Aggressive, Repetitive, Intentional, Visible, and Imbalanced: Refining Representations for Cyberbullying Classification

Caleb Ziems
Emory University

Ymir Vigfusson
Emory University

Fred Morstatter
USC ISI

EMORY

USC Viterbi
Information Sciences Institute

ICWSM-2020
Atlanta, Georgia, USA June 8-11, 2020

Cyberbullying is a growing **public health** problem.

Exclusively human moderation is **infeasible**.

Automatic Cyberbullying Detection

$$f\left(w_1 \times \boxed{\text{f*ck}} + w_2 \times \boxed{\text{idiot}} + \ldots + w_n \times \boxed{\text{b*tch}}\right) =$$

Automatic Cyberbullying Detection

$$f\left(w_1 \times \boxed{\text{f*ck}} + w_2 \times \boxed{\text{idiot}} + \ldots + w_n \times \boxed{\text{b*tch}}\right)$$

$$f\left(w_1 \times \boxed{\text{(f*ck, you)}} + w_2 \times \boxed{\text{(an, idiot)}} + \ldots + w_n \times \boxed{\text{(b*tch, face)}}\right) =$$

Automatic Cyberbullying Detection

$$f\left(w_1 \times \boxed{\text{f*ck}} + w_2 \times \boxed{\text{idiot}} + \ldots + w_n \times \boxed{\text{b*tch}}\right)$$

$$f\left(w_1 \times \boxed{\text{(f*ck, you)}} + w_2 \times \boxed{\text{(an, idiot)}} + \ldots + w_n \times \boxed{\text{(b*tch, face)}}\right)$$

$$f\left(w_1 \times \boxed{\text{affect}} + w_2 \times \boxed{\text{bio}} + \ldots + w_n \times \boxed{\text{negemo}}\right) =$$

Automatic Cyberbullying Detection

LIWC

* affective processes : {happy, cried}
* biological processes : {eat, blood, pain}
* negative emotion : {hurt, ugly, nasty}

$$f\left(w_1 \times \boxed{\text{affect}} + w_2 \times \boxed{\text{bio}} + \ldots + w_n \times \boxed{\text{negemo}}\right) =$$

Automatic Cyberbullying Detection

LIWC

* affective processes : {happy, cried}
* biological processes : {eat, blood, pain}
* negative emotion : {hurt, ugly, nasty}

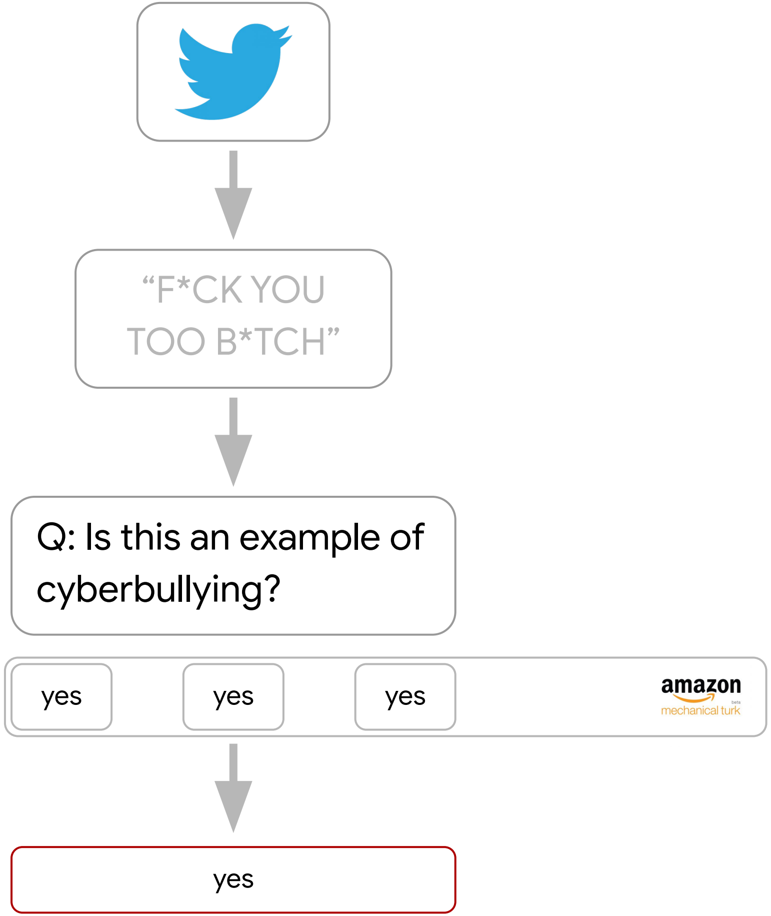
$$f\left(w_1 \times \boxed{\text{affect}} + w_2 \times \boxed{\text{bio}} + \ldots + w_n \times \boxed{\text{negemo}}\right) = ?$$

[11] Hosseinmardi et al. 2015

Existing Cyberbullying Datasets

Work	Source	Size	Balance	Context
Al-garadi et al. [1]	Twitter	10,007	6.0%	✗
Chatzakou et al. [3]	Twitter	9,484	-	✓
Hosseinmardi et al. [11]	Instagram	1,954	29.0%	✓
Huang et al. [13]	Twitter	4,865	1.9%	✗
Reynolds et al. [26]	Formspring	3,915	14.2%	✗
Rosa et al. [27]	Formspring	13,160	19.4%	✗
Sugandhi et al. [34]	Mixed	3,279	12.0%	✗
Van Hee et al. [35]	AskFM	113,698	4.7%	✓

Ground Truth?



Ground Truth?

"F*CK YOU TOO B*TCH"

Q: Is this an example of cyberbullying?

yes | yes | yes

amazon mechanical turk

yes

[redacted] · Jun 11

BAD BITCH 🗣️😛🤞🏽💙

> [redacted] Jun 11
>
> Replying to [redacted]

2 1

[redacted] · Jun 11

Wheew Don't My Head Bigger Than It Already Is 😂😭🗣️ THANK YOU BITCHH 💯💙🅿️

1 1 1

[redacted] · Jun 11

😭😭😭😭😭 you do have big head 1-95 ho 😂😂😂😂 you're welcomeeeee 💙😛

1 1

[redacted] · Jun 11

Bitchhh Ya Know What 🗣️ FUCK YA 😂😂😂😭💙

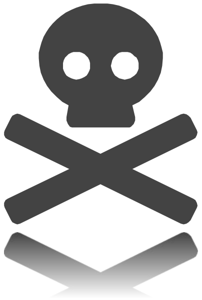
1 1

[redacted]

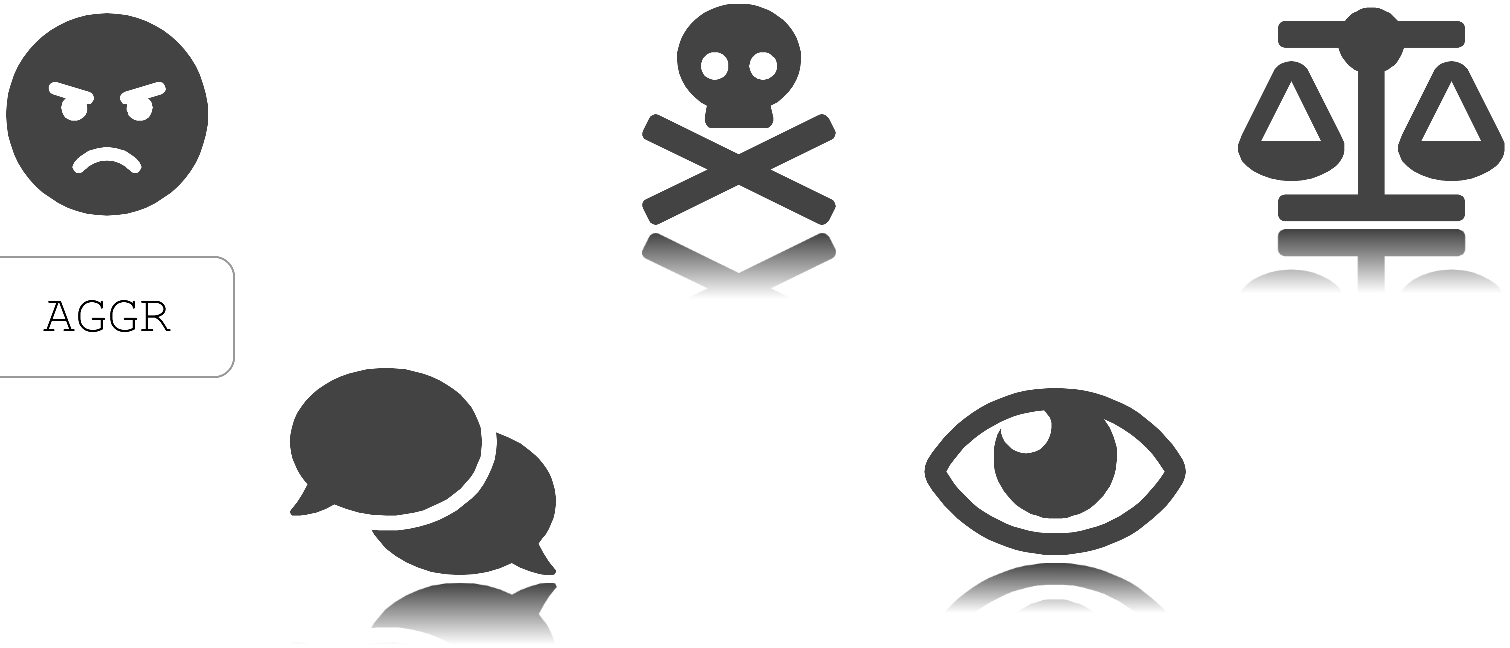
Replying to [redacted]

😂😂😂😂😂 FUCK YOU TOO BITCH

Defining Cyberbullying

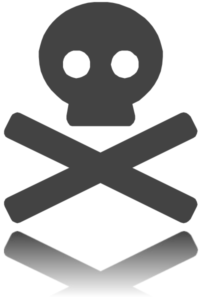


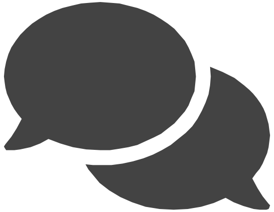
Defining Cyberbullying



Defining Cyberbullying

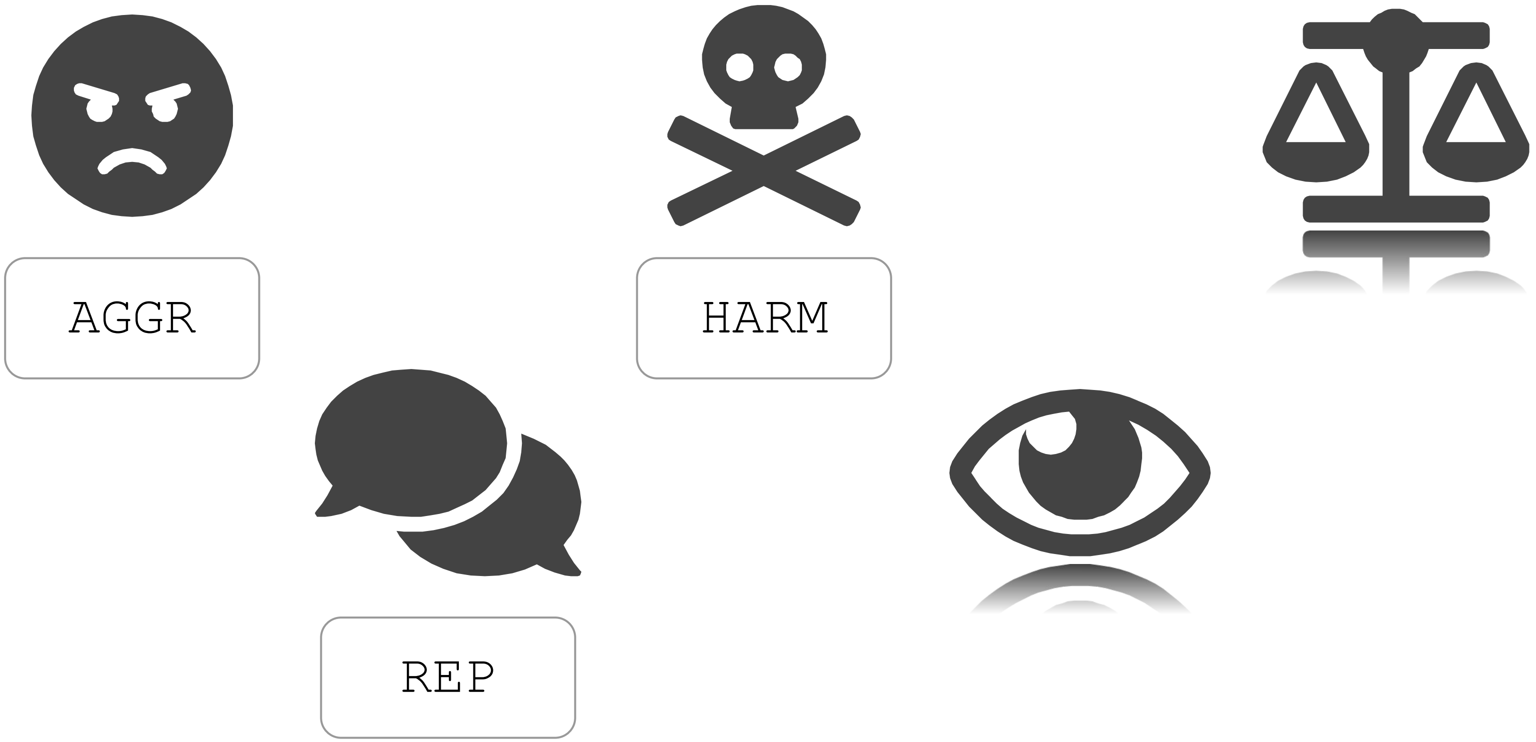
AGGR



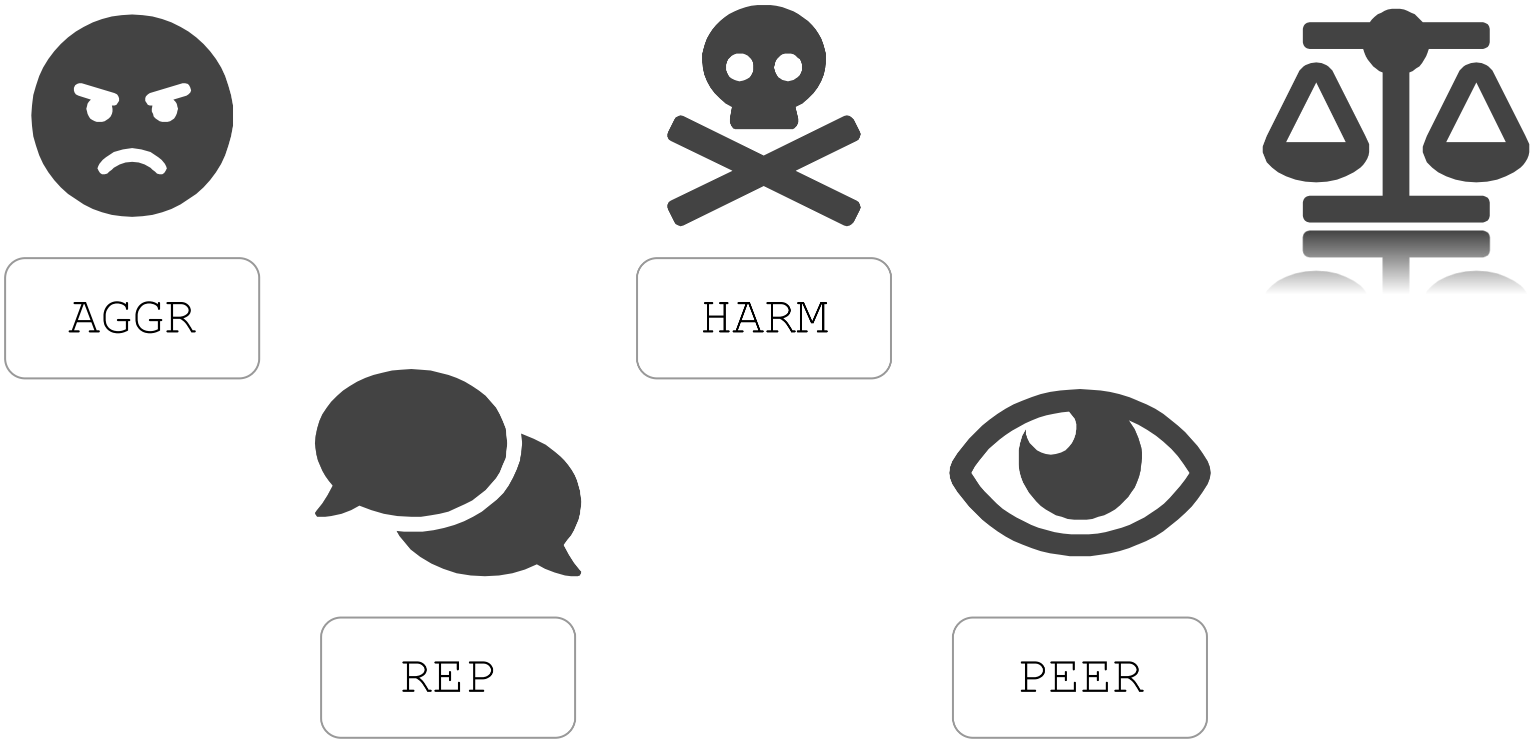


REP

Defining Cyberbullying



Defining Cyberbullying

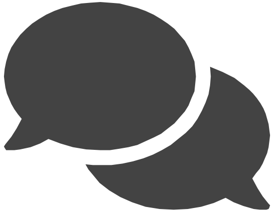


Defining Cyberbullying

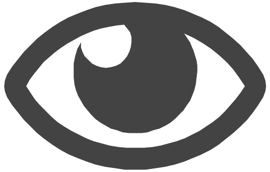
AGGR

HARM

POWER



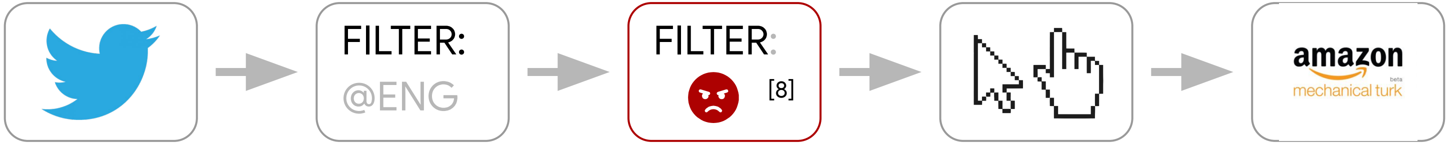
REP



PEER

Curating a Comprehensive Cyberbullying Dataset

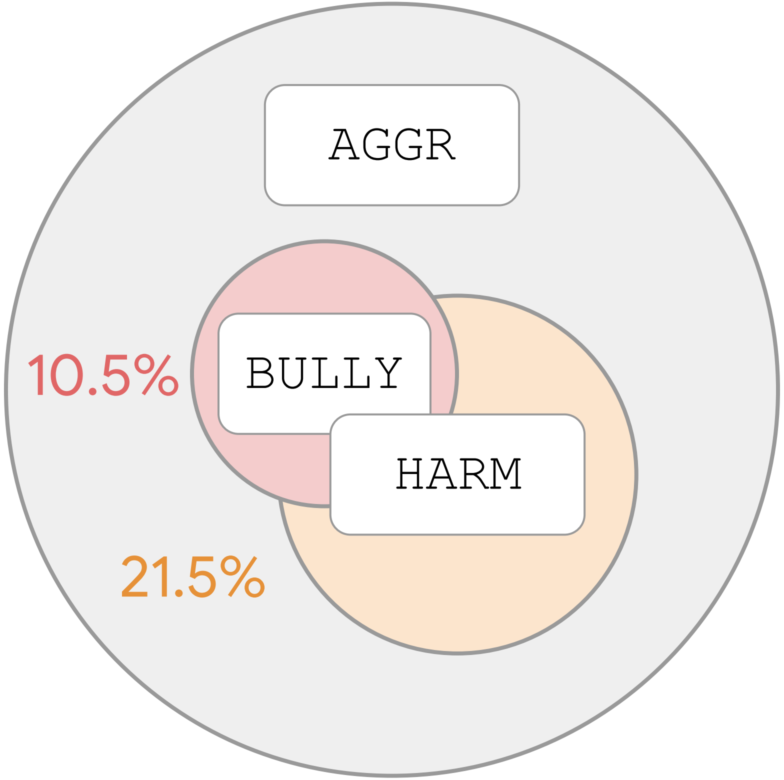
Data Collection



1.3 million Tweets → 280,301 → 9,803 → 6,897

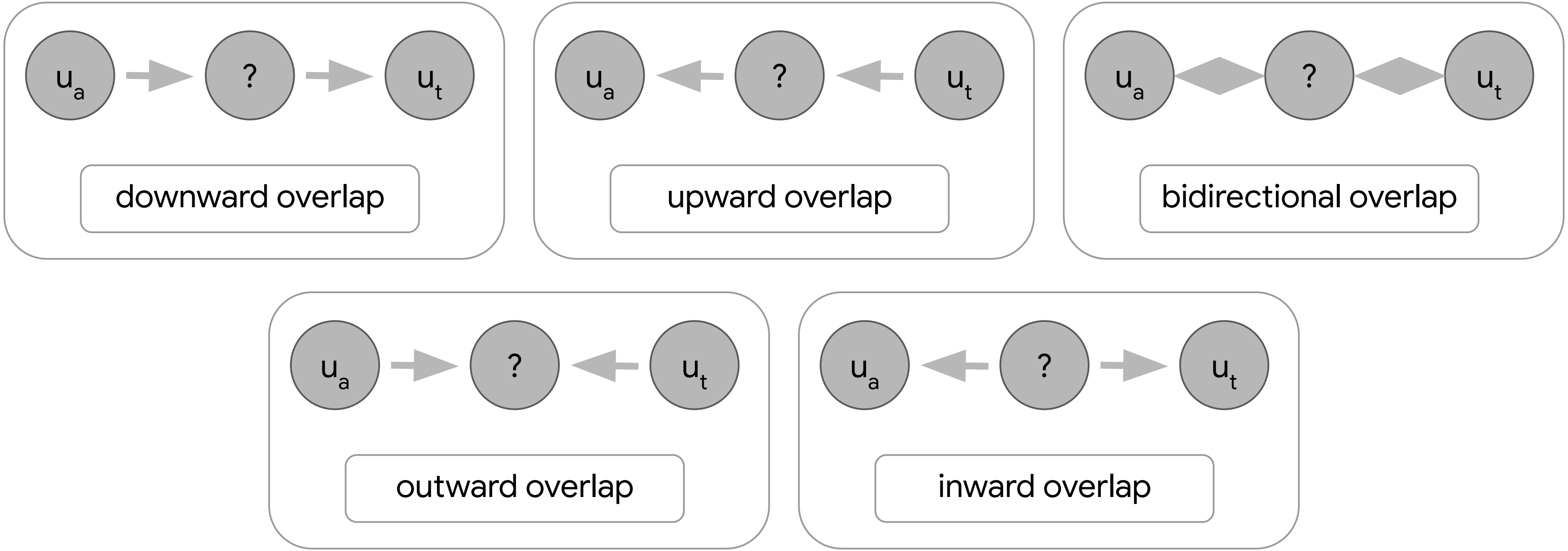
[8] Davidson et al. 2017

Analysis of Labeled Data



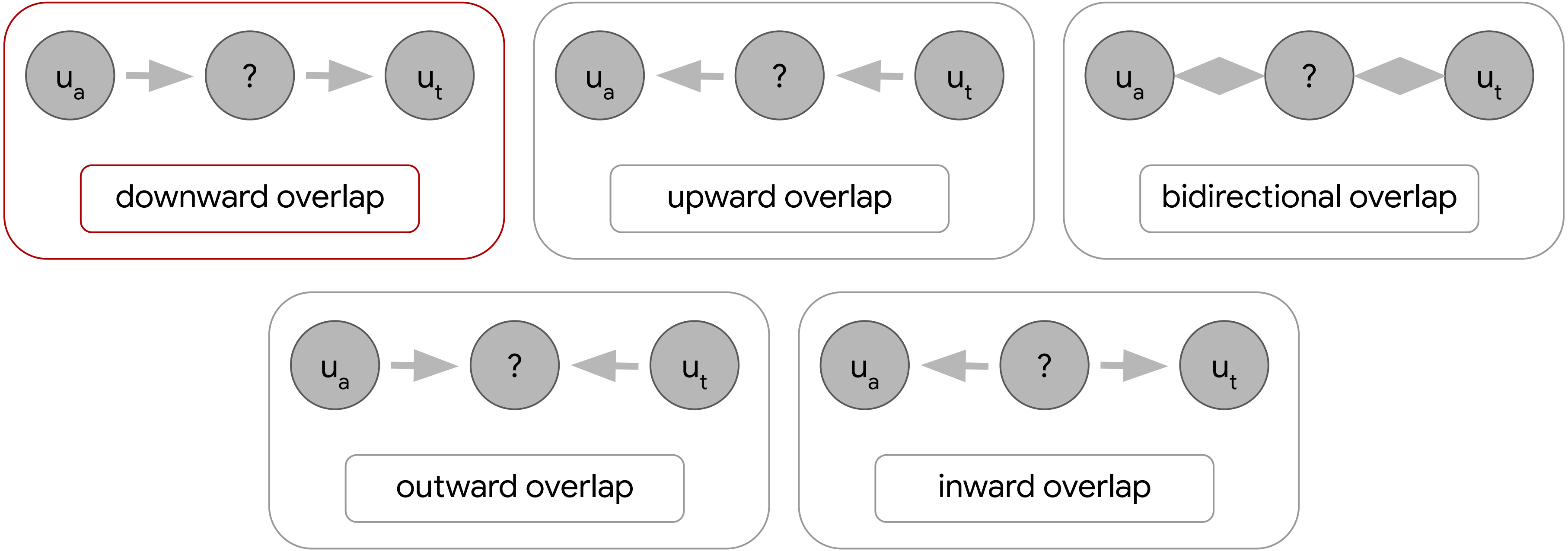
Feature Engineering

Social Network Features



$$\mathbf{overlap}(u_a, u_t) := \frac{|N(u_a) \cap N(u_t)|}{|N(u_a) \cup N(u_t)|}$$

Social Network Features



$$\textbf{overlap}(u_a, u_t) := \frac{|N(u_a) \cap N(u_t)|}{|N(u_a) \cup N(u_t)|}$$

Timeline Features

`auth -> targ`

downward mentions

`targ -> auth`

upward mentions

$$\frac{|M_a \cap M_t|}{|M_a \cup M_t|}$$

mention overlap

Timeline Features

$$\cos\theta = \frac{\vec{A}\cdot\vec{T}}{\|\vec{A}\|\|\vec{T}\|}$$

timeline similarity

```
new_words /
length of msg
```

new words ratio

$$\mathrm{H}(m) = -\frac{1}{N}\sum_{i=1}^{N}\log P(b_i)$$

linguistic surprise

Thread Visibility Features

total messages

reply messages

reply users

max author favorites

max author RTs

Thread Aggression Features

39

aggr[8] messages

7

auth aggr[8] msgs

aggr[8] users

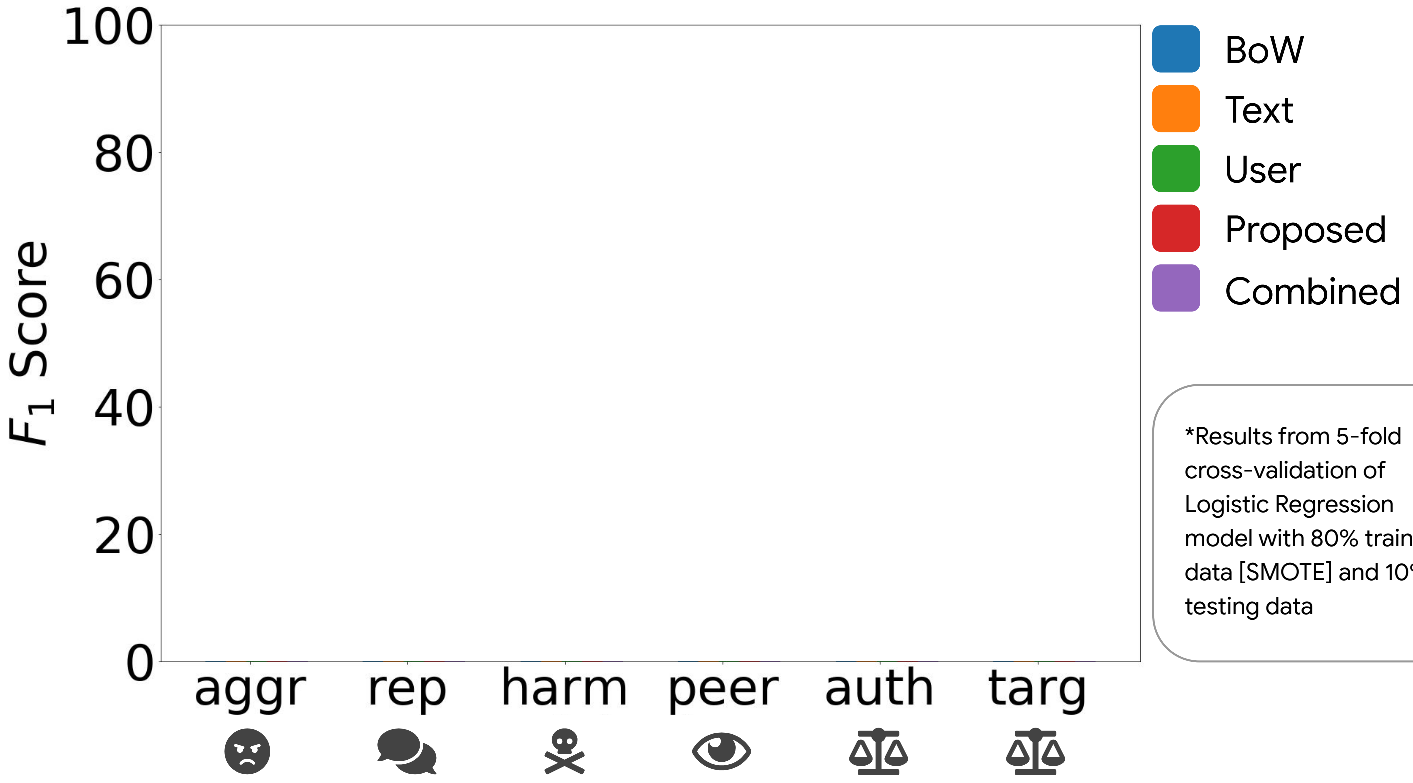
[8] Davidson et al. 2017

Experimental Evaluation

Feature Combinations

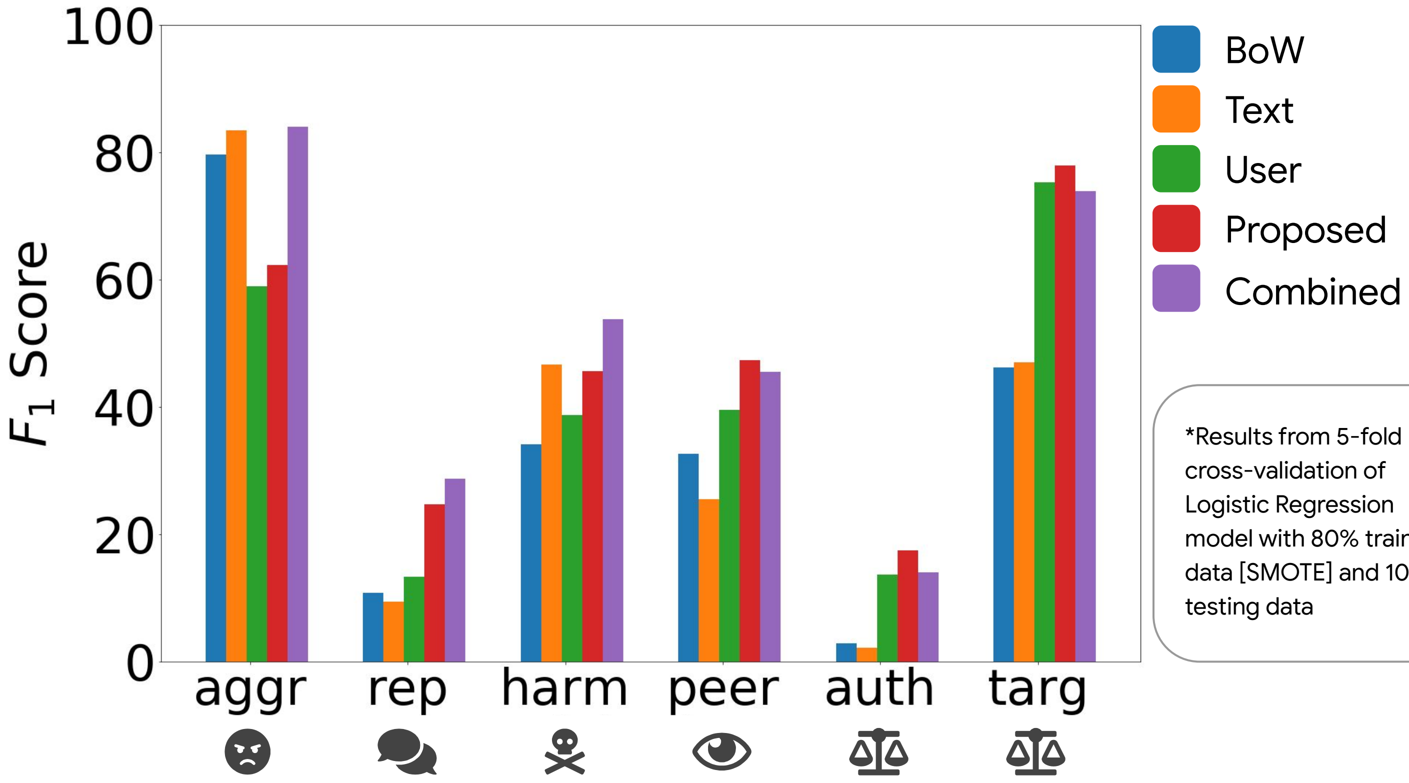
Feature	*BoW*	*Text*	*User*	*Proposed*	*Combined*
n-grams	✓	✓			✓
LIWC, VADER, Flesch-Kincaid		✓			✓
Friend/following counts, tweet count, verified			✓	✓	✓
Neighborhood overlap measures			✓	✓	✓
Mention counts and overlaps			✓	✓	✓
Timeline similarity			✓	✓	✓
New words ratio, cross-entropy			✓	✓	✓
Thread visibility features				✓	✓
Thread aggression features				✓	✓

Model Evaluation



*Results from 5-fold cross-validation of Logistic Regression model with 80% training data [SMOTE] and 10% testing data

Model Evaluation



*Results from 5-fold cross-validation of Logistic Regression model with 80% training data [SMOTE] and 10% testing data

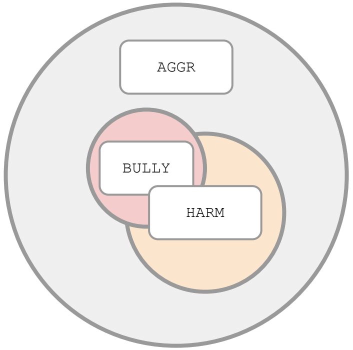
Limitations

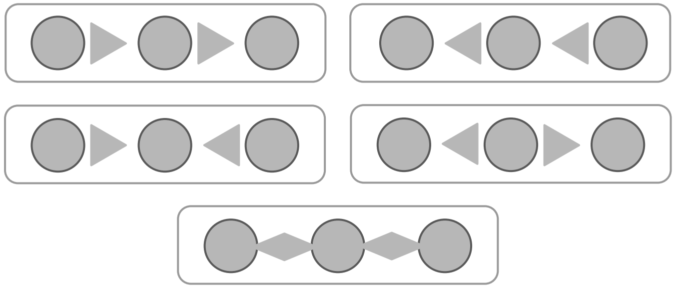
- Bias
  - Sampling bias [imperfect aggressive tweet filter]
  - Algorithmic bias [class imbalance]
- Subjectivity in the labeling process
  - low inter-rater agreement
  - *harmful intent* and *power imbalance* may depend on conventions or norms of a particular community
- Correlation, not implication
  - [cyberbullying] $\nLeftrightarrow$ [cyberbullying criteria]
  - *Cyberbullying* still hasn't been unambiguously defined

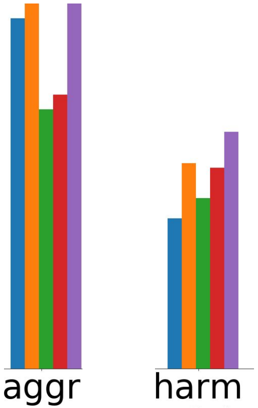
Cyberbullying detection remains an open research problem

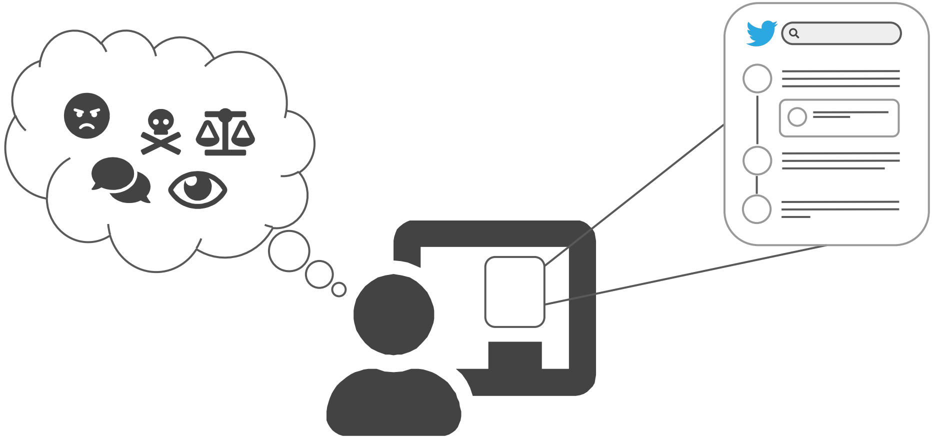
Conclusion

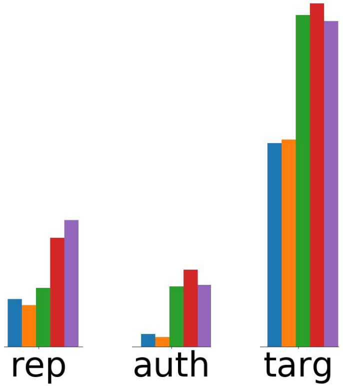
github.com/cjziems/cyberbullying-representations











Acknowledgements

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Agreement No. HR0011890019.

The project was completed in part at the **USC** *Information Sciences Institute*, supported by NSF Grant No. 1659886

and at **Emory University**, supported by NSF Grant No. 1553579.

References

[1] Mohammed Ali Al-garadi, 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: 433–443, 2016.

[2] Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit.* " O'Reilly Media, Inc.", 2009.

[3] Despoina Chatzakou, 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*, pages 13–22. ACM, 2017.

[4] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. SMOTE: synthetic minority over-sampling technique. *JAIR*, 16:321–357, 2002.

[5] Charalampos Chelmis, Daphney-Stavroula Zois, and Mengfan Yao. Mining patterns of cyberbullying on twitter. In *ICDMW*, pages 126–133. IEEE, 2017.

[6] Maral Dadvar, Dolf Trieschnigg, Roeland Ordelman, and Franciska de Jong. Improving cyberbullying detection with user context. In *European Conference on Information Retrieval*, pages 693–696. Springer, 2013.

[7] Corinne David-Ferdon and Marci F Hertz. Electronic media and youth violence; a CDC issue brief for researchers. 2009.

[8] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. Automated hate speech detection and the problem of offensive language. In *Eleventh International AAAI Conference on Web and Social Media*, 2017.

[9] Karthik Dinakar, Roi Reichart, and Henry Lieberman. Modeling the detection of textual cyberbullying. In *fifth international AAAI conference on weblogs and social media*, 2011.

[10] Sameer Hinduja and Justin W Patchin. Cyberbullying: An exploratory analysis of factors related to offending and victimization. *Deviant behavior*, 29(2):129–156, 2008.

[11] Homa Hosseinmardi, 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*, pages 49–66. Springer, 2015.

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

[13] Qianjia Huang, Vivek Kumar Singh, and Pradeep Kumar Atrey. Cyber bullying detection using social and textual analysis. In *Proceedings of the 3rd International Workshop on Socially-Aware Multimedia*, pages 3–6. ACM, 2014.

[14] Yun-yin Huang and Chien Chou. An analysis of multiple factors of cyberbullying among junior high school students in taiwan. *Computers in Human Behavior*, 26(6):1581–1590, 2010.

[15] Clayton J Hutto and Eric Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *ICWSM*, 2014.

[16] Robin M Kowalski and Susan P Limber. Psychological, physical, and academic correlates of cyberbullying and traditional bullying. *Journal of Adolescent Health*, 53(1):S13–S20, 2013.

[17] Qing Li. Cyberbullying in schools: A research of gender differences. *School psychology international*, 27(2):157–170, 2006.

[18] Kimberly Miller. Cyberbullying and its consequences: How cyberbullying is contorting the minds of victims and bullies alike, and the law's limited available redress. *S. Cal. Interdisc. LJ*, 26:379, 2016.

[19] Vinita Nahar, Xue Li, Chaoyi Pang, and Yang Zhang. Cyberbullying detection based on text-stream classification. In *The 11th Australasian Data Mining Conference (AusDM 2013)*, 2013.

[20] Vinita Nahar, Sanad Al-Maskari, Xue Li, and Chaoyi Pang. Semi-supervised learning for cyberbullying detection in social networks. In *Australasian Database Conference*, pages 160–171. Springer, 2014.

[21] Dan Olweus. Bullying at school. In *Aggressive behavior*, pages 97–130. Springer, 1994.

[22] Dan Olweus. Cyberbullying: An overrated phenomenon? *European Journal of Developmental Psychology*, 9(5):520–538, 2012.

[23] James W Pennebaker, Roger J Booth, and Martha E Francis. Liwc2007: Linguistic inquiry and word count. *Austin, Texas: liwc. net*, 2007.

[24] Megan Price, John Dalgleish, et al. Cyberbullying: Experiences, impacts and coping strategies as described by australian young people. *Youth Studies Australia*, 29(2):51, 2010.

[25] Juliana Raskauskas and Ann D Stoltz. Involvement in traditional and electronic bullying among adolescents. *Developmental psychology*, 43(3):564, 2007.

References

[26] Kelly Reynolds, April Kontostathis, and Lynne Edwards. Using machine learning to detect cyberbullying. In *2011 10th International Conference on Machine learning and applications and workshops*, volume 2, pages 241–244. IEEE, 2011.

[27] H Rosa, N Pereira, R Ribeiro, PC Ferreira, JP Carvalho, S Oliveira, L Coheur, P Paulino, AM Veiga Simão, and I Trancoso. Automatic cyberbullying detection: A systematic review. *Computers in Human Behavior*, 93:333–345, 2019.

[28] Semiu Salawu, Yulan He, and Joanna Lumsden. Approaches to automated detection of cyberbullying: A survey. *IEEE Transactions on Affective Computing*, 2017.

[29] Hugues Sampasa-Kanyinga, Paul Roumeliotis, and Hao Xu. Associations between cyberbullying and school bullying victimization and suicidal ideation, plans and attempts among canadian schoolchildren. *PloS one*, 9(7):e102145, 2014.

[30] Vivek K Singh, Qianjia Huang, and Pradeep K Atrey. Cyberbullying detection using probabilistic socio-textual information fusion. In *Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 884–887. IEEE Press, 2016.

[31] Robert Slonje and Peter K Smith. Cyberbullying: Another main type of bullying? *Scandinavian journal of psychology*, 49(2):147–154, 2008.

[32] Robert Slonje, Peter K Smith, and Ann Frisén. The nature of cyberbullying, and strategies for prevention. *Computers in human behavior*, 29(1):26–32, 2013.

[33] Devin Soni and Vivek Singh. Time reveals all wounds: Modeling temporal characteristics of cyberbullying. In *Twelfth International AAAI Conference on Web and Social Media*, 2018.

[34] Rekha Sugandhi, Anurag Pande, Abhishek Agrawal, and Husen Bhagat. Automatic monitoring and prevention of cyberbullying. *International Journal of Computer Applications*, 8:17–19, 2016.

[35] Cynthia Van Hee, Gilles Jacobs, Chris Emmery, Bart Desmet, Els Lefever, Ben Verhoeven, Guy De Pauw, Walter Daelemans, and Véronique Hoste. Automatic detection of cyberbullying in social media text. *PloS one*, 13(10):e0203794, 2018.

[36] Tracy E Waasdorp and Catherine P Bradshaw. The overlap between cyberbullying and traditional bullying. *Journal of Adolescent Health*, 56(5):483–488, 2015.

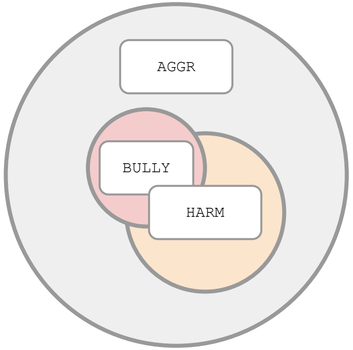
[37] Jun-Ming Xu, Kwang-Sung Jun, Xiaojin Zhu, and Amy Bellmore. Learning from bullying traces in social media. In *Proceedings of the 2012 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pages 656–666. Association for Computational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

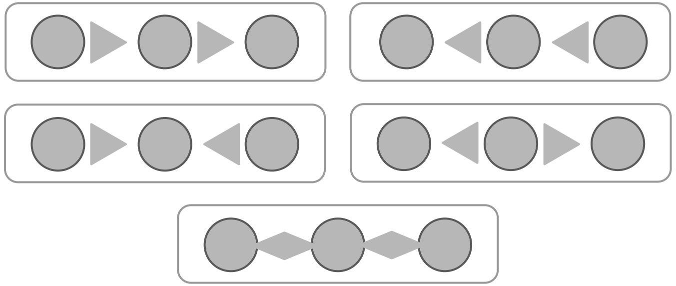
[38] Mengfan Yao, Charalampos Chelmis, Daphney Zois, et al. Cyberbullying ends here: Towards robust detection of cyberbullying in social media. In *The World Wide Web Conference*, pages 3427–3433. ACM, 2019.

[39] Xiang Zhang, Jonathan Tong, Nishant Vishwamitra, Elizabeth Whittaker, Joseph P Mazer, Robin Kowalski, Hongxin Hu, Feng Luo, Jamie Macbeth, and Edward Dillon. Cyberbullying detection with a pronunciation based convolutional neural network. In *2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 740–745. IEEE, 2016.

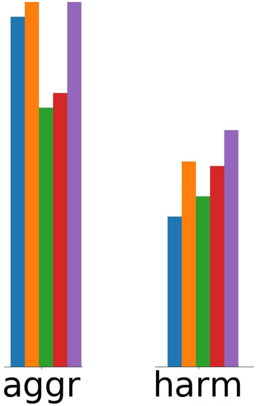
[40] Rui Zhao, Anna Zhou, and Kezhi Mao. Automatic detection of cyberbullying on social networks based on bullying features. In *Proceedings of the 17th international conference on distributed computing and networking*, page 43. ACM, 2016.

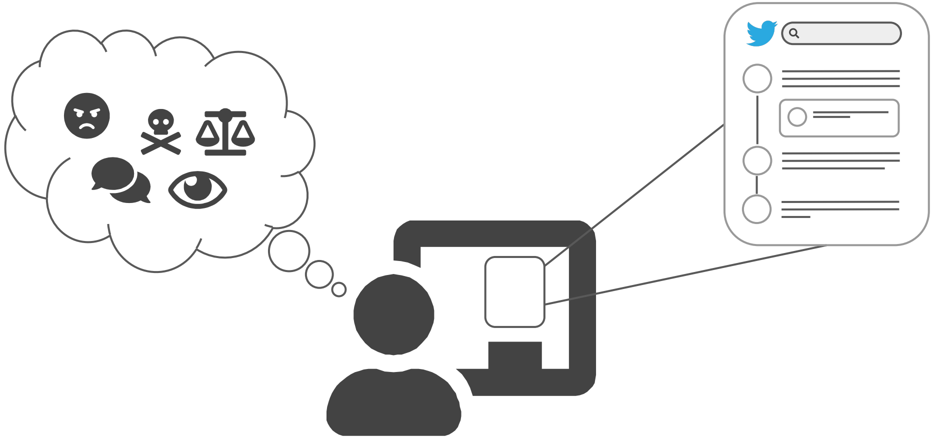
[Refining Representations for Cyberbullying Classification]

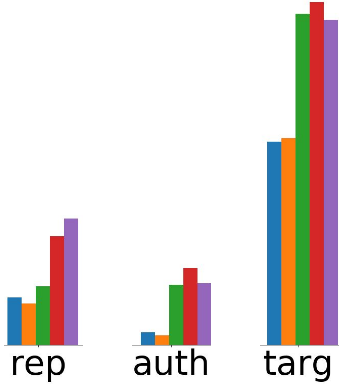




github.com/cjziems/cyberbullying-representations



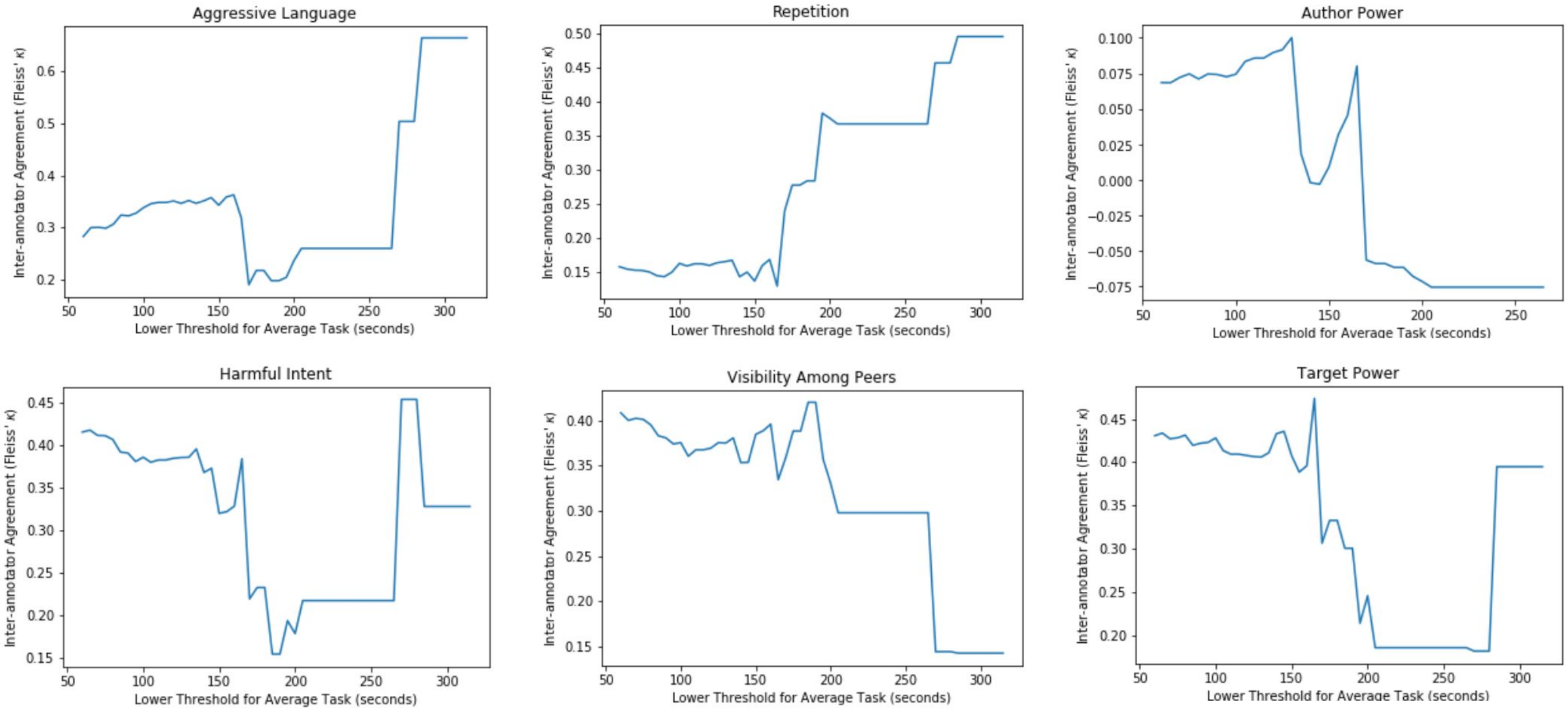




Spotlight Session II Q&A

Aggressive, Repetitive, Intentional, Visible, and Imbalanced: Refining Representations for Cyberbullying Classification

Low Inter-Annotator Agreement



Low Inter-Annotator Agreement

Table R.1: Inter-annotator agreement of Mechanical Turk workers on comments sourced from the /r/MTurk subreddit. These scores are lower than those obtained from our Twitter dataset.

Criterion	Positive Balance	Inter-annotator Agreement	Cyberbullying Correlation
aggression	69.0%	0.07	0.34
repetition	7.1%	0.21	0.53
harmful intent	17.8%	0.43	0.73
visibility among peers	48.2%	0.03	0.17
target power	2.0%	0.03	0.12
author power	1.0%	0.03	0.17
equal power	94.9%	0.05	-0.20
cyberbullying	14.7%	0.33	-

Other Classifiers

Table 17: Random Forest F_1

Criterion	*BoW*	*Text*	*User*	*Proposed*	*Combined*
aggression	65.2%	**79.3%**	56.0%	57.5%	77.9%
repetition	11.0%	10.6%	13.2%	**25.8%**	15.8%
harmful intent	25.6%	31.1%	46.6%	46.8%	**47.7%**
visibility among peers	35.7%	30.8%	41.2%	**46.1%**	33.6%
target power	47.4%	39.9%	**78.4%**	78.0%	72.8%

Table 19: AdaBoost F_1

Criterion	*BoW*	*Text*	*User*	*Proposed*	*Combined*
aggression	78.6%	**83.9%**	71.0%	77.5%	**83.9%**
repetition	11.7%	5.6%	11.5%	**21.6%**	20.9%
harmful intent	35.1%	41.6%	42.8%	45.4%	**55.0%**
visibility among peers	34.9%	21.0%	39.1%	**44.3%**	37.8%
target power	48.3%	42.7%	**79.8%**	79.6%	76.7%

Table 18: SVM F_1

Criterion	*BoW*	*Text*	*User*	*Proposed*	*Combined*
aggression	16.9%	37.7%	60.9%	**65.4%**	42.1%
repetition	12.6%	13.0%	11.8%	24.8%	**28.9%**
harmful intent	28.1%	33.8%	45.6%	**45.8%**	43.3%
visibility among peers	44.3%	46.1%	41.4%	**47.4%**	28.6%
target power	52.0%	35.8%	74.1%	**75.4%**	63.1%

Table 20: MLP F_1

Criterion	*BoW*	*Text*	*User*	*Proposed*	*Combined*
aggression	72.2%	**82.5%**	70.7%	72.4%	81.8%
repetition	12.0%	7.6%	12.4%	**20.7%**	15.2%
harmful intent	35.7%	37.3%	45.0%	**45.8%**	41.3%
visibility among peers	38.0%	27.7%	39.2%	**45.5%**	31.4%
target power	48.2%	41.0%	**75.4%**	74.0%	67.0%

Real-World Class Distribution

Criterion	Positive Balance	Inter-annotator Agreement	Cyberbullying Correlation
aggression	6.3%	0.23	0.68
repetition	0.9%	0.04	0.46
harmful intent	1.4%	0.31	0.75
visibility among peers	0.17%	0.51	0.11
target power	22.5%	0.23	0.11
author power	3.6%	0.04	0.06
equal power	64.7%	0.15	-0.14
cyberbullying	2.7%	0.25	-

Detection at the Intersection of Criteria

Cyberbullying Criteria	*BoW*	*Text*	*User*	*Proposed*	*Combined*
AGGR, REP	10.3%	7.8%	13.8%	26.6%	26.5%
AGGR, HARM	34.5%	47.3%	43.4%	44.4%	54.3%
AGGR, PEER	25.0%	21.7%	34.0%	38.3%	30.0%
AGGR, POWER	38.3%	39.1%	67.5%	67.8%	65.4%
REP, HARM	5.8%	5.2%	7.7%	15.0%	13.8%
REP, PEER	1.9%	2.9%	5.2%	10.8%	4.7%
REP, POWER	2.4%	4.2%	10.3%	9.9%	12.1%
HARM, PEER	10.5%	13.8%	17.5%	17.9%	20.5%
HARM, POWER	20.6%	37.0%	49.8%	49.4%	55.8%
PEER, POWER	15.2%	10.4%	34.4%	33.2%	23.3%
AGGR, REP, HARM	5.8%	5.2%	7.7%	15.0%	13.8%
AGGR, REP, PEER	3.7%	0.9%	5.0%	10.8%	3.5%
AGGR, REP, POWER	5.3%	4.4%	9.6%	9.7%	9.8%
AGGR, HARM, PEER	9.3%	18.3%	18.3%	19.5%	25.5%
AGGR, HARM, POWER	23.6%	34.9%	49.8%	49.2%	56.4%
AGGR, PEER, POWER	11.1%	11.5%	31.9%	29.7%	19.1%
REP, HARM, PