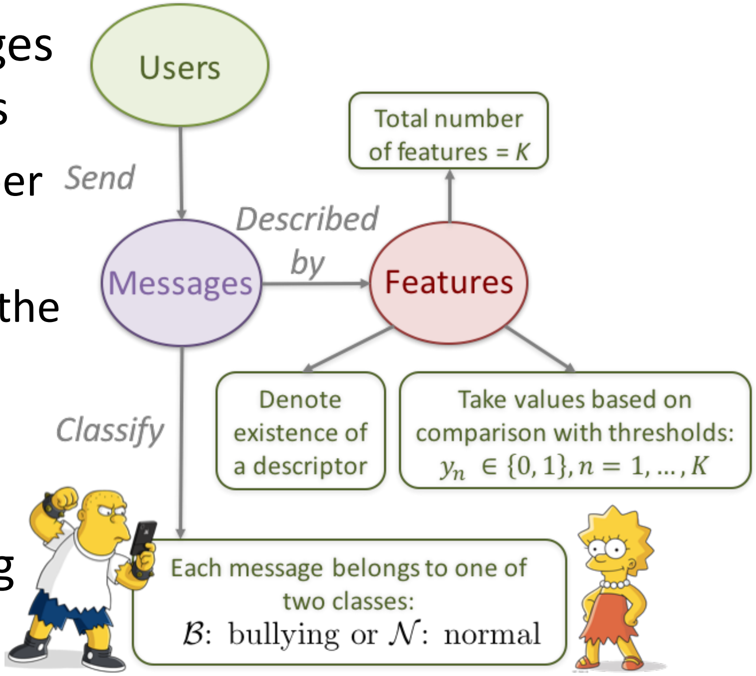
-  Recommendations to improve performance and use of resources
  - Stop monitoring sessions that need >500 hours to get their first comment
  - Purge out all media sessions that are one year old

Optimal Online Cyberbullying Detection

[Yao2018, Zois2018]

- <u>**Goal**</u>: Accurately detect cyberbullying messages using text (& some network) – based features
  - Solution should be **scalable** to the large number of media sessions
  - Detection should be **timely** (i.e., shortly after the event)
  - Decision without sacrificing classification performance
- Formulated as a sequential hypothesis testing problem
  - Use additive feature score to encode belief that a comment is an instance of bullying (or not)
  - Enables 

implementation & meets the goals



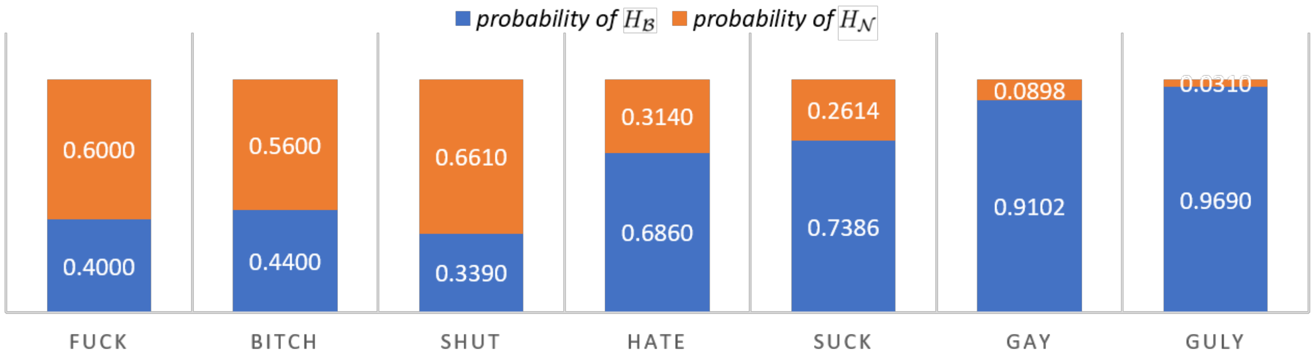
Optimal Online Cyberbullying Detection

[Yao2018, Zois2018]

BULLYING SESSION SAMPLE COMMENTS:

1.BITCHES TALK SHIT ABOUT JIN ALL FUCKIN DAY YO. BITCH GET OFF HIS DICK! GO GET A LIFE IR A JOB OR SOMETHING. GET THE FUCK OFF HIS INSTAGRAM!!!!

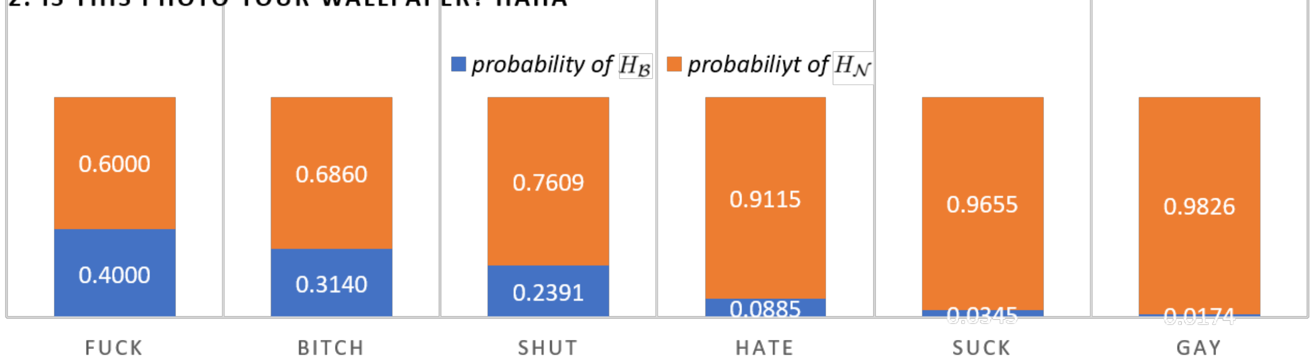
2.THAT SHIT WOULDA ON DA NEWS HOEEEEEEE



NON-BULLYING SESSION SAMPLE COMMENTS:

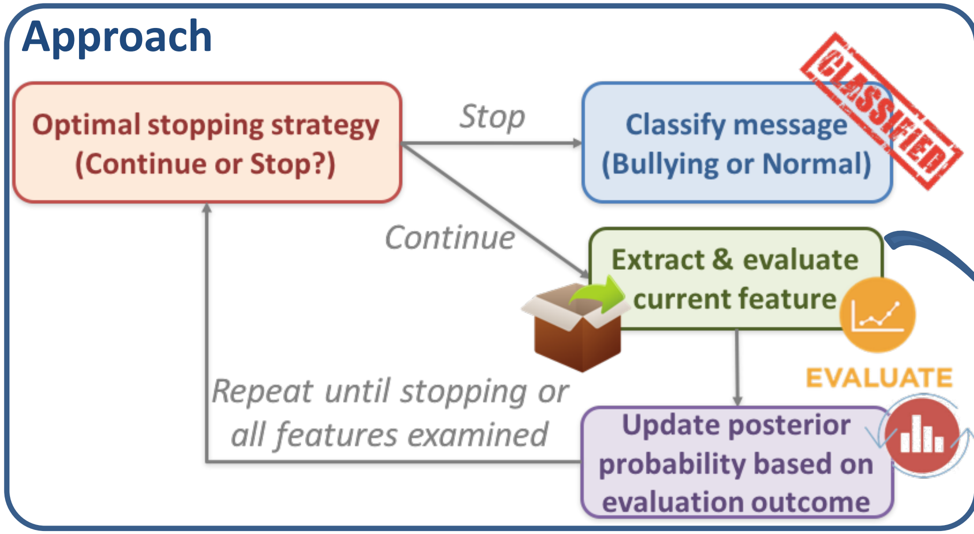
1. I THOUGHT THEY WERE ALL RUMORS HEHE GUESSING WHAT HAPPENED BETWEEN YOU TWO IS TRUE

2. IS THIS PHOTO YOUR WALLPAPER? HAHA



Optimal Online Cyberbullying Detection

[Yao2018, Zois2018]



Type	Features
	# of exclamation marks, # of uppercase letters, # of emoticons, # of acronyms, # of second person pronouns, # of curse hashtags, # of curse words, density of curse words
	mean value of valence, arousal and dominance respectively

Important

A different subset of features may be examined for each comment

offline

Features have been ordered using the heuristic:

$$c_n(p(y_n = 0|H_{\mathcal{N}}) + p(y_n = 1|H_{\mathcal{B}}))$$

Promotes low cost features that at the same time result in few errors

ONLINE

Posterior probability π_0 is set to prior probability of bullying message ρ

- Features are evaluated one at a time
- Update posterior probability as:

$$\pi_n = \frac{p(y_n|H_{\mathcal{B}})\pi_{n-1}}{\pi_{n-1}p(y_n|H_{\mathcal{B}}) + (1-\pi_{n-1})p(y_n|H_{\mathcal{N}})}$$

Optimal Online Cyberbullying Detection

[Yao2018, Zois2018]

Optimization Problem

- **Goal**: use **least number of features** for detecting a cyberbullying message without loss of accuracy

Minimize cost function

$$\min_{R\geqslant 0} \tilde{J}(R) = \min_{R\geqslant 0} \mathbb{E}\left[\sum_{n=1}^{R} c_n + g(\pi_R)\right]$$

- Optimal stopping theory for Markov processes

Optimal Solution

- Optimal solution via dynamic programming (DP):

$$\underbrace{\bar{J}_n(\pi_n)}_{\text{Optimal cost–to–go}} = \min\left[\overbrace{g(\pi_n)}^{\text{Cost of stopping}}, \underbrace{c_{n+1} + \sum_{y_{n+1}} A_n(y_{n+1}) \times \bar{J}_{n+1}\left(\frac{p(y_{n+1}|H_{\mathcal{B}})\pi_n}{A_n(y_{n+1})}\right)}_{\text{Cost of continuing}}\right]$$

Classification Strategy

- Optimal classification strategy:

$$\mathcal{D}_R^{optimal} = \arg\min_{1\leqslant j\leqslant L}\left[C_{\mathcal{B}j}\pi_R + C_{\mathcal{N}j}(1-\pi_R)\right]$$

- Results to the smallest average cost:

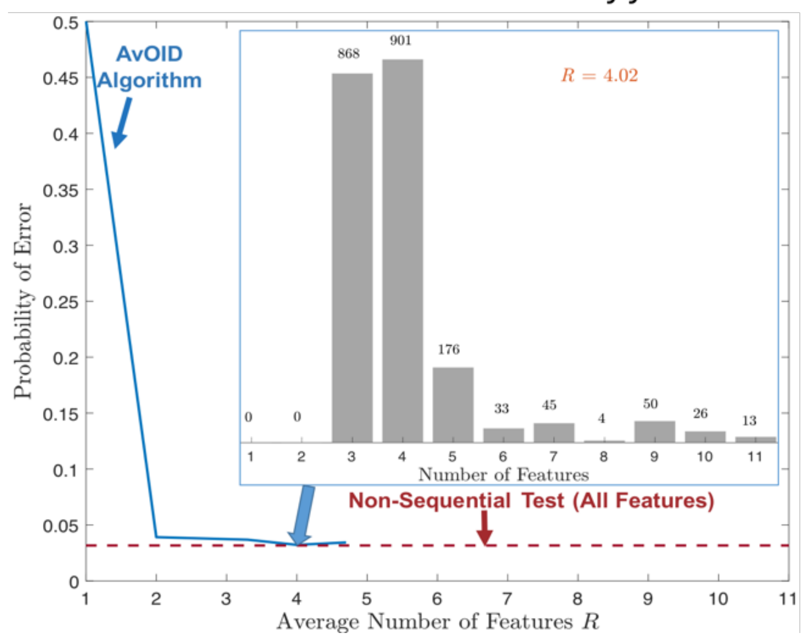
$$\tilde{J}(R) = J(R, \mathcal{D}_R^{optimal}) = \mathbb{E}\left[\sum_{n=1}^{R} c_n + g(\pi_R)\right]$$

Optimal Online Cyberbullying Detection

[Yao2018, Zois2018]

- Evaluation on Twitter dataset [Zois2018]: 10,600 tweets
- Evaluation on Instagram dataset [Yao2018]:
  - 2,218 media sessions in total
    - 19.74% cyberbullying sessions
  - Set0+: 1,296 media sessions with ≥ 0 but $< 40\%$ negativity
    - Unbalanced (15/85 normal/cyberbullying)
  - Set40+: 922 media sessions with 40% of comments containing ≥ 1 profane word
    - Balanced (49/51 normal/cyberbullying)

Performance: *Error? Number of features?*

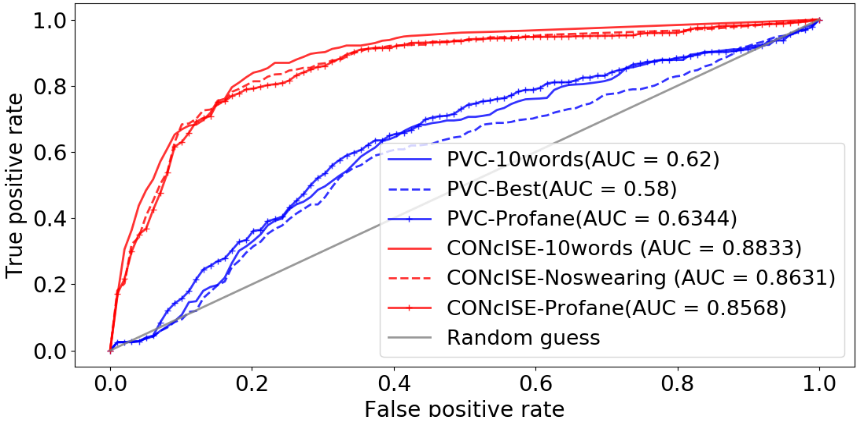


- 3 - 4 features suffice for accurate classification on Twitter
- ~7 features on Instagram

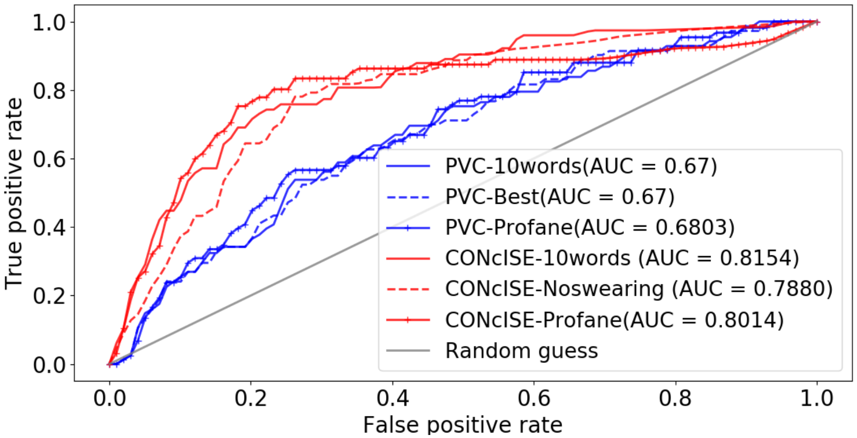
Optimal Online Cyberbullying Detection

[Yao2018]

- Approach is robust to class imbalance



Imbalance ratio 1.5%

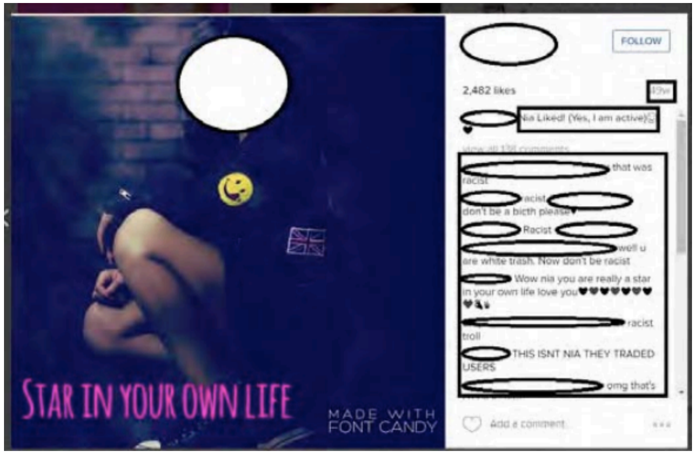


Imbalance ratio 15.7%

Prediction of Cyberbullying Incidents in a Media-Based Social Network

[Hosseinmardi2016]

- **Goal**: predict the occurrence of **cyberaggression / cyberbullying** before it happens by utilizing only initial user data
- **Dataset:**
  - Set0: 1,164 randomly selected media sessions whose comments do not contain any profane words
  - Set0+: 1,296 media sessions with ≥ 0 but $< 40\%$ negativity
    - Unbalanced (15/85 %ratio of normal to cyberbullying sessions
  - Set40+: 922 media sessions with 40% of comments containing ≥ 1 profane word
    - Balanced (49/51 % ratio of normal to cyberbullying sessions)



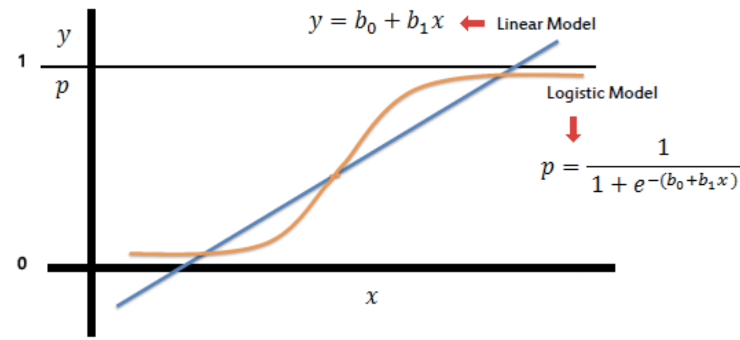
Typical Instagram profile

- **Ground truth:**
  - Each media was labelled by five CrowdFlower contributors

Prediction of Cyberbullying Incidents in a Media-Based Social Network

[Hosseinmardi2016]

- Approach: a logistic regression classifier with forward feature selection
  - Find the feature f_1 that achieves best classification performance
  - Find feature f_2 s.t. (f_1, f_2) achieves best performance
  - Repeat until performance cannot be improved
- Features used
  - Post-time
  - Text caption
  - First few comments
  - Profile (# of shared media)
  - Network features (# of followers/followees)



98% of cyberbullying incidents were captured in Set0+ using the image content feature alone

Adding network features boosts performance significantly for Set0

Prediction of Cyberbullying Incidents in a Media-Based Social Network

[Hosseinmardi2016]

Features	Set	F1-measure	Precision	Recall	False Positive
Image content	Set40+	0.56	0.62	0.51	0.37
Image content	Set0+	0.27	0.15	0.98	0.83
Image content	Set0	-	-	-	0.24
Following, Image content	Set40+	0.62	0.68	0.51	0.18
Following, Image content	Set0+	0.37	0.23	0.91	0.48
Following, Image content	Set0	-	-	-	0.03
Followers, Following, Image content	Set40+	0.68	0.75	0.60	0.22
Followers, Following , Image content	Set0+	0.42	0.28	0.88	0.34
Followers, Following, Image content	Set0	-	-	-	0.05
Media objects ,Followers, Following, Image content	Set40+	0.69	0.77	0.62	0.21
Media objects ,Followers, Following, Image content	Set0+	0.45	0.31	0.87	0.3
Media objects ,Followers, Following, Image content	Set0	-	-	-	0.04
Post time ,User properties, Image content	Set40+	0.67	0.76	0.61	0.22
Post time ,User properties, Image content	Set0+	0.52	0.38	0.88	0.23
Post time ,User properties, Image content	Set0	-	-	-	0.04
Caption ,Post time ,User properties, Image content	Set40+	0.67	0.76	0.61	0.22
Caption,Post time ,User properties, Image content	Set0+	0.57	0.40	0.99	0.23
Caption ,Post time ,User properties, Image content	Set0	-	-	-	0.03
Early Comments, Caption,Post time ,User properties, Image content	Set40+	0.75	0.78	0.72	0.22
Early Comments, Caption,Post time ,User properties, Image content	Set0+	0.66	0.50	1.00	0.14
Early Comments, Caption,Post time ,User properties, Image content	Set0	-	-	-	0.01

Forecasting Hostility on Instagram using Linguistic and Social Features

[Liu2018]

- **Goal**: predict the presence and intensity of hostile comments
  - **Hostile comment**: one that contains harassing, threatening, or offensive language directed toward a specific individual or group

Post-hostility detection

-------------------------------Time------------------------------->

Comment1, Comment2, Comment3, Comment4, stfu, Comment5

Forecast

hostile comment

- **Focus**: teenager community
  - This determines the choice of social media platform
- **Tasks**:
  - Hostility presence forecasting
  - Hostility intensity forecasting
- **Dataset**: ~1K Instagram media sessions

	posts	comments	hostile comments
hostile posts	591	21,608	4,083
non-hostile posts	543	9,379	0
total	1,134	30,987	4,083

Forecasting Hostility on Instagram using Linguistic and Social Features

[Liu2018]

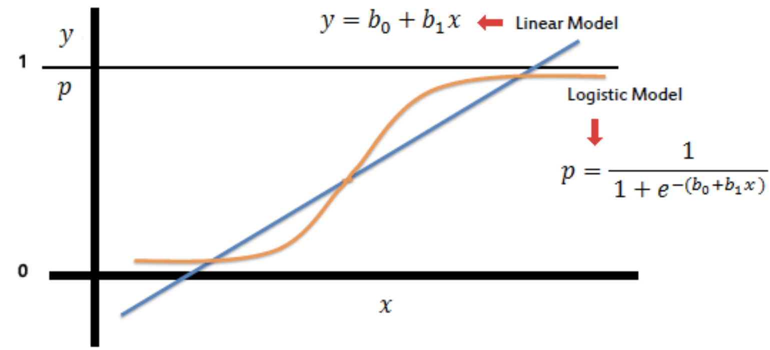
- Hostility presence forecasting
  - **Given** the **initial sequence** of non-hostile comments in a media session
  - **Predict** whether some future comment will be **hostile**
- Hostility intensity forecasting
  - **Given** the **first hostile comment** in a media session
  - **Predict** whether the media session will receive more than N hostile comments in the future

- Solutions to the first task could be used to eliminate all hostile comments from the system
- Solutions to the second task could be used for targeted interventions on the most extreme cases

Forecasting Hostility on Instagram using Linguistic and Social Features

[Liu2018]

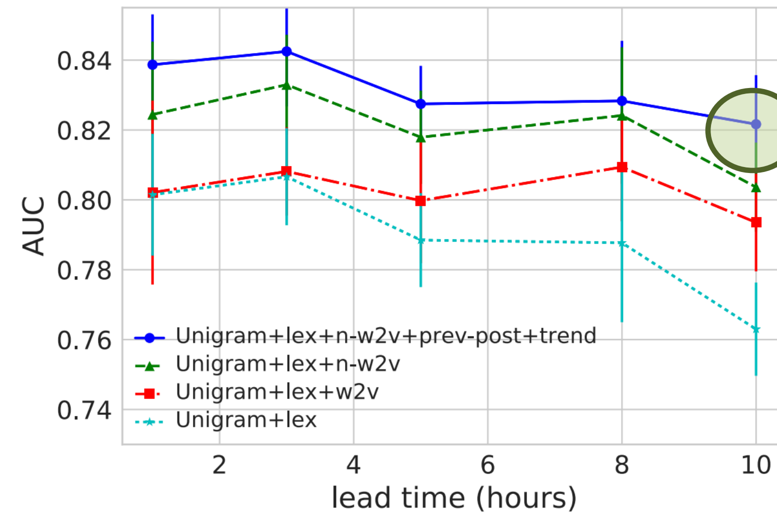
- Approach:
  - Logistic regression trained on first N comments of each media session
- Features:
  - Unigrams
  - Word2vec [Mikolov2013]
  - N–gram character word2vec [Bojanowski2017]
  - Hatebase (www.hatebase.org)
  - ProfaneLexicon (www.cs.cmu.edu/~biglou/resources/)
  - Comments from previous media sessions
  - Comments on previous media sessions by the author
  - Trend: conversation trajectory
  - User activity: participant diversity

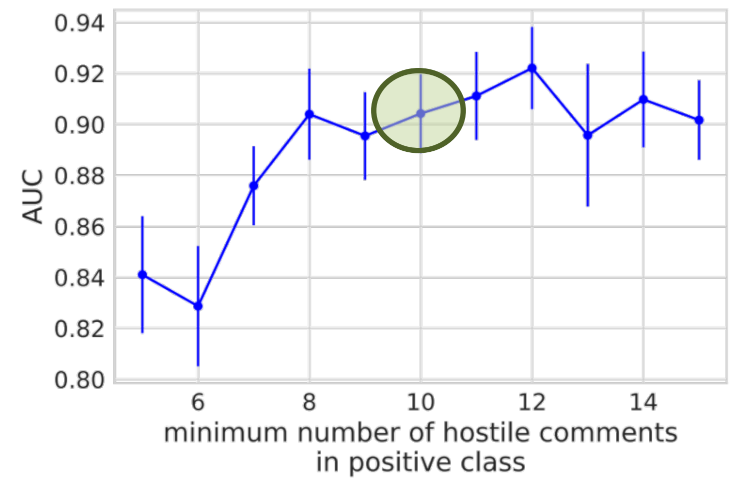


Forecasting Hostility on Instagram using Linguistic and Social Features

[Liu2018]

- Evaluation Methodology
  - 10-fold cross-validation experiments to measure the forecasting accuracy for each task
- Evaluation Results
  - Presence
    - Can predict that a hostile comment will arrive 10 hours in the future with ~.82 AUC
  - Intensity
    - Distinguishes between posts that will have 1 versus 10 or more hostile comments with ~.90 AUC

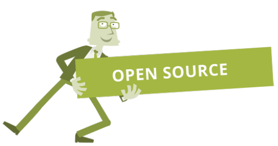




Forecasting Hostility on Instagram using Linguistic and Social Features

[Liu2018]

- Prominent predictors of future hostility on Instagram media sessions
  - Whether the author of the media session has received hostile comments in the past
  - Use of user–directed profanity
  - Number of distinct users participating in a media session
  - Trends in hostility over time

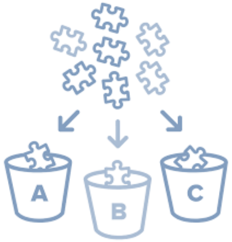


Code available at: https://github.com/tapilab/icwsm-2018-hostility

Section

Mitigation Strategies

Taxonomy of Mitigation Strategies



[AlMazari2013]

- Prevention/mitigation strategies can be adopted at different levels

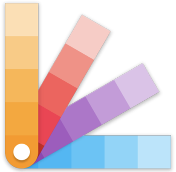
Technological Approach
parental control services | online services rules | online memberships rules | Firewall blocking services | text-messaging control | mobile parental control | anti-spam and malware | slanderous emails blocking | online reporting facilities | online information services | IP address hiding and back tracking applications

Educational and Awareness Approach
educating of end-users | coping strategy | improving the technical skills | improving the cognitive skills | awareness active campaigns | awareness workshops | social responsibilities | awareness training | awareness forums | media channels

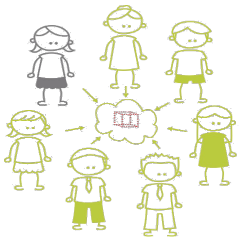
Psychological Approach
talking and listening to cyber-victims | making new relations | joining social clubs | minimize self-transcendence and self-oriented | improve levels of trust | open communications channels | create trusted social groups | build confidence | create comfortable environments | improving mental health | enhance self-esteem | provide counseling services

Administrative Approach
policy development | enhance workplace environment | regulate using free services | identify and apply penalties of misuse | regulations and laws | developing mentoring programs | proper training | bully-box and locked containers strategy

Broad Themes of Mitigation Research



- Psychology, public health, sociology, criminology, and other related behavioral and social sciences (e.g., [Kraft2009], [Kazerooni2018])
  - consider prevention/mitigation scenarios
  - conduct **surveys** and **focus groups**
  - analyze findings and report correlations between different variables



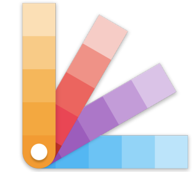
Summary of Cyber bullying Prevention Strategies

Strategy Number	Question	Highlight of strategy	Strategy stated on questionnaire
1	Q-11-a	No computer use in school and home for offender	Cyber bullies would not be allowed to use the computer at home and school. Any assignments for school that required using the library would have to be done at the library using books.
2	Q-11-b	Sending offender to another school	Sending cyber bullies to an "alternative" school away from their regular school as punishment.
3	Q-11-c	Parent taking away offender's computers and cell phones	Parents would take away a cyber bully's cell phone and computer.
4	Q-11-d	Offender paying victim money	Cyber bullies would have to get a job and pay money to the person they bullied online.
5	Q-11-e	One year delay to a 4-year college for offender	Repeat cyber bullies would not go to 4 year colleges. They would have to spend at least one year at a community college before going to a 4 year college. It would not matter how well they did in high school.

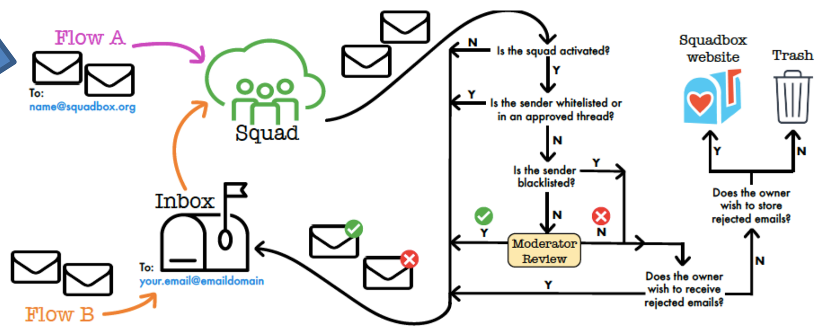
[Kraft2009]

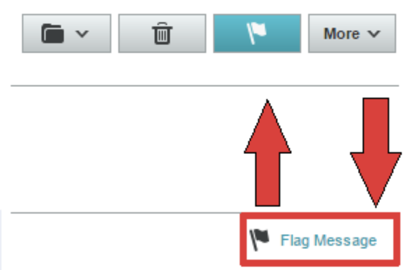
[Kazerooni2018]

Mitigation Themes



- Computer and information sciences, and engineering develop **technological solutions** to prevent/mitigate cyberbullying
  - Report/control/warn about message content (e.g., [Vishwamitra2017], [Bowler2014], [Dinakar2012], [Ashktorab2016], [Cohen2014], [Mahar2018], [Fan2016])
  - Provide support for victims (e.g., [Vishwamitra2017], [Dinakar2012], [Ashktorab2016], [Cohen2014], [vanderZwaan2013], [Fan2016])
  - Educate both victims and bullies (e.g., [Vala2012], [Dinakar2012], [Ashktorab2016], [Bowler2014])





Dealing with Cyberbullying

Cyberbullying typically has a detrimental effect on its victims. Victims often feel helpless and as a result suffer from depression, anxiety, and social isolation. There are many practices that you can take to prevent cyberbullying from happening to you or anyone else in your environment.

How to react to cyberbullying

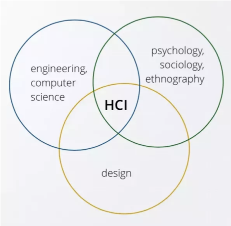
- **Cut off the bully**—If the bully is making direct communication with you, tell them to stop. If he or she refuses to stop, block him or her from the communication channel he or she is using to harass you. Studies have shown that bullies typically bullying to seek attention and will often stop if they are ignored.
- **Record**—If the bully continues to harass you, keep records of all the communication, i.e. phone calls, messages, posts, e-mails, sent. If the bullying is physical as well, record the time of the event and what happened. For phone calls, dialing *57 before the end of a call will have the bully's phone number recorded by the phone company. These records will serve as important evidence against the bully.
- **Reach out**—Report cyberbullying to someone in authority such as your administrators, teachers, or managers. You can also report cyberbullying to the police, as undesired repeated harassment is considered a criminal offence. It may also be helpful to talk to close friends and family for emotional support. There are also many helplines and counselors that you can reach out to to seek help.
- **Report to Service Provider**—Many service providers have terms of use agreements that its users are required to follow regarding decorum and etiquette. Reporting the cyberbullying incident can get them banned from the platform. Moreover, the service provider may also be able to track down the identity of anonymous bullies and remove defamatory content.
- **Obtain a Civil Restraining Order**—You may be able to obtain a restraining order so the bully can no longer interact with you legally.

What Not to Do

- **Become a cyberbully yourself**—Sinking to the bully's level will not help to solve the problem. You are only becoming a bully yourself and will make other suffer as you have.
- **Broadcast the message**—Do not forward or share the message with others who are not aware of the situation. Messages forwarded to people who are not aware of context can exacerbate the problem greatly.
- **Let the bully get to you**—No one deserves to be bullied or harassed at all. The inappropriate behavior of bullies often has nothing to do with the victim. Bullies tend to be insecure people with problems who are taking it out on other people as a means of release. They are cowards who have no courage to deal with their own problems.

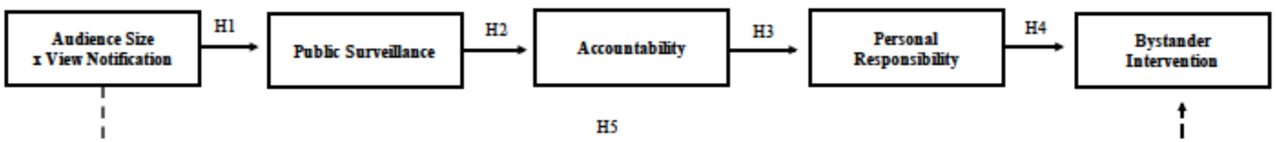
References

[1] R. C. Lohmann, "Taking on Cyberbullying", 2014, Psychology Today, http://www.psychologytoday.com/blog/teen-angst/201011/taking-cyberbullying.

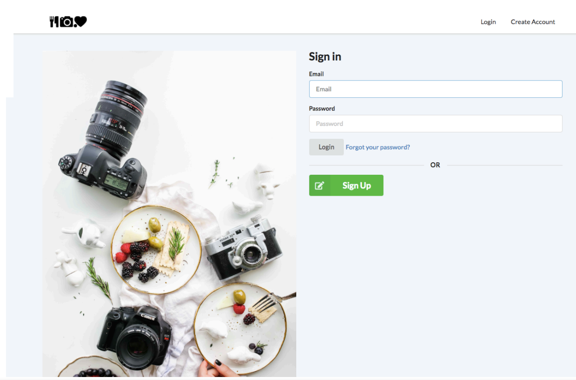


Mitigation Themes

- Joint effort between computer and social scientists to understand behavior of users in realistic environments (e.g., [Ashktorab2017], [DiFranzo2018])
  - Design/Develop **social media site** for experimentation
  - Perform controlled study
  - **Post-study survey**
  - Analyze findings and report correlations between different variables (e.g., bystander engagement and number of views of a post) to **prove/disprove hypotheses**



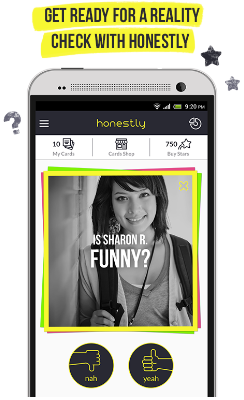
Conceptual Model of Bystander Intervention in Cyberbullying
[DiFranzo2018]



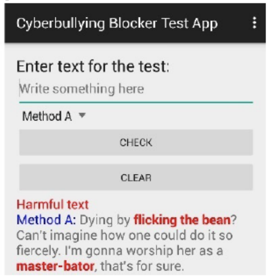
EatSnap.Love social networking site

Existing Mitigation Technology

- Apps to promote well-being of social media users
  - “You’re Valued” searches Twitter for tweets that say “nobody likes me” and then sends a response tweet with messages like “I like you”, “You’re valued”, or “You matter” [White2014]
  - “Honestly” asks friends of a particular user question like “Can I sing well?” and shares positive responses with a user [Shaul2015]
  - “No More Bullying Me!” provides online meditation techniques to support victims [NoMoreBullyingMeApp]



- Apps to inform user of harmfulness of a message before sending
  - “ReThink” shows pop-up warning message when user tries to send harmful message [ReThinkApp]
  - “Cyberbullying Blocker” warns user of harmful message while indicating harmful words [Lempa2015]



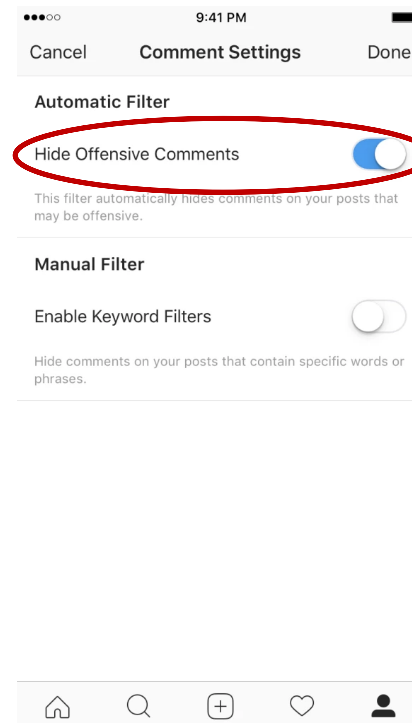
Existing Mitigation Technology

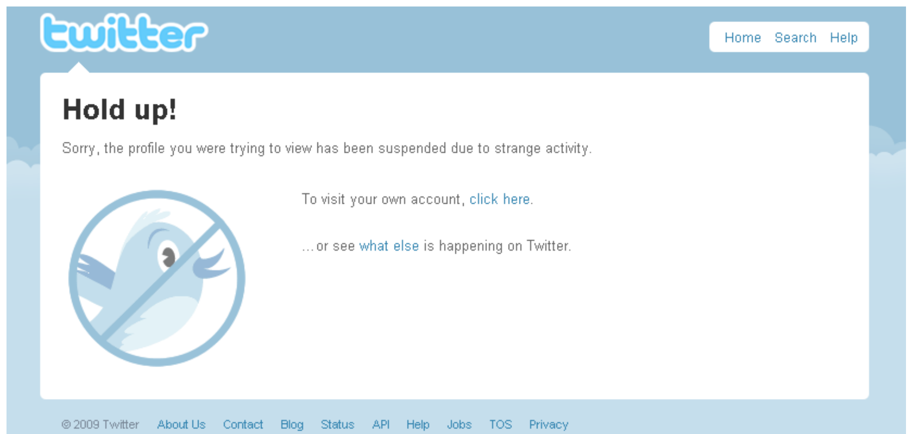
- Report/monitoring of cyberbullying messages, e.g.,
  - Apps such as "PocketGuardian" [PocketGuardianApp] and "Bark–Monitor.Detect.Alert" [BarkApp] report inappropriate material to parents
  - Twitter allows users to report harassment tweets and blocks accounts of bullies until they erase these tweets
  - App "Anonymous Alerts" helps students anonymously submit bullying incidents to school officials [AnonymousAlertsApp]
  - Facebook allows reporting, unfriending and blocking individuals [FBStopBullying]
  - Instagram allows reporting and blocking individuals

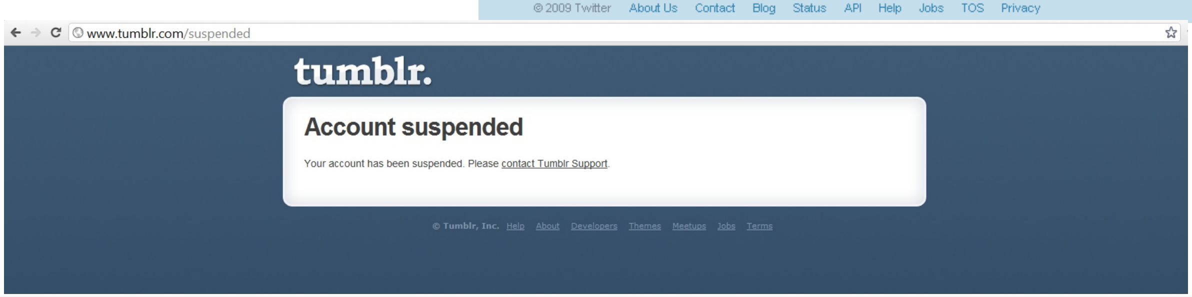
- Improve awareness about cyberbullying, e.g.,
  - App "Cyberbully Zombies Attack" helps individuals learn how to handle cyberbullying [CyberbullyZombiesAttackApp]
  - App "Cyber-Bullying First-Aid App" provides resources to combat cyberbullying [CBFirstAidApp]

Existing Mitigation Technology (2)

- Review/take actions on content and inform administrators
  - Instagram automatically hides toxic comments and alerts administrators [InstagramHideComments]
  - Ask.fm reviews images for harmful content before upload [Askfmhelp]
  - Twitter suspends accounts that violate Twitter rules [TwitterRules]
  - …





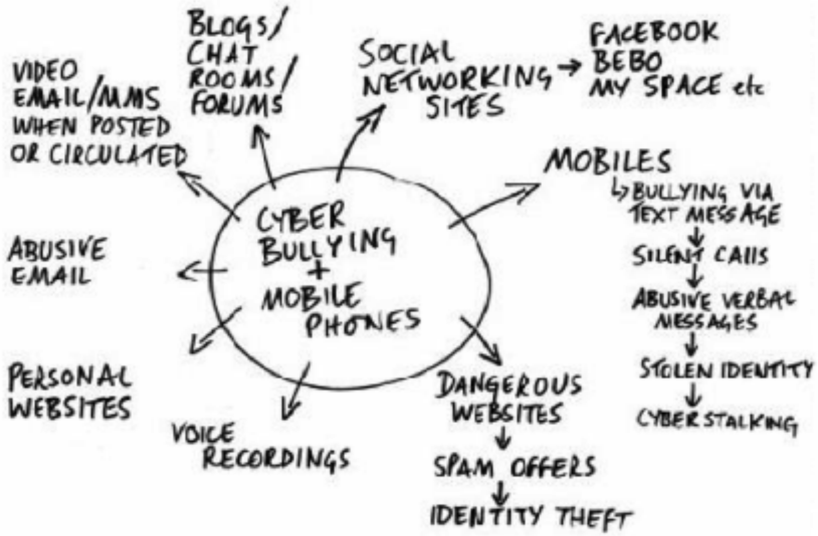


Using Computer Technology to Address the Problem of Cyberbullying

[Cohen2014]

- **Goal**: provide assistance for victims and bullies
  - **Detect** cyberbullying incidents
  - Report of cyberbullying incidents
  - Integrate third-party assistance when cyberbullying is detected
  - Facilitate authorities to take actions against detected bullies

Mitigation



Instances of cyberbullying

Using Computer Technology to Address the Problem of Cyberbullying

[Cohen2014]

- **Cyberbullying detection**:
  - Label malicious messages: *model* ***reputation*** *of each message using* ***users' feedback*** *and assign warning label to potential instance of bullying*
    - Score r_i of message *i*:

$$r_i = \frac{\#\text{positive votes} + 1}{\#\text{negative votes} + 1}$$

    - User *u*'s reputation score:

$$\text{reputation}(u) = \frac{1}{n}\sum_{i=1}^{n} r_i$$

< 1 → User *u* is not malicious

> 1 → User *u* is malicious

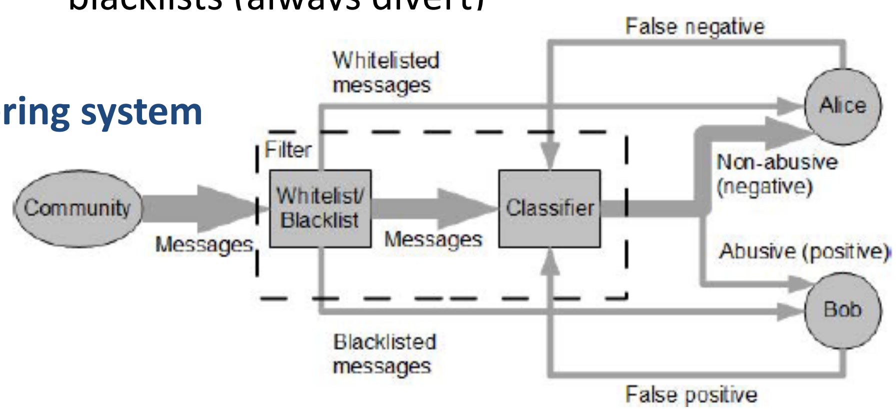
  - Proposed approach combines:
    - Positive and negative reviews of messages by user's social network audience, and
    - Standard machine learning methods based on textual feature
  - Assertion: Reputation scores can potentially help identify bullies and victims (e.g., user with many friends that have negative score can be a victim)

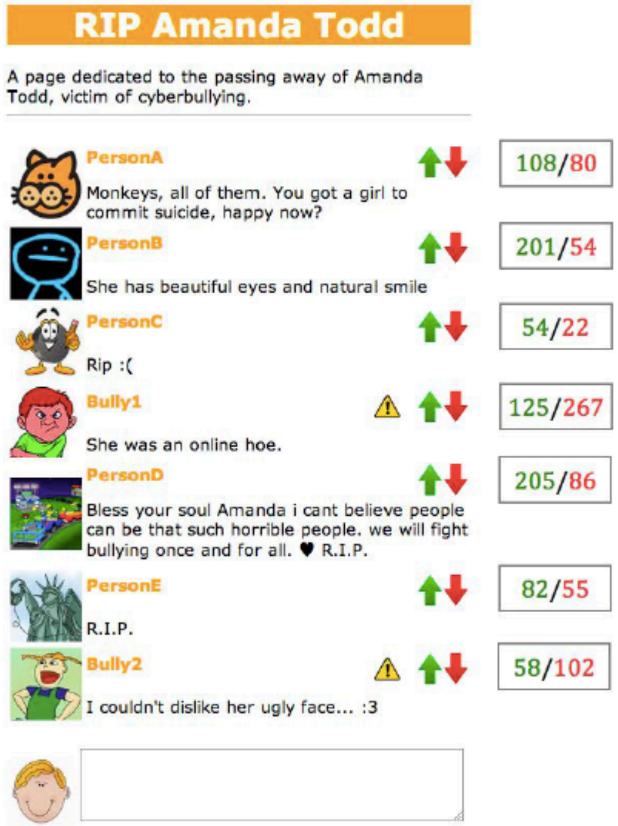
Using Computer Technology to Address the Problem of Cyberbullying

[Cohen2014]

- Filter suspected messages: *classify messages as abusive or non-abusive using bag-of-word, sentiment and sender information features incorporating* ***trusted third party***
  - Divert possibly abusive messages to a trusted third party (e.g., parent, friend)
  - Third party can
    - delete or report abusive message
    - inform filter of non-abusive message
  - Users may create whitelists (always deliver) and blacklists (always divert)

Filtering system





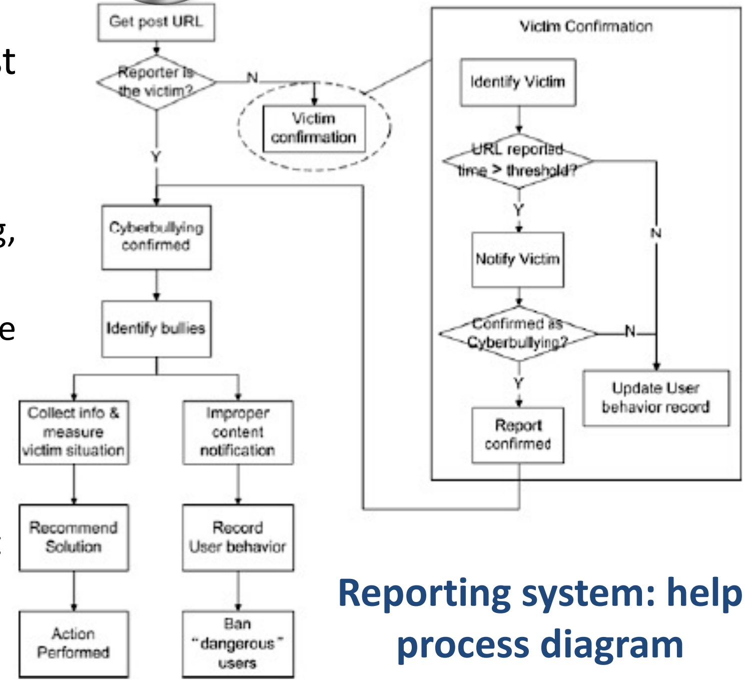
Message thread with flagged malicious messages and reputation scores

Using Computer Technology to Address the Problem of Cyberbullying

[Cohen2014]

- **Mitigation**:
  - Reporting system with third party assistance: *victims or their friends can report bullies and their messages (user reports)*

  - **Reporting phase**: provide source of improper post and define user role (victim or friend)
  - **Victim confirmation phase**: affirm reported post as improper
  - **Victim helping phase (protection)**:
    - Identify type of harassment (e.g., bullying, stalking, privacy leaking)
    - Select solution (e.g., access legal aid, disable sharing of post, blacklist message)
  - **Improper online behavior phase (monitoring)**:
    - Notify bullies of improper behavior
    - Constrain/ban account or notify law enforcement

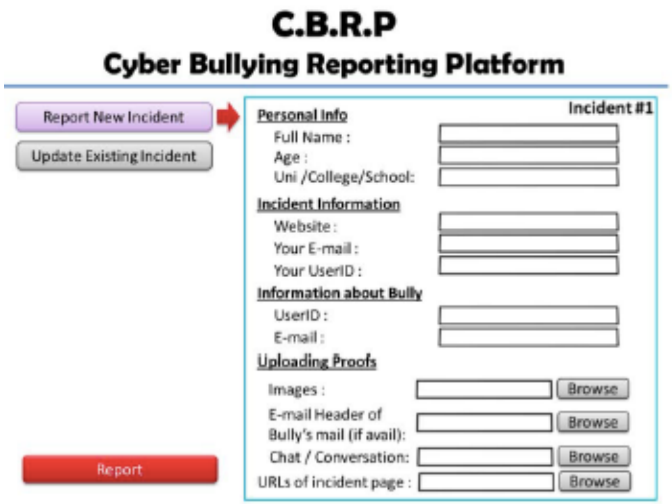


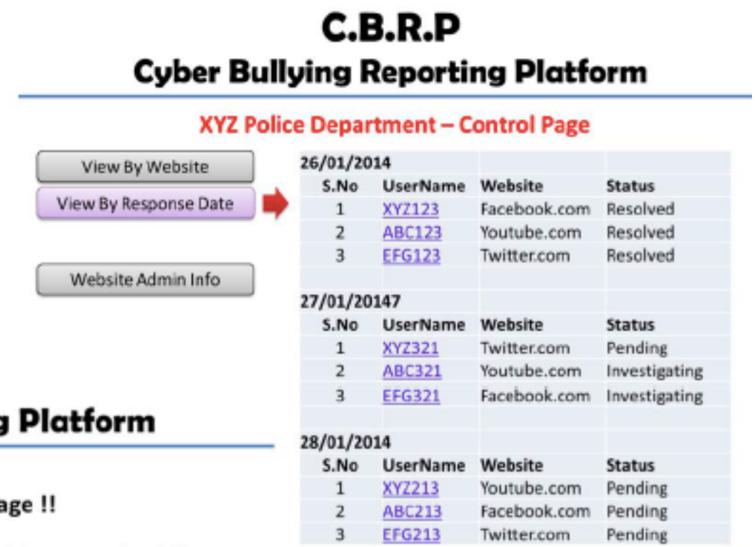
Reporting system: help process diagram

Using Computer Technology to Address the Problem of Cyberbullying

[Cohen2014]

– Centralized reporting platform: *web portal managed by authorities where victims and witnesses can report incidents*





Acknowledge and provide encouragement

Using Computer Technology to Address the Problem of Cyberbullying

[Cohen2014]

- Education: *provide educational resources to both victims and bullies, e.g.,*
  - Be mindful and thoughtful of message contents
  - Phone number of support centers
  - Educational tests for bullies

Cyberbullying

Cyberbullying is defined as the use of technology to support deliberate, hostile and hurtful behaviour towards an individual or group of individuals [2].

Why People Cyberbully

Just like other forms of bullying, cyberbullying is about gaining power and control. Those who bully others are trying to establish dominance over people they perceive to be weaker than them. While technology can be used as a positive communication tool it can also be used to hurt others [1].

In scientific studies, it has been found that people engage in cyberbullying activities to direct their frustration, anger, hurt, and difficulty they are experiencing elsewhere. Some also do so due to lack of attention from friends and family. Others bully to fit in with their friends, in cases of group bullying [1].

Impact of Cyberbullying

- Feel helpless, angry, depressed, and/or anxious
- Feel unsafe in cases that the bully is anonymous
- Feel shame and embarrassment in a worldwide venue
- Surprise at how communicate and content can be blown out of context
- Have a tendency to isolate oneself from social group
- Feel that harassment cannot be avoided because technology is easily accessible
- More susceptible to self-inflicted harm and even suicide

The Law

Some forms of online bullying are considered criminal acts. Under the Criminal Code of Canada it is a crime to:

- Communicate repeatedly with someone if the communication casues them to fear their own safety or the safety of others
- Write something that is designed to insult a person or likely to injure a person's reputation by exposing them to hatred, contempt or ridicule.

A person may also be violating the Canadian Human Rights Act, if he or she spreads hate or discrimination based on race, national or ethnic origin, colour, religion, age, sex, sexual orientation, marital status, family status or disability [3].

What to Do

You can be prosecuted for involvement in cyberbullying. Here are some tips to not be a cyberbully:

- Think before you click! Consider the recipient's feelings before sending the message. Chances are that if you would not say it person-to-person, then you should not be posting the message.
- If a group of your friends are cyberbullying an individual, do not participate. Notify an authority.
- Private messages between you and another person should not be publically shared.
- If you are bullying to seek attention or because of difficulties in your life, speak with an adult and seek the proper social support needed.

References

[1] D. Bridgett, A. Grippo, J. Magliano, Why Do Some Kids Cyberbully Other?, 2013, Psychology Today, http://www.psychologytoday.com/blog/the-wide-wide-world-psychology/201304/why-do-some-kids-cyberbully-others

[2] J. Will and C. Clayburn, The Psychological Impact of Cyber Bullying, 2010, University Business, http://www.universitybusiness.com/article/psychological-impact-cyber-bullying

[3] Stalking, criminal harassment and cyberbullying, 2013, The Canadian Bar Association – British Columbia, http://www.cba.org/dev/BC/public_media/criminal/206.aspx

Cyberbullying Discipline Quiz

This is a quiz to ensure that you have reviewed the cyberbullying document. You are required to score at least 80% before being able to return to the social network.

LIKE

What can you do to not be a cyberbully?

- None of the Above
- Create hate websites about an individual
- Take revenge on ex by posting naked pictures of him or her.
- Send offensive messages repeatedly to a person who you do not like.

Why do people cyberbully?

- To direct their frustration, anger, hurt, and difficulty they are experiencing elsewhere
- To fit in with their, in cases of group bullying
- Due to lack of attention from family and friends
- All of the Above

Dealing with Cyberbullying

Cyberbullying typically has a detrimental effect on its victims. Victims often feel helpless and as a result suffer from depression, anxiety, and social isolation. There are many practices that you can take to prevent cyberbullying from happening to you or anyone else in your environment.

How to react to cyberbullying

- **Cut off the bully**—If the bully is making direct communication with you, tell them to stop. If he or she refuses to stop, block him or her from the communication channel he or she is using to harass you. Studies have shown that bullies typically bullying to seek attention and will often stop if they are ignored.
- **Record**—If the bully continues to harass you, keep records of all the communication, i.e. phone calls, messages, posts, e-mails, sent. If the bullying is physical as well, record the time of the event and what happened. For phone calls, dialing *57 before the end of a call will have the bully's phone number recorded by the phone company. These records will serve as important evidence against the bully.
- **Reach out**—Report cyberbullying to someone in authority such as your administrators, teachers, or managers. You can also report cyberbullying to the police, as undesired repeated harassment is considered a criminal offence. It may also be helpful to talk to close friends and family for emotional support. There are also many helplines and counselors that you can reach out to to seek help.
- **Report to Service Provider**—Many service providers have terms of use agreements that its users are required to follow regarding decorum and etiquette. Reporting the cyberbullying incident can get them banned from the platform. Moreover, the service provider may also be able to track down the identity of anonymous bullies and remove defamatory content.
- **Obtain a Civil Restraining Order**—You may be able to obtain a restraining order so the bully can no longer interact with you legally.

What Not to Do

- **Become a cyberbully yourself**—Sinking to the bully's level will not help to solve the problem. You are only becoming a bully yourself and will make other suffer as you have.
- **Broadcast the message**—Do not forward or share the message with others who are not aware of the situation. Messages forwarded to people who are not aware of context can exacerbate the problem greatly.
- **Let the bully get to you**—No one deserves to be bullied or harassed at all. The inappropriate behavior of bullies often has nothing to do with the victim. Bullies tend to be insecure people with problems who are taking it out on other people as a means of release. They are cowards who have no courage to deal with their own problems.

References

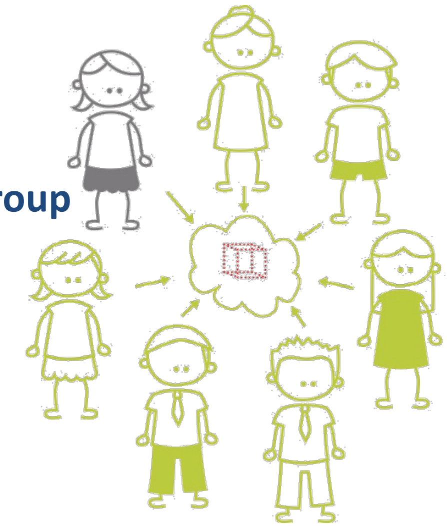
[1] R. C. Lohmann, "Taking on Cyberbullying", 2014, Psychology Today, http://www.psychologytoday.com/blog/teen-angst/201011/taking-cyberbullying.

Designing Cyberbullying Mitigation and Prevention Solutions Through Participatory Design with Teenagers

[Ashktorab2016]

- Goal: design cyberbullying mitigation solutions
  - Participatory design with two high school student groups (9th and 12th grade) in spring 2015 (five design sessions per group)
    - Participants shared their experiences, iteratively designed potential solutions and identified challenges
  - Discussion of findings and presentation of potential cyberbullying mitigation solutions

Focus group



Designing Cyberbullying Mitigation and Prevention Solutions Through Participatory Design with Teenagers

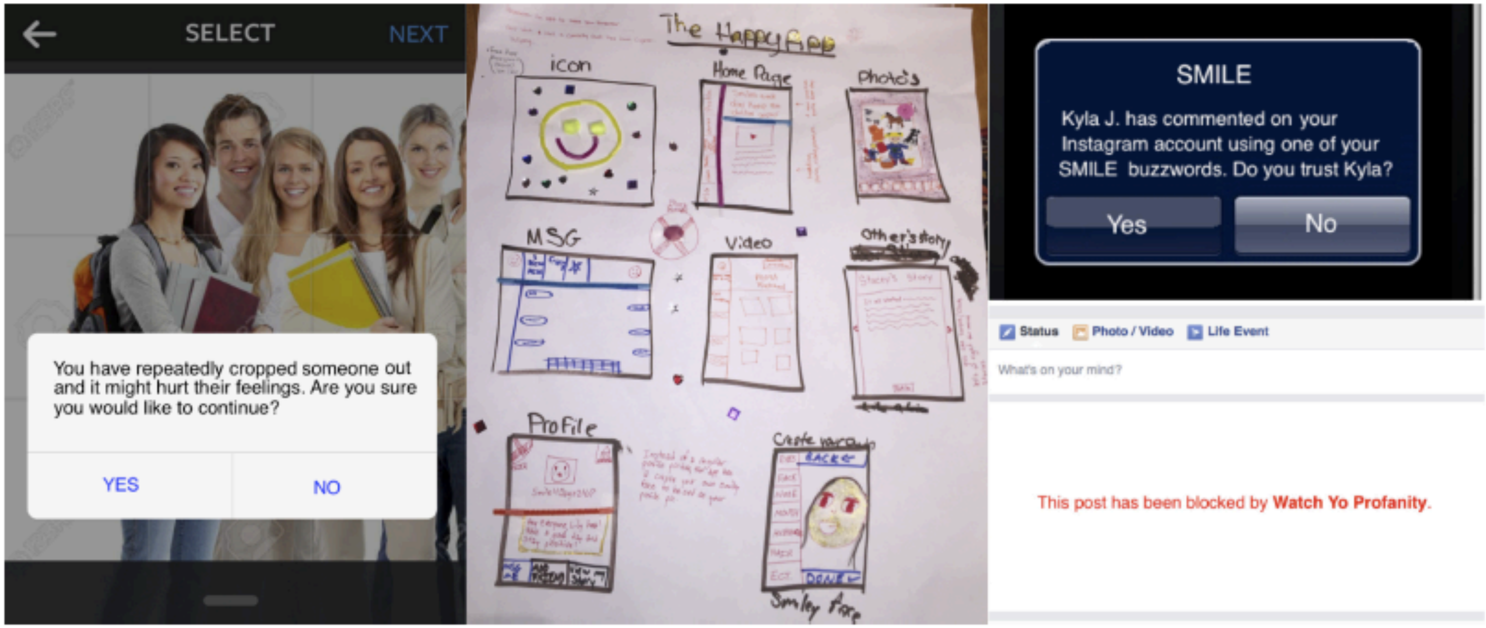
[Ashktorab2016]

- Hypothesis: children who are experiencing and engaging in cyberbullying can be viewed as **domain experts** of cyberbullying
- Design activities:
  - Focus groups: *how participants interact with online social media platforms and how these platforms are used for cyberbullying*
  - Scenario centers: *think technological and non-technological solutions to mitigate negative behaviors in online social media platforms based on scenarios*
  - Bags of staff: *participants were asked to design solution for specific cyberbullying event*
  - Mixing ideas: *encourage participants to think about common themes between their solutions to create better solutions and prototypes*
  - Evaluating prototypes: *discuss feasibility and limitations of each solution*

Designing Cyberbullying Mitigation and Prevention Solutions Through Participatory Design with Teenagers

[Ashktorab2016]

- Findings:
  - Cyberbullying victims either do nothing or turn to a friend for support
  - Focus on social media platforms that teenagers are mostly using (i.e., Instagram, Snapchat)



Designing Cyberbullying Mitigation and Prevention Solutions Through Participatory Design with Teenagers

[Ashktorab2016]

- Nine (9) design applications
  - Control posted content (“SMILE”, “Watch Yo Profanity”, “Reporting Bullies with Feedback”, “Hate Page Prevention”)
  - Emotional support and respond back strategies for victims (“Happy App”, “Fight Back”, “Positivity Generator”, “The Broiler”)
  - Education of bullies (“Exclusion Prevention”)
- Timely support after cyberbullying occurs is vital part of mitigation
- Limitations
  - Trust in accuracy of filtering algorithms
  - “Bullying the bullies” is not ethically sound solution
  - Evaluation of effectiveness of cyberbullying prevention mechanisms in practice

Common Sense Reasoning for Detection, Prevention, and Mitigation of Cyberbullying

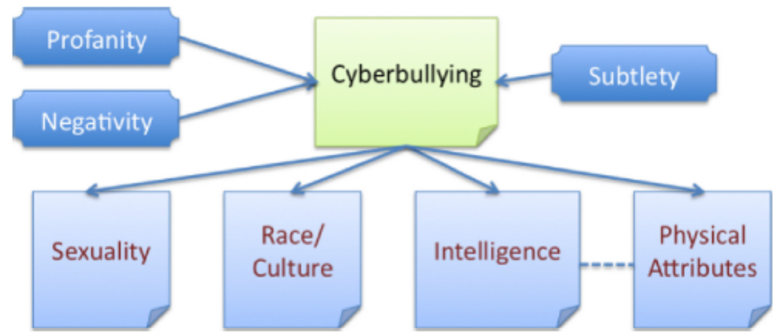
[Dinakar2012]

- **Goals**:
  - Design techniques for **effective cyberbullying detection**
  - Develop **reflective user interfaces** that encourage users to reflect upon their behavior and their choices
- Cyberbullying detection: *combine state-of-the-art natural language processing with common sense reasoning (AnalogySpace) based on common sense knowledge base (BullySpace)*
  - Evaluation on formspring.me and YouTube datasets
- “Air traffic control”-like dashboard: *alert moderators to large-scale cyberbullying outbreaks and facilitate prioritization*
- Educational materials for victims: *how to cope with situation and connect with emotional support*

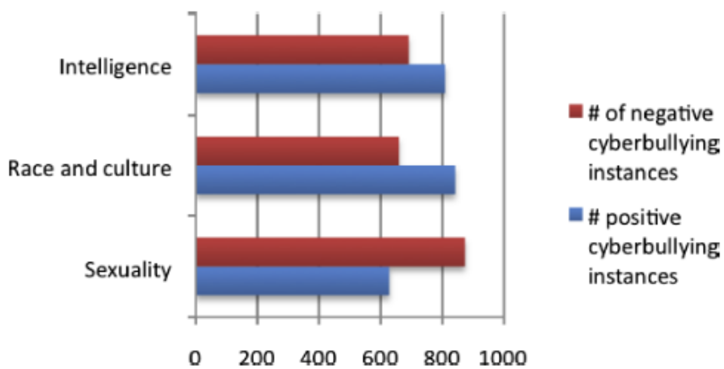
Common Sense Reasoning for Detection, Prevention, and Mitigation of Cyberbullying

[Dinakar2012]

Cyberbullying topics sensitive to victim



- **Cyberbullying detection**:
  - Focus on **textual cyberbullying**
  - How to find insulting language when there is no explicit profane or negative language?



Label/Annotation	# of positive cyberbullying instances	# of negative cyberbullying instances
Sexuality	627	873
Race & Culture	841	659
Intelligence	809	691

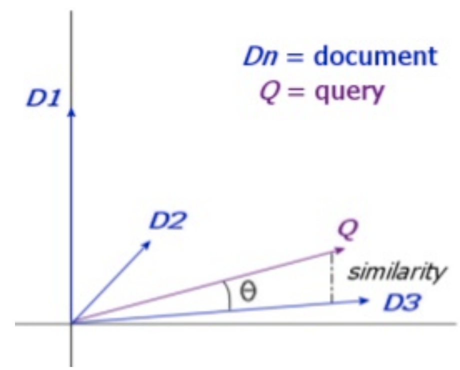
  - Datasets: manually labeling process (3 annotators)
    - YouTube: comments of controversial and non-controversial topics
    - FormSpring: actual user- or moderator-flagged cyberbullying instances
  - Methods:
    - Naïve Bayes
    - JRip (incrementally learn rules and optimize them)
    - J48 (tree-based classifier)
    - SVM

Common Sense Reasoning for Detection, Prevention, and Mitigation of Cyberbullying

[Dinakar2012]

- Features common among sexuality, race and culture, and intelligence, as well as specific features for each of them separately
- BullySpace: (*based on Formspring dataset*)
  - Knowledge base about commonly used stereotypes employed to bully individuals based on their sexuality
- AnalogySpace:
  - Each question about a concept can be thought of as a "dimension" of a concept space
  - Answering a question can be thought of as projecting the concept onto a specific dimension
  - Singular Value Decomposition (SVD) is used for dimensionality reduction
  - Resulting space helps determine which concepts are similar

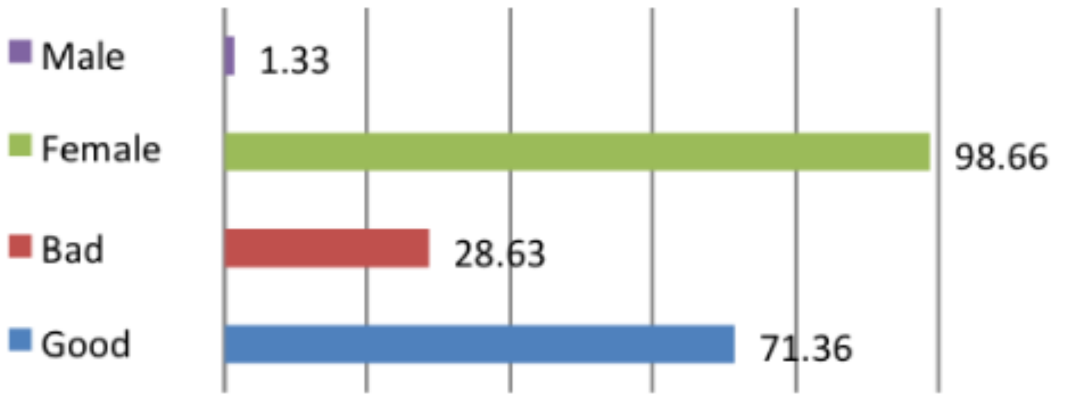
Feature	Type
TF-IDF	General
Ortony lexicon for negative affect	General
List of profane words	General
POS bigrams: JJ_DT, PRP_VBP, VB_PRP	General
Topic-specific unigrams and bigrams	Label-specific



Common Sense Reasoning for Detection, Prevention, and Mitigation of Cyberbullying

[Dinakar2012]

- Common sense reasoning example:
  - “**Hey Brandon, you look gorgeous today. What beauty salon did you visit?**”
  - If this comment is aimed at a boy, it might be an implicit way of accusing the boy of being effeminate (cyberbullying instance candidate)



	Good	Bad	Female	Male
Similarity	71.36	28.63	98.66	1.33

Analysis of sentence relationship with certain concepts

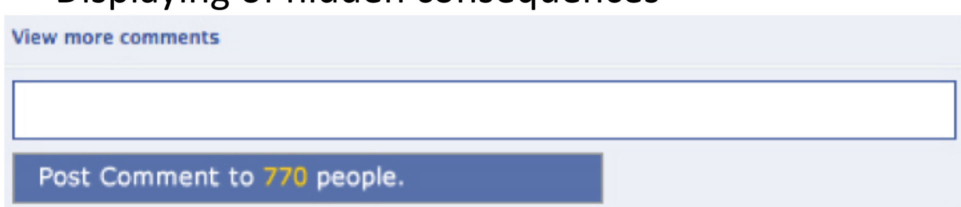
Common Sense Reasoning for Detection, Prevention, and Mitigation of Cyberbullying

[Dinakar2012]

- **Intervention strategies**:
  - Reflective user interfaces: *encourage positive digital behavioral norms*
    - Notifications (i.e., reflect on consequences)
    - Interactive tailored education

    - Action delays

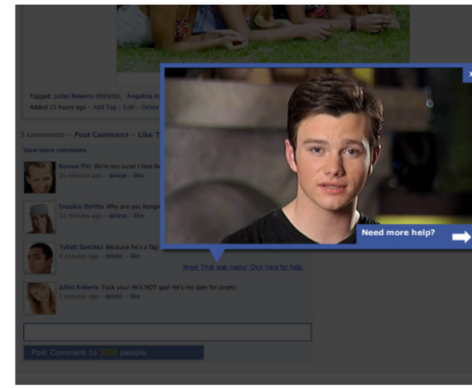
    - Displaying of hidden consequences



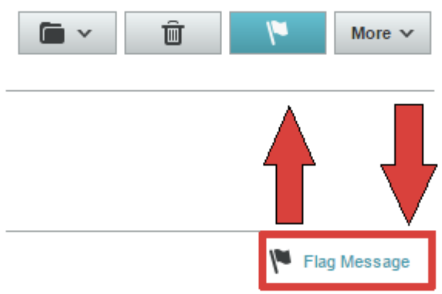
Common Sense Reasoning for Detection, Prevention, and Mitigation of Cyberbullying

[Dinakar2012]

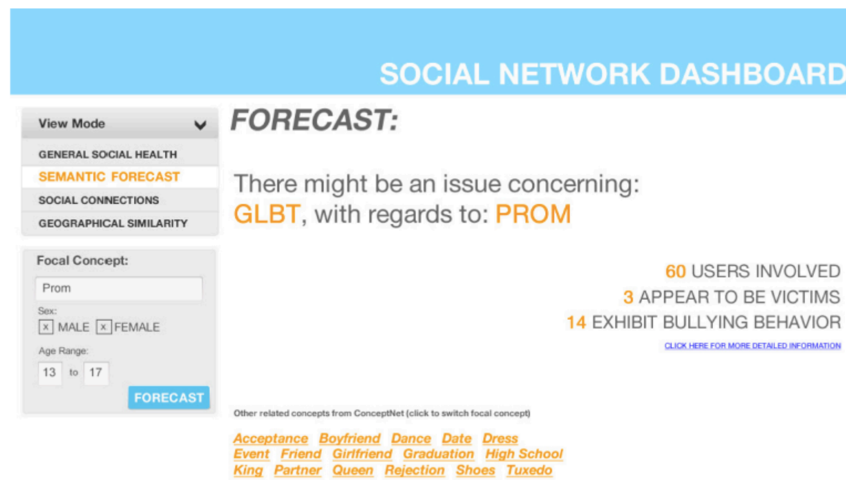
- Interactive educational support



- System-suggested flagging



- Visualization
  – Assist authorities to monitor

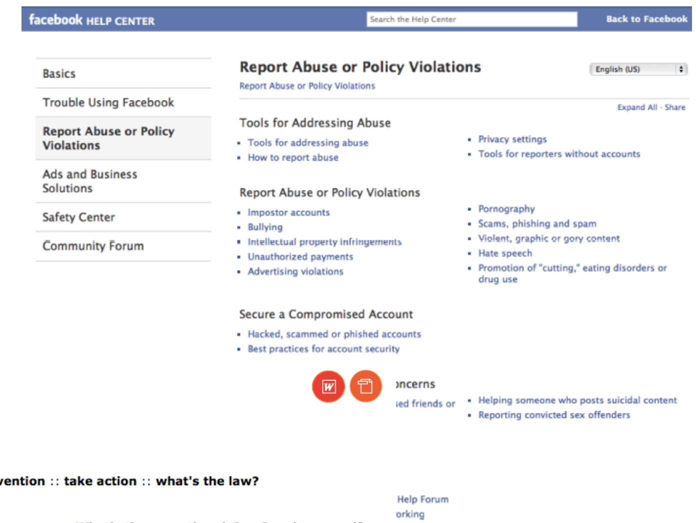


Common Sense Reasoning for Detection, Prevention, and Mitigation of Cyberbullying

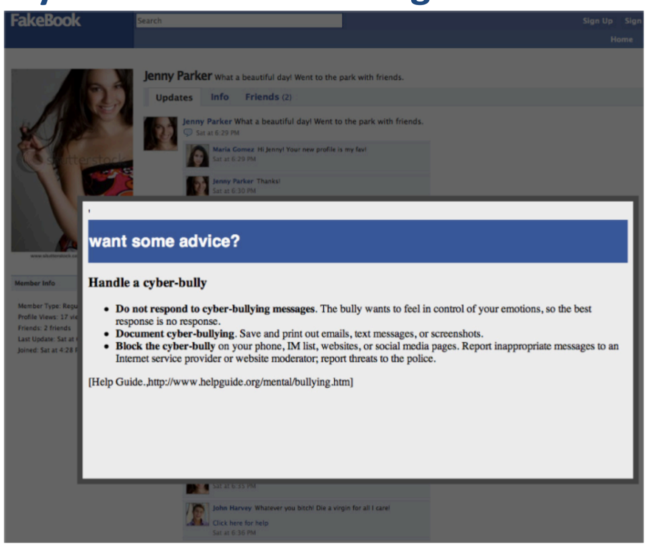
[Dinakar2012]

- Evaluation of suggesting educational materials:
  - Small study with <u>five participants</u> on fully functional hypothetical social network (Fakebook)
  - Test differences between 3 scenarios

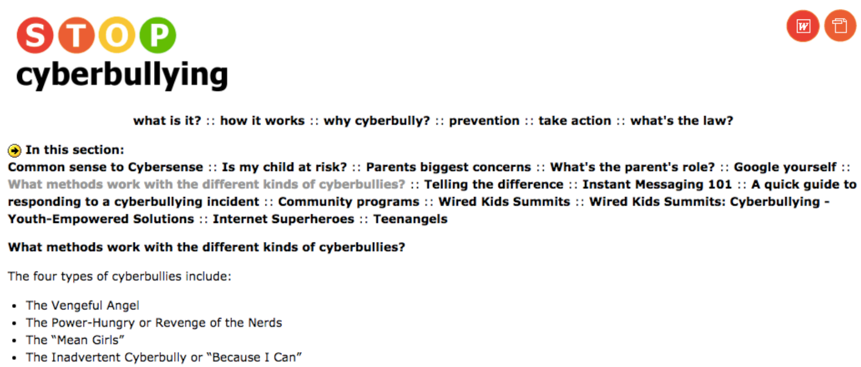
Typical "help" link user interaction



Dynamic in-context targeted advice



Targeted static advice



Common Sense Reasoning for Detection, Prevention, and Mitigation of Cyberbullying

[Dinakar2012]

- Each participant took a survey after reading fictional cyberbullying incident, imagining themselves as one of the characters and clicking on the links for help
- Participants **preferred the interface with targeted in-context advice**

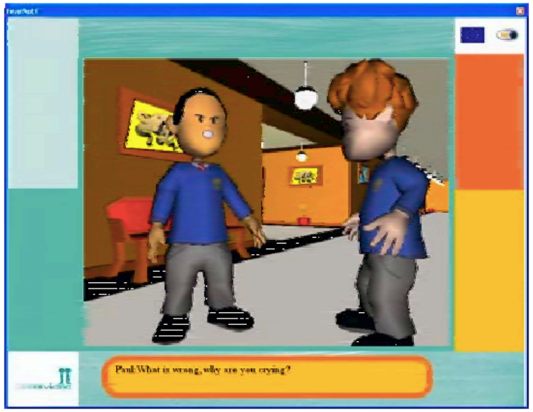
Interface 1: In-Context Dynamic Targeted Help					
	Strongly Disagree	Disagree	Neither Agree nor Disagree	Agree	Strongly Agree
Imagine you are Jenny. Assuming Jenny is the victim, when I clicked on the advice links I considered the advice helpful in the situation.	0%	0%	0%	20%	**80%**
Imagine you are John. Assuming John is the bully, when I clicked on the help links, I felt reflective about my behavior and how it might have affected Jenny.	0%	20%	0%	**40%**	**40%**
Imagine you are Maria. Assuming the Maria is a bystander, when I clicked on the links, I reflected on how the messages might have affected Jenny.	0%	0%	0%	20%	**80%**

FearNot! Demo - A Virtual Environment with Synthetic Characters to help Bullying

[Vala2012]

- **Goal**: teach 8–12 years old children coping strategies in bullying situations based on synthetic characters on virtual learning environments

- Interactive storytelling with animated on-screen characters
  - User gets to play one of the participants in the bullying scenario
  - User may select any one of a number of response strategies to a bullying challenge (e.g., fight back, run away, tell a teacher)

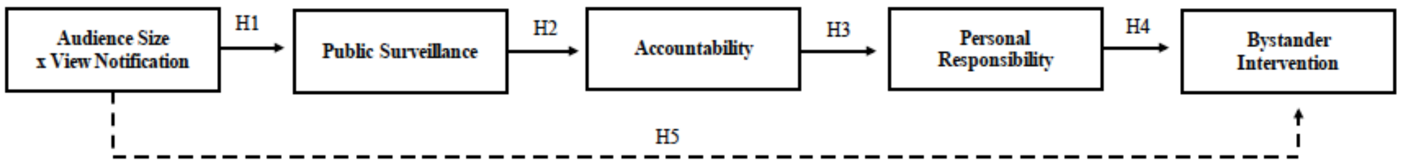


Bullying situation in FearNot!

Upstanding by Design: Bystander Intervention in Cyberbullying

[DiFranzo2018]

- **Goal**: explore effects of interface design on bystander intervention through simulated custom-made social media platform
  - Understand bystander behavior in cyberbullying
  - Design and implement interfaces aimed at encouraging bystander intervention based on bystander intervention model [Darley1968]
- If bystanders feel personally responsible, they tend to intervene
  - Interface designs that **heighten self-awareness via public surveillance** should indirectly increase cyberbystander intervention
- Two design interventions:
  - "You have already viewed this message" notification
  - Information about audience size ("this many people have seen this message")

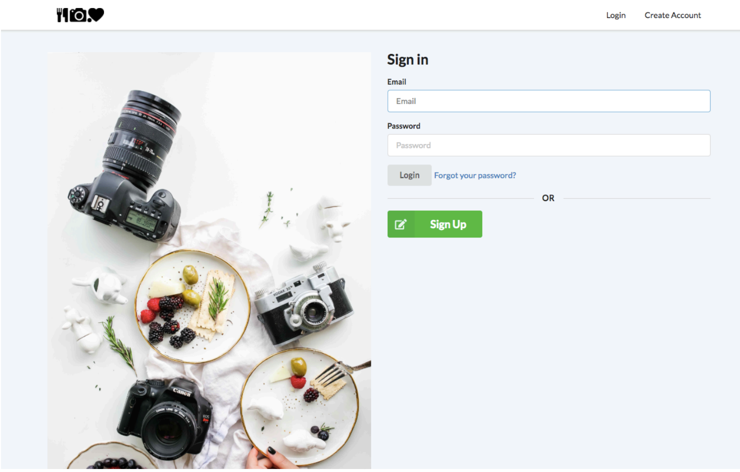


Conceptual Model of Bystander Intervention

Upstanding by Design: Bystander Intervention in Cyberbullying

[DiFranzo2018]

- Approach:
  - Developed EatSnap.Love social networking site (share, like, react to food pictures)
  - Created platform to control social interactions between users
    - Each participant was exposed to same social interactions, users, posts, and responses within controlled environment
    - Participants did not interact with each other, but with bots



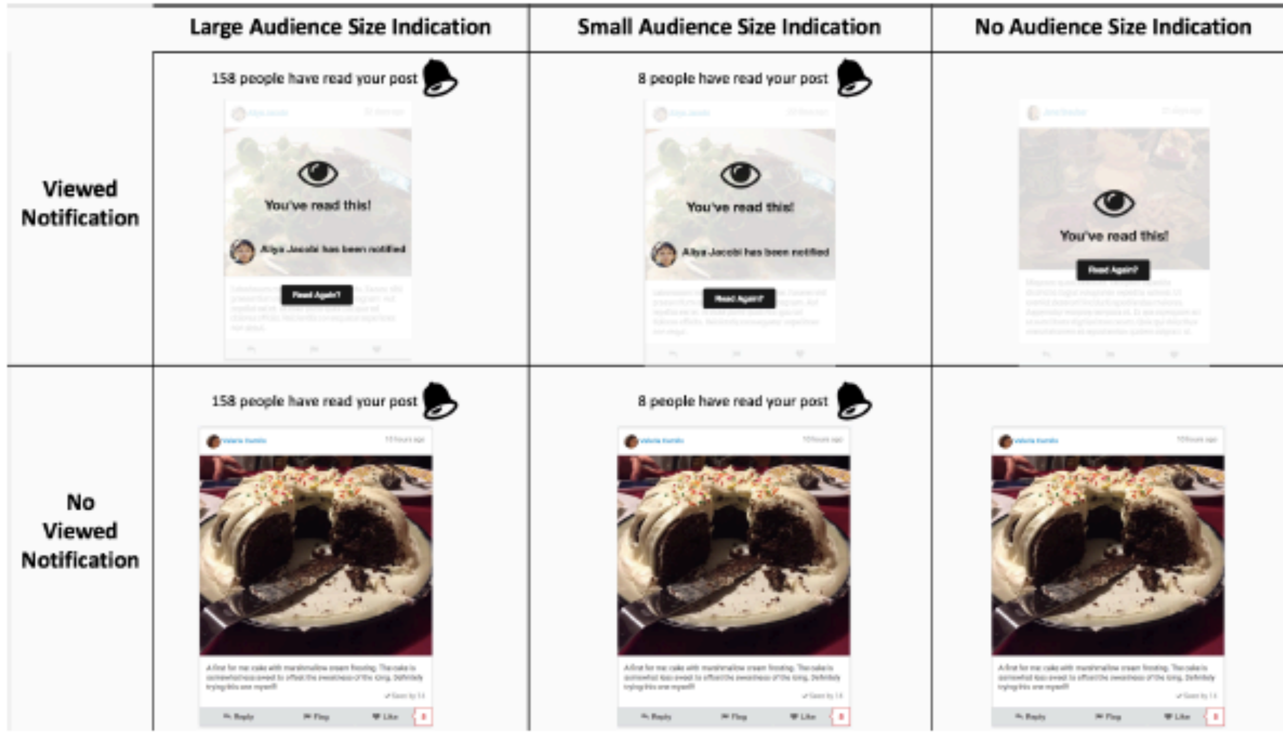
EatSnap.Love social networking site

  - 400 participants from Amazon Mechanical Turk (attrition rate: 41%)
    - Participants were exposed to several cyberbullying incidents during 3 days
    - Participants received different information about audience size and viewing notifications

Upstanding by Design: Bystander Intervention in Cyberbullying

[DiFranzo2018]

Design intervention scenarios

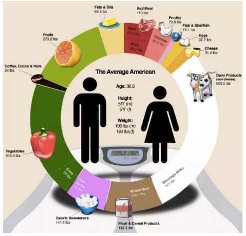


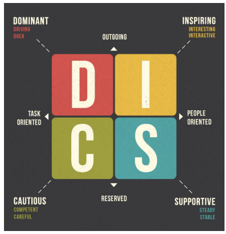
– Participants were provided
  - Community guidelines governing the site
  - What to do if they witnessed someone breaking those rules

Upstanding by Design: Bystander Intervention in Cyberbullying

[DiFranzo2018]

- Pre-study survey:
  - Demographics, personality measures, and filler questions
  - General food consumption patterns





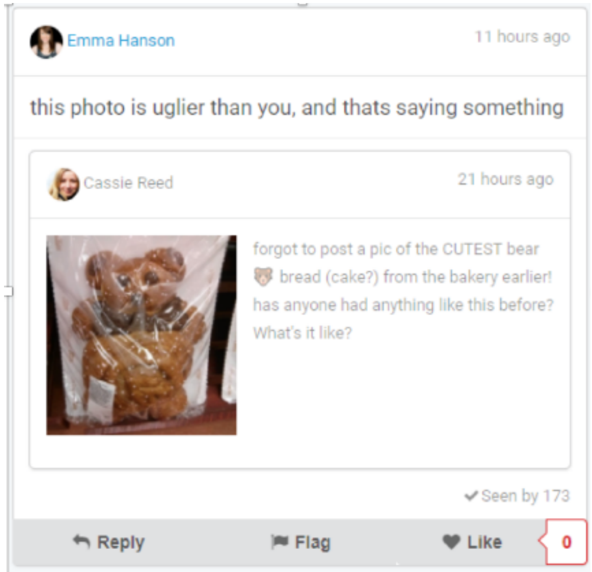
- During study:
  - Post a photo and message at least once per day during 3-day period
  - Read posts
  - Interact with posts

- Post-study survey:
  - Reflect on experience using the site
  - Whether cyberbullying incidents were observed

Upstanding by Design: Bystander Intervention in Cyberbullying

[DiFranzo2018]

- Each participant was exposed to 4 cyberbullying instances
- Measures:
  - Bystander intervention (direct or indirect)
  - Public Surveillance (7-point agree/disagree scale)
    - "Users of EatSnap.Love are aware that I viewed their posts"
    - "The other people using EatSnap.Love know when I see their posts and replies"
  - Accountability (7-point agree/disagree scale)
    - "I was held accountable for my behavior on EatSnap.Love"
    - "I would have to answer to others if I acted inappropriately on EatSnap.Love"
  - Personal Responsibility (7-point agree/disagree scale)
    - "Helping other users of EatSnap.Love who are teased or left out was my responsibility"



Cyberbullying instance example

Upstanding by Design: Bystander Intervention in Cyberbullying

[DiFranzo2018]

– Observations:
  - 74.5% of the cyberbullying bystanders did not intervene in any form
  - Indirect interventions were more common than direct ones
  - 96% of interventions involved flagging the cyberbullying post
  - < 3% blocked or notified administrator
  - Participants who felt greater accountability also tended to report more personal responsibility for cyberbullying behaviors and ended up flagging the content
  - Small audience increases likelihood of bystander intervention

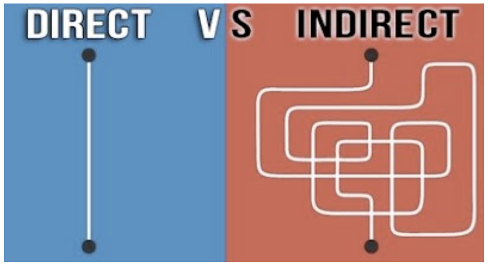
Condition (vs. control)	Serial Mediators	Outcome	
		Effect (*SE*)	95% CI
Low Bystander, No View	(direct effect)	.79 (.65)	[-.54, 1.07]
	→ public surveillance → accountability →	-.09 (.08)	[-.30, -.004]
	→ public surveillance → accountability → responsibility →	.08 (.05)	[.01, .22]
Low Bystander, View	(direct effect)	1.09 (.73)	[-.34, 2.52]
	→ public surveillance → accountability →	-.15 (.12)	[-.46, -.01]
	→ public surveillance → accountability → responsibility →	.13 (.08)	[.03, .33]
High Bystander, View	(direct effect)	.48 (.76)	[-1.0, 1.95]
	→ public surveillance → accountability →	-.14 (.11)	[-.41, -.01]
	→ public surveillance → accountability → responsibility →	.12 (.07)	[.03, .29]

Note. This table reports only mediation models tested with confidence intervals that did not include zero.

Analysis: most probable paths to intervention

Evaluation of Mitigation Tools

- **No evaluation** in most cases
  - e.g., [Vala2012], [Cohen2014], [Ashktorab2016], [Vishwamitra2017], [Fan2016]
- Indirect evaluation (e.g., [Dinakar2012])
  - Hypothesis that strategy will work based on insights drawn from the literature such as psychology [Walther2005], criminology [Madlock2011]
- Qualitative evaluation
  - Pre/post **surveys** (e.g., [Dinakar2012], [Ashktorab2017], [DiFranzo2018, [Kazerooni2018])
  - **Focus groups** (e.g., [Bowler2014], [vanderZwaan2013]) on artificially constructed scenarios
- Quantitative/Direct evaluation is hard!



Section

Interactive Session

Divide into Groups of 3-5 people

- Imagine you are a research group that wants to study bullying on two online social media
- You have access to:

Twitter Dataset: a sample of 20 tweets

- Your task is to label each <u>**tweet**</u> as **normal**, **spam**, **hateful**, or **abusive**

Instagram Dataset: You are provided 4 sample Instagram media sessions

- Your task is to label each <u>**session**</u> as **normal**, **abusive** or **bullying**

URL http://www.cs.albany.edu/~cchelmis/icwsm2018tutorial/interactivesessionmatterials.zip

- Attempt the tasks <u>individually</u> first
- Once each member of your group is done, <u>aggregate</u> your annotations
  - Try to reach consensus on as many items (i.e., tweets and sessions) as possible
- Chose a representative to briefly explain contention points (if any)

- Let me know if you have any questions/issues/concerns!

Twitter Dataset (~5 mins)

- Mark a **<u>tweet</u>** (i.e., single post) as follows:
  - **Abusive**: Strongly impolite, rude or hurtful language using profanity
  - **Hateful**: Hatred, or derogatory, insulting, humiliating statements towards an individual or members of the group, on the basis of attributes such as race, disability, or gender
  - **Spam**: Advertising/marketing, linking to malicious websites, unwanted information
  - **Normal**: None of the above

- Definitions are from [Founta2018]
- A dictionary of profane words is not given

Sample tweets [Founta2018]

Alex Brosas another idiot
#ALDUBKSGoesToUS

Mama_Dub🇨🇦™💚 @star_58
Paid journalist..WHAT A WASTE OF PAPER... CHECK YOUR FACTS BAKLANG BROSAS. G NA G KA NA TAKAGA KAYA KAHIT WALANG KWENTA IMBENTO MO!!!
twitter.com/cindyharvard/s...

Niggas keep talking about women wearing weave but be sick when a bitch up a fro on they ass. 😭

#Insulin a key molecule for health, evidence also shows side effects (e.g. #inflammation): take an #InsulinHoliday:

Insulin Holiday – Allen Tien – Medium
What is an 'insulin holiday'? It is a period of time with low or very low insulin levels. How does one take an 'insulin holiday'? One way...
medium.com

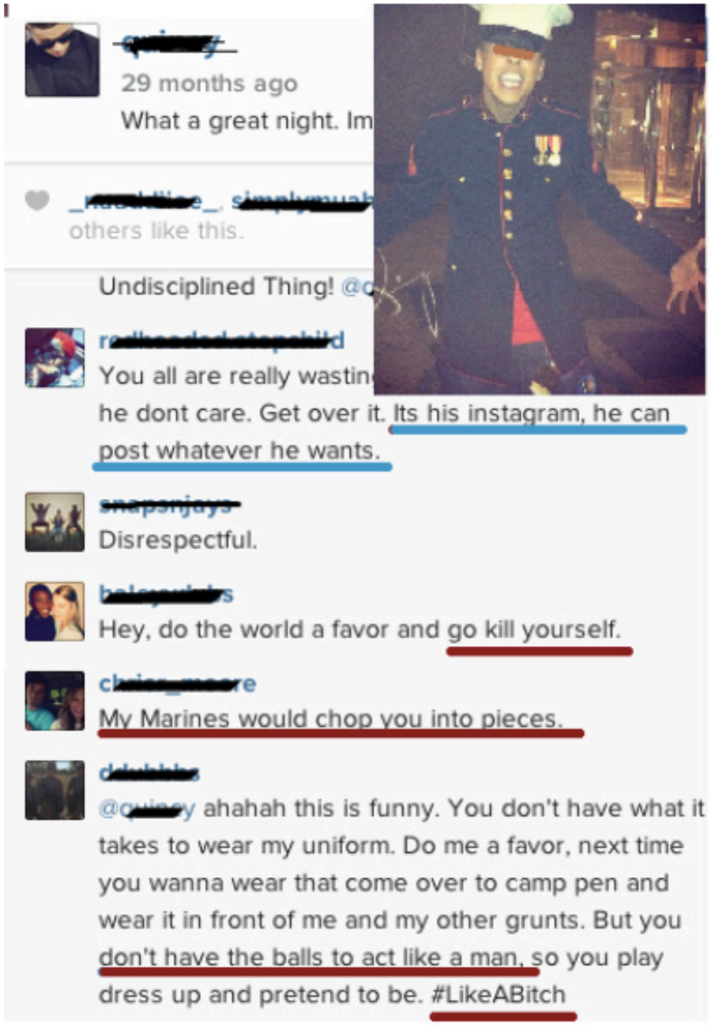
The Nazi death gas so horrific even Hitler feared using it

Instagram Dataset (~10 mins)

- Mark a **session** (i.e., collection of comments) as an instance of:
  - Cyberaggression if there is at least one negative word/comment and or content with intent to harm someone
  - Cyberbullying if two (2) or more comments include negative words/content with intent to harm someone
  - **Normal**: None of the above

- Definitions are from [Hosseinmardi2015]
- Images and user profiles are not provided
  - Labeling associated comments may be harder
- A dictionary of profane words is not given

Sample media session [Hosseinmardi2015]



Discussion (~5mins)

Tweet Id	Normal	Abusive	Hateful	Spam	True Label
1					A
2					A
3					A
4					A
5					A
6					N
7					N
8					N
9					N
10					N
11					H
12					H
13					H
14					H
15					H

DISCLAIMER Labels are from [Founta2018]

Discussion (~5mins)

Session Id	Normal	Aggressive	Bullying	True Label
1				N
2				B
3				B
4				A
5				A

- Use workbook_group_agreement.xlx to measure consensus between your group members
  - Use 0 for Normal, 1 for Bullying and 2 for Aggressive

Important Note

In practice interrater agreement is measured using statistical measures such as Cohen's kappa [James1984, McHugh2012]

DISCLAIMER

Labels are from [Hosseinmardi2015]

Discussion (~10mins)

- What was the main difficulty when going through the tasks?
- How easy was it to distinguish between different categories?
  - e.g., hate speech vs. abusive language
- What would be the implications of possible annotation mistakes?
  - What metrics/inferences are they likely to impact the most?
- Can you imagine scaling the multi-labeled annotation process to thousand comments (tweets, posts, …)?
  - What would be the issues?
- Think about the implications of trying to sample/analyze data from certain online social networking platforms
  - Bias?
    - Keyword-based sampling?
    - Occurrence rates for different categories introduced by the sample
  - Anonymity?
  - User population demographics ?

Do these influence discovered patterns?

Issues with Annotation

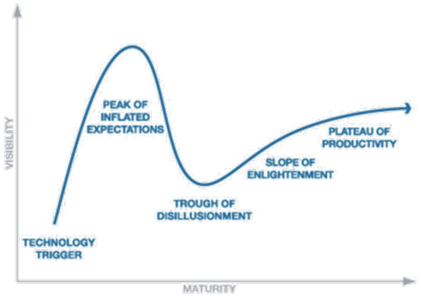
Source: Zeerak Waseem's 2018 Turing Institute presentation

Section

Summary and Concluding Remarks

Summary and Conclusions

- Characterizing, detecting (or predicting) and mitigating cyberbullying instances is a **hard problem**!
  - Very **active** research area
    - Still in an incipient phase of the hype cycle!
  - We have identified more than a dozen **challenges**



- Fascinating field at the **intersection** of many disciplines
  - Psychology and Sociology
  - (Computational) Social Science
  - Computer Science
  - Electrical Engineering
  - …

- Overall, cyberbullying is a function of a complex social system
  - Notions of bullying behavior and the use of technology coevolve

Tutorial Slides

DISCLAIMER

We recognize that our coverage of the state-of-the-art and the challenges we identify are not exhaustive

- Some important topics we did not cover include (but are not limited to)
  - Expanding cyberbullying detection beyond bullies and victims
  - Determining victim's emotional state after cyberbullying
- References are provided for additional reading

The slides can be found at:
http://www.cs.albany.edu/~cchelmis/icwsm2018tutorial/

Suggested citation:

Charalampos Chelmis, Daphney–Stavroula Zois, Characterization, Detection, and Mitigation of Cyberbullying, Tutorial at the 12th International Conference on Web and Social Media, Stanford, CA, June 2018.

References

[Darley1968] Darley, John M., and Bibb Latane. "Bystander intervention in emergencies: Diffusion of responsibility." *Journal of personality and social psychology* 8, no. 4 (1968): 377.

[Granovetter1976] Granovetter, Mark. "Network sampling: Some first steps." American journal of sociology 81, no. 6 (1976): 1287-1303.

[Olweus1978] Olweus, Dan. "Aggression in the schools: Bullies and whipping boys." Hemisphere, 1978.

[Biernacki1981] Biernacki, Patrick, and Dan Waldorf. "Snowball sampling: Problems and techniques of chain referral sampling." Sociological methods & research 10, no. 2 (1981): 141-163.

[James1984] James, Lawrence R., Robert G. Demaree, and Gerrit Wolf. "Estimating within-group interrater reliability with and without response bias." Journal of applied psychology 69, no. 1 (1984): 85.

[Olweus1994] Olweus, Dan. "Bullying at school." In Aggressive behavior, pp. 97-130. Springer, Boston, MA, 1994.

[Greenwood1996] Greenwood, Priscilla E., and Michael S. Nikulin. A guide to chi-squared testing. Vol. 280. John Wiley & Sons, 1996.

[Yang1997] Yang, Yiming, and Jan O. Pedersen. "A comparative study on feature selection in text categorization." In Icml, vol. 97, pp. 412-420. 1997.

[Salmivalli1999] Salmivalli, Christina. "Participant role approach to school bullying: Implications for interventions." *Journal of adolescence* 22, no. 4 (1999): 453-459.

[Hołyst2000] Hołyst, Janusz A., Krzysztof Kacperski, and Frank Schweitzer. "Phase transitions in social impact models of opinion formation." Physica A: Statistical Mechanics and its Applications 285, no. 1-2 (2000): 199-210.

[Atkinson2001] Atkinson, Rowland, and John Flint. "Accessing hidden and hard-to-reach populations: Snowball research strategies." Social research update 33, no. 1 (2001): 1-4.

References (2)

[Chawla2002] Chawla, Nitesh V., Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. "SMOTE: synthetic minority over-sampling technique." Journal of artificial intelligence research16 (2002): 321-357.

[Guyon2003] Guyon, Isabelle, and André Elisseeff. "An introduction to variable and feature selection." Journal of machine learning research 3, no. Mar (2003): 1157-1182.

[Batista2004] Batista, Gustavo EAPA, Ronaldo C. Prati, and Maria Carolina Monard. "A study of the behavior of several methods for balancing machine learning training data." ACM SIGKDD explorations newsletter 6, no. 1 (2004): 20-29.

[Bocij2004] Bocij, Paul. Cyberstalking: Harassment in the Internet age and how to protect your family. Greenwood Publishing Group, 2004.

[Rigby2004] Rigby, Ken. "Addressing bullying in schools: Theoretical perspectives and their implications." School Psychology International 25, no. 3 (2004): 287-300.

[Campbell2005] Campbell, Marilyn A. "Cyber bullying: An old problem in a new guise?." Journal of Psychologists and Counsellors in Schools15, no. 1 (2005): 68-76.

[Han2005] Han, Hui, Wen-Yuan Wang, and Bing-Huan Mao. "Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning." In International Conference on Intelligent Computing, pp. 878-887. Springer, Berlin, Heidelberg, 2005.

[Li2005] Li, Tanya Beran Qing. "Cyber-harassment: A study of a new method for an old behavior." Journal of educational computing research 32, no. 3 (2005): 265-277.

[Peng2005] Peng, Hanchuan, Fuhui Long, and Chris Ding. "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy." IEEE Transactions on pattern analysis and machine intelligence 27, no. 8 (2005): 1226-1238.

References (3)

[Walther2005] Walther, Joseph B., Tracy Loh, and Laura Granka. "Let me count the ways: The interchange of verbal and nonverbal cues in computer-mediated and face-to-face affinity." *Journal of language and social psychology* 24, no. 1 (2005): 36-65.

[Boivin2006] Boivin, Jean, and Serena Ng. "Are more data always better for factor analysis?." Journal of Econometrics 132, no. 1 (2006): 169-194.

[Davis2006] Davis, Jesse, and Mark Goadrich. "The relationship between Precision-Recall and ROC curves." In Proceedings of the 23rd international conference on Machine learning, pp. 233-240. ACM, 2006.

[Fawcett2006] Fawcett, Tom. "An introduction to ROC analysis." Pattern recognition letters 27, no. 8 (2006): 861-874.

[Pittaro2007] Pittaro, Michael L. "Cyber stalking: An analysis of online harassment and intimidation." International Journal of Cyber Criminology 1, no. 2 (2007): 180-197.

[Saeys2007] Saeys, Yvan, Iñaki Inza, and Pedro Larrañaga. "A review of feature selection techniques in bioinformatics." bioinformatics 23, no. 19 (2007): 2507-2517.

[Sheridan2007] Sheridan, Lorraine P., and Tim Grant. "Is cyberstalking different?." Psychology, crime & law 13, no. 6 (2007): 627-640.

[Vala2007] Vala, Marco, Pedro Sequeira, Ana Paiva, and Ruth Aylett. "FearNot! demo: a virtual environment with synthetic characters to help bullying." In *Proceedings of the 6th international joint conference on Autonomous agents and multiagent systems*. ACM, 2007.

[Slonje2008] Slonje, Robert, and Peter K. Smith. "Cyberbullying: Another main type of bullying?." Scandinavian journal of psychology49, no. 2 (2008): 147-154.

[Vandebosch2008] Vandebosch, Heidi, and Katrien Van Cleemput. "Defining cyberbullying: A qualitative research into the perceptions of youngsters." CyberPsychology & Behavior 11, no. 4 (2008): 499-503.

References (4)

[Benevenuto2009] Benevenuto, Fabrício, Tiago Rodrigues, Meeyoung Cha, and Virgílio Almeida. "Characterizing user behavior in online social networks." In Proceedings of the 9th ACM SIGCOMM Conference on Internet Measurement, pp. 49-62. ACM, 2009.

[Bunkhumpornpat2009] Bunkhumpornpat, Chumphol, Krung Sinapiromsaran, and Chidchanok Lursinsap. "Safe-level-smote: Safe-level-synthetic minority over-sampling technique for handling the class imbalanced problem." In Pacific-Asia conference on knowledge discovery and data mining, pp. 475-482. Springer, Berlin, Heidelberg, 2009.

[Dooley2009] Dooley, Julian J., Jacek Pyżalski, and Donna Cross. "Cyberbullying versus face-to-face bullying: A theoretical and conceptual review." Zeitschrift für Psychologie/Journal of Psychology 217, no. 4 (2009): 182-188.

[Kraft2009] Kraft, Ellen M., and Jinchang Wang. "Effectiveness of cyber bullying prevention strategies: A study on students' perspectives." *International Journal of Cyber Criminology* 3, no. 2 (2009): 513.

[Yin2009] Yin, Dawei, Zhenzhen Xue, Liangjie Hong, Brian D. Davison, April Kontostathis, and Lynne Edwards. "Detection of harassment on web 2.0." Proceedings of the Content Analysis in the WEB 2 (2009): 1-7.

[Erdur-Baker2010] Erdur-Baker, Özgür. "Cyberbullying and its correlation to traditional bullying, gender and frequent and risky usage of internet-mediated communication tools." New media & society12, no. 1 (2010): 109-125.

[Grigg2010] Grigg, Dorothy Wunmi. "Cyber-aggression: Definition and concept of cyberbullying." Journal of Psychologists and Counsellors in Schools 20, no. 2 (2010): 143-156.

[Nadali2010] Nadali, Samaneh, Masrah Azrifah Azmi Murad, Nurfadhlina Mohamad Sharef, Aida Mustapha, and Somayeh Shojaee. "A review of cyberbullying detection: An overview." In Intelligent Systems Design and Applications (ISDA), 2013 13th International Conference on, pp. 325-330. IEEE, 2013.

References (5)

[Tokunaga2010] Tokunaga, Robert S. "Following you home from school: A critical review and synthesis of research on cyberbullying victimization." Computers in human behavior 26, no. 3 (2010): 277-287.

[Bayzick2011] Bayzick, Jennifer, April Kontostathis, and Lynne Edwards. "Detecting the presence of cyberbullying using computer software." In 3rd Annual ACM Web Science Conference (WebSci '11), pp. 1-2. 2011.

[Biel2011] Biel, Joan-Isaac, Oya Aran, and Daniel Gatica-Perez. "You Are Known by How You Vlog: Personality Impressions and Nonverbal Behavior in YouTube." In ICWSM. 2011.

[Harabagiu2011] Harabagiu, Sanda, and Andrew Hickl. "Relevance Modeling for Microblog Summarization." In Fifth International AAAI Conference on Weblogs and Social Media. 2011.

[Madlock2011] Madlock, Paul E., and David Westerman. "Hurtful Cyber-Teasing and Violence: Who's Laughing Out Loud?." *Journal of interpersonal violence* 26, no. 17 (2011): 3542-3560.

[Powers2011] Powers, David Martin. "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation." (2011).

[Reyns2011] Reyns, Bradford W., Billy Henson, and Bonnie S. Fisher. "Being pursued online: Applying cyberlifestyle–routine activities theory to cyberstalking victimization." Criminal justice and behavior 38, no. 11 (2011): 1149-1169.

[Ahmed2012] Ahmed, Nesreen K., Jennifer Neville, and Ramana Rao Kompella. "Network Sampling Designs for Relational Classification." In ICWSM. 2012.

[Dinakar2012] Dinakar, Karthik, Birago Jones, Catherine Havasi, Henry Lieberman, and Rosalind Picard. "Common sense reasoning for detection, prevention, and mitigation of cyberbullying." ACM Transactions on Interactive Intelligent Systems (TiiS) 2, no. 3 (2012): 18.

References (6)

[Espelage2012] Espelage, Dorothy L., Mrinalini A. Rao, and Rhonda G. Craven. "Theories of cyberbullying." Principles of cyberbullying research: Definitions, measures, and methodology (2012): 49-67.

[Kowalski2012] Kowalski, Robin M., Susan P. Limber, Sue Limber, and Patricia W. Agatston. Cyberbullying: Bullying in the digital age. John Wiley & Sons, 2012.

[McCreadie2012] McCreadie, Richard, Ian Soboroff, Jimmy Lin, Craig Macdonald, Iadh Ounis, and Dean McCullough. "On building a reusable Twitter corpus." In Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval, pp. 1113-1114. ACM, 2012.

[McHugh2012] McHugh, Mary L. "Interrater reliability: the kappa statistic." Biochemia medica: Biochemia medica 22, no. 3 (2012): 276-282.

[Mishna2012] Mishna, Faye, Mona Khoury-Kassabri, Tahany Gadalla, and Joanne Daciuk. "Risk factors for involvement in cyber bullying: Victims, bullies and bully–victims." Children and Youth Services Review 34, no. 1 (2012): 63-70.

[Navarro2012] Navarro, Jordana N., and Jana L. Jasinski. "Going cyber: Using routine activities theory to predict cyberbullying experiences." Sociological Spectrum 32, no. 1 (2012): 81-94.

[Smith2012] Smith, Peter K. "Cyberbullying and cyber aggression." In Handbook of school violence and school safety, pp. 111-121. Routledge, 2012.

[Volk2012] Volk, Anthony A., Joseph A. Camilleri, Andrew V. Dane, and Zopito A. Marini. "Is adolescent bullying an evolutionary adaptation?." Aggressive behavior 38, no. 3 (2012): 222-238.

[Xu2012] Xu, Jun-Ming, Kwang-Sung Jun, Xiaojin Zhu, and Amy Bellmore. "Learning from bullying traces in social media." In Proceedings of the 2012 conference of the North American chapter of the association for computational linguistics: Human language technologies, pp. 656-666. Association for Computational Linguistics, 2012.

References (7)

[AlMazari2013] Al Mazari, Ali. "Cyber-bullying taxonomies: Definition, forms, consequences and mitigation strategies." In *Computer Science and Information Technology (CSIT), 2013 5th International Conference on*, pp. 126-133. IEEE, 2013.

[Blei2013] Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." Journal of machine Learning research 3, no. Jan (2003): 993-1022.

[Dadvar2013] Dadvar, Maral, Dolf Trieschnigg, Roeland Ordelman, and Franciska de Jong. "Improving cyberbullying detection with user context." In European Conference on Information Retrieval, pp. 693-696. Springer, Berlin, Heidelberg, 2013.

[Jones2013] Jones, Lisa M., Kimberly J. Mitchell, and David Finkelhor. "Online harassment in context: Trends from three youth internet safety surveys (2000, 2005, 2010)." Psychology of violence 3, no. 1 (2013): 53.

[Kontostathis2013] Kontostathis, April, Kelly Reynolds, Andy Garron, and Lynne Edwards. "Detecting cyberbullying: query terms and techniques." In Proceedings of the 5th annual acm web science conference, pp. 195-204. ACM, 2013.

[Mikolov2013] Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. "Distributed representations of words and phrases and their compositionality." In Advances in neural information processing systems, pp. 3111-3119. 2013.

[Morstatter2013] Morstatter, Fred, Jürgen Pfeffer, Huan Liu, and Kathleen M. Carley. "Is the Sample Good Enough? Comparing Data from Twitter's Streaming API with Twitter's Firehose." In ICWSM. 2013.

[VanderZwaan2013] van der Zwaan, Janneke M., and Virginia Dignum. "Robin, an empathic virtual buddy for social support." In *Proceedings of the 2013 international conference on Autonomous agents and multi-agent systems*, pp. 1413-1414. International Foundation for Autonomous Agents and Multiagent Systems, 2013.

References (8)

[Ahmed2014] Ahmed, Nesreen K., Nick Duffield, Jennifer Neville, and Ramana Kompella. "Graph sample and hold: A framework for big-graph analytics." In Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 1446-1455. ACM, 2014.

[Bourigault2014] Bourigault, Simon, Cedric Lagnier, Sylvain Lamprier, Ludovic Denoyer, and Patrick Gallinari. "Learning social network embeddings for predicting information diffusion." In Proceedings of the 7th ACM international conference on Web search and data mining, pp. 393-402. ACM, 2014.

[Bowler2014] Bowler, Leanne, Eleanor Mattern, and Cory Knobel. "Developing design interventions for cyberbullying: A narrative-based participatory approach." *iConference 2014 Proceedings* (2014).

[Cohen2014] Cohen, Robin, Disney Yan Lam, Nikhil Agarwal, Michael Cormier, Jasmeet Jagdev, Tianqi Jin, Madhur Kukreti et al. "Using computer technology to address the problem of cyberbullying." *ACM SIGCAS Computers and Society* 44, no. 2 (2014): 52-61.

[Dalessandro2014] Dalessandro, Brian, Claudia Perlich, and Troy Raeder. "Bigger is better, but at what cost? Estimating the economic value of incremental data assets." Big data 2, no. 2 (2014): 87-96.

[Frénay2014] Frénay, Benoît, and Michel Verleysen. "Classification in the presence of label noise: a survey." IEEE transactions on neural networks and learning systems 25, no. 5 (2014): 845-869.

[González-Bailón2014] González-Bailón, Sandra, Ning Wang, Alejandro Rivero, Javier Borge-Holthoefer, and Yamir Moreno. "Assessing the bias in samples of large online networks." Social Networks 38 (2014): 16-27.

[Hosseinmardi2014] Li, Homa Hosseinmardi Shaosong, Zhili Yang, Qin Lv, Rahat Ibn Rafiq Richard Han, and Shivakant Mishra. "A comparison of common users across instagram and ask. fm to better understand cyberbullying." In Big Data and Cloud Computing (BdCloud), 2014 IEEE Fourth International Conference on, pp. 355-362. IEEE, 2014.

References (9)

[Hosseinmardi2014corr] Hosseinmardi, Homa, Richard Han, Qin Lv, Shivakant Mishra, and Amir Ghasemianlangroodi. "Analyzing negative user behavior in a semi-anonymous social network." CoRR abs1404 (2014).

[Liu2014] Liu, Yabing, Chloe Kliman-Silver, and Alan Mislove. "The Tweets They Are a-Changin: Evolution of Twitter Users and Behavior." ICWSM 30 (2014): 5-314.

[Olteanu2014] Olteanu, Alexandra, Carlos Castillo, Fernando Diaz, and Sarah Vieweg. "CrisisLex: A Lexicon for Collecting and Filtering Microblogged Communications in Crises." In ICWSM. 2014.

[Potha2014] Potha, Nektaria, and Manolis Maragoudakis. "Cyberbullying detection using time series modeling." In *Data Mining Workshop (ICDMW), 2014 IEEE International Conference on*, pp. 373-382. IEEE, 2014.

[Tsytsarau2014] Tsytsarau, Mikalai, Themis Palpanas, and Malu Castellanos. "Dynamics of news events and social media reaction." In Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 901-910. ACM, 2014.

[Tufekci2014] Tufekci, Zeynep. "Big Questions for Social Media Big Data: Representativeness, Validity and Other Methodological Pitfalls." ICWSM 14 (2014): 505-514.

[White2014] White, Molly. "The challenges of building a compassionate robot." http://blog.mollywhite.net/the-challenges-ofbuilding-a-compassionate-robot, November 2014.

[Balakrishnan2015] Balakrishnan, Vimala. "Cyberbullying among young adults in Malaysia: The roles of gender, age and Internet frequency." Computers in Human Behavior 46 (2015): 149-157.

[Bellmore2015] Bellmore, Amy, Angela J. Calvin, Jun-Ming Xu, and Xiaojin Zhu. "The five w's of "bullying" on twitter: who, what, why, where, and when." Computers in human behavior 44 (2015): 305-314.

References (10)

[Corcoran2015] Corcoran, Lucie, Conor Mc Guckin, and Garry Prentice. "Cyberbullying or cyber aggression?: A review of existing definitions of cyber-based peer-to-peer aggression." Societies5, no. 2 (2015): 245-255.

[Farajtabar2015] Farajtabar, Mehrdad, Yichen Wang, Manuel Gomez Rodriguez, Shuang Li, Hongyuan Zha, and Le Song. "Coevolve: A joint point process model for information diffusion and network co-evolution." In Advances in Neural Information Processing Systems, pp. 1954-1962. 2015.

[Hosseinmardi2015] Hosseinmardi, Homa, Sabrina Arredondo Mattson, Rahat Ibn Rafiq, Richard Han, Qin Lv, and Shivakant Mishra. "Analyzing labeled cyberbullying incidents on the instagram social network." In International Conference on Social Informatics, pp. 49-66. Springer, Cham, 2015.

[Lempa2015] Lempa, Pawel, Michal Ptaszynski, and Fumito Masui. "Cyberbullying Blocker Application for Android." In *7th Language & Technology Conference (LTC'15), Poznan, Poland*. 2015.

[Rafiq2015] Rafiq, Rahat Ibn, Homa Hosseinmardi, Richard Han, Qin Lv, Shivakant Mishra, and Sabrina Arredondo Mattson. "Careful what you share in six seconds: Detecting cyberbullying instances in Vine." In Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015, pp. 617-622. ACM, 2015.

[Shaul2015] Shaul, Brandy. “Honestly Looks to Combat Cyberbullying on iOS, Android.” http://www.adweek.com/socialtimes/honestly-looksto-combat-cyberbullying-on-ios-android/615873, 2015.

[Al-garadi2016] Al-garadi, Mohammed Ali, Kasturi Dewi Varathan, and Sri Devi Ravana. "Cybercrime detection in online communications: The experimental case of cyberbullying detection in the Twitter network." Computers in Human Behavior 63 (2016): 433-443.

[Ashktorab2016] Ashktorab, Zahra, and Jessica Vitak. "Designing cyberbullying mitigation and prevention solutions through participatory design with teenagers." In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pp. 3895-3905. ACM, 2016.

References (11)

[Edwards2016] Edwards, Lynne, April Edwards Kontostathis, and Christina Fisher. "Cyberbullying, race/ethnicity and mental health outcomes: A review of the literature." Media and Communication 4, no. 3 (2016): 71-78.

[Fan2016] Fan, Mingyue, Liyue Yu, and Leanne Bowler. "Feelbook: A social media app for teens designed to foster positive online behavior and prevent cyberbullying." In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, pp. 1187-1192. ACM, 2016.

[Hosseinmardi2016] Hosseinmardi, Homa, Rahat Ibn Rafiq, Richard Han, Qin Lv, and Shivakant Mishra. "Prediction of cyberbullying incidents in a media-based social network." In Advances in Social Networks Analysis and Mining (ASONAM), 2016 IEEE/ACM International Conference on, pp. 186-192. IEEE, 2016.

[Liu2016] Liu, Leqi, Daniel Preotiuc-Pietro, Zahra Riahi Samani, Mohsen Ebrahimi Moghaddam, and Lyle H. Ungar. "Analyzing Personality through Social Media Profile Picture Choice." In ICWSM, pp. 211-220. 2016.

[Singh2016] Singh, Vivek K., Qianjia Huang, and Pradeep K. Atrey. "Cyberbullying detection using probabilistic socio-textual information fusion." In Advances in Social Networks Analysis and Mining (ASONAM), 2016 IEEE/ACM International Conference on, pp. 884-887. IEEE, 2016.

[Squicciarini2016] Squicciarini, A., S. Rajtmajer, Y. Liu, and Christopher Griffin. "Identification and characterization of cyberbullying dynamics in an online social network." In Advances in Social Networks Analysis and Mining (ASONAM), 2015 IEEE/ACM International Conference on, pp. 280-285. IEEE, 2015.

[Zhong2016] Zhong, Haoti, Hao Li, Anna Cinzia Squicciarini, Sarah Michele Rajtmajer, Christopher Griffin, David J. Miller, and Cornelia Caragea. "Content-Driven Detection of Cyberbullying on the Instagram Social Network." In IJCAI, pp. 3952-3958. 2016.

[Anderson2017] Anderson, Emma Louise, Eloisa Steen, and Vasileios Stavropoulos. "Internet use and Problematic Internet Use: A systematic review of longitudinal research trends in adolescence and emergent adulthood." International Journal of Adolescence and Youth 22, no. 4 (2017): 430-454.

References (12)

[Ashktorab2017] Ashktorab, Zahra. "Designing Cyberbullying Prevention and Mitigation Tools." PhD diss., University of Maryland, College Park, 2017.

[Bojanowski2017] Bojanowski, Piotr, Edouard Grave, Armand Joulin, and Tomas Mikolov. "Enriching Word Vectors with Subword Information." Transactions of the Association of Computational Linguistics5, no. 1 (2017): 135-146.

[Chatzakou2017] Chatzakou, Despoina, Nicolas Kourtellis, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, and Athena Vakali. "Mean birds: Detecting aggression and bullying on twitter." In Proceedings of the 2017 ACM on Web Science Conference, pp. 13-22. ACM, 2017.

[Chelmis2017] Chelmis, Charalampos, Daphney-Stavroula Zois, and Mengfan Yao. "Mining Patterns of Cyberbullying on Twitter." In Data Mining Workshops (ICDMW), 2017 IEEE International Conference on, pp. 126-133. IEEE, 2017.

[Gosling2017] Gosling, Samuel D., Sam Gaddis, and Simine Vazire. "Personality impressions based on facebook profiles." ICWSM7 (2007): 1-4.

[Raisi2017] Raisi, Elaheh, and Bert Huang. "Cyberbullying detection with weakly supervised machine learning." In Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017, pp. 409-416. ACM, 2017.

[Raisi2017b] Raisi, Elaheh, and Bert Huang. "Co-trained Ensemble Models for Weakly Supervised Cyberbullying Detection." NIPS 2017 Workshop on Learning with Limited Labeled Data: Weak Supervision and Beyond.

[Salawu2017] Salawu, Semiu, Yulan He, and Joanna Lumsden. "Approaches to Automated Detection of Cyberbullying: A Survey." IEEE Transactions on Affective Computing (2017).

References (13)

[Vishwamitra2017] Vishwamitra, Nishant, Xiang Zhang, Jonathan Tong, Hongxin Hu, Feng Luo, Robin Kowalski, and Joseph Mazer. "MCDefender: Toward Effective Cyberbullying Defense in Mobile Online Social Networks." In *Proceedings of the 3rd ACM on International Workshop on Security And Privacy Analytics*, pp. 37-42. ACM, 2017.

[Zhao2017] Zhao, Rui, and Kezhi Mao. "Cyberbullying Detection based on Semantic-Enhanced Marginalized Denoising Auto-Encoder." *IEEE Transactions on Affective Computing* 8, no. 3 (2017): 328-339.

[DiFranzo2018] DiFranzo, Dominic, Samuel Hardman Taylor, Franccesca Kazerooni, Olivia D. Wherry, and Natalya N. Bazarova. "Upstanding by design: Bystander intervention in cyberbullying." In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 2018.

[Founta2018] Founta, Antigoni-Maria and Djouvas, Constantinos and Chatzakou, Despoina and Leontiadis, Ilias and Blackburn, Jeremy and Stringhini, Gianluca and Vakali, Athena and Sirivianos, Michael and Kourtellis, Nicolas. "Large Scale Crowdsourcing and Characterization of Twitter Abusive Behavior." In 12th International Conference on Web and Social Media, ICWSM 2018

[Kazerooni2018] Kazerooni, Franccesca, Samuel Hardman Taylor, Natalya N. Bazarova, and Janis Whitlock. "Cyberbullying Bystander Intervention: The Number of Offenders and Retweeting Predict Likelihood of Helping a Cyberbullying Victim." *Journal of Computer-Mediated Communication* 23, no. 3 (2018): 146-162.

[Liu2018] Liu Ping, Guberman Joshua, Hemphill Libby, Culotta Aron. "Forecasting the presence and intensity of hostility on Instagram using linguistic and social features." In 12th International Conference on Web and Social Media, ICWSM 2018

[Mahar2018] Mahar, Kaitlin, Amy X. Zhang, and David Karger. "Squadbox: A Tool to Combat Email Harassment Using Friendsourced Moderation." (2018).

References (14)

[Rafiq2018] Rafiq, Rahat Ibn, Homa Hosseinmardi, Richard Han, Qin Lv, and Shivakant Mishra. "Scalable and Timely Detection of Cyberbullying in Online Social Networks." In Symposium on Applied Computing (2018).

[Rezvan2018] Rezvan, Mohammadreza, Saeedeh Shekarpour, Lakshika Balasuriya, Krishnaprasad Thirunarayan, Valerie L. Shalin, and Amit Sheth. "A Quality Type-aware Annotated Corpus and Lexicon for Harassment Research." In *Proceedings of the 10th ACM Conference on Web Science*, pp. 33-36. ACM, 2018.

[Ribeiro2018] Ribeiro, Manoel Horta, Pedro H. Calais, Yuri A. Santos, Virgílio AF Almeida, and Wagner Meira Jr. "Characterizing and Detecting Hateful Users on Twitter." arXiv preprint arXiv:1803.08977 (2018).

[Soni 2018] Soni Devin, and Singh Vivek. "Time Reveals all Wounds: Modeling Temporal Dynamics of Cyberbullying Sessions." In 12th International Conference on Web and Social Media, ICWSM 2018

[Teh2018] Teh, Phoey Lee, Chi-Bin Cheng, and Weng Mun Chee. "Identifying and Categorising Profane Words in Hate Speech." In Proceedings of the 2nd International Conference on Compute and Data Analysis, pp. 65-69. ACM, 2018

[Yao2018] Yao Mengfan, Chelmis Charalampos, and Zois, Daphney-Stavroula. " Cyberbullying Detection on Instagram with Optimal Online Feature Selection." In IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), 2018

[Zois2018] Zois, Daphney-Stavroula, Angeliki Kapodistria, Mengfan Yao, and Charalampos Chelmis. "Optimal online cyberbullying detection." In IEEE International Conference on Acoustics, Speech and Signal Processing, 2018

[AnonymousAlertsApp] http://www.anonymousalerts.com/webcorp/

[Askfm] https://safety.ask.fm/ask-fm-safety-guidelines-for-parents/

[BarkApp] https://www.bark.us/

References (15)

[CyberbullyZombiesAttackApp] https://itunes.apple.com/us/app/cyberbully-zombies-attack/id682193534?mt=8

[CBFirstAidApp] https://play.google.com/store/apps/details?id=de.teamdna.cybermobbing&hl=en_US

[FBStopBullying] https://www.facebook.com/safety/bullying

[InstagramHideComments] https://instagram-press.com/blog/2018/05/01/protecting-our-community-from-bullying-comments-2/

[NoMoreBullyingMeApp] https://nomorebullyingme.com/app

[PocketGuardianApp] https://gopocketguardian.com/

[ReThinkApp] http://www.rethinkwords.com/

[TwitterRules] https://help.twitter.com/en/managing-your-account/suspended-twitter-accounts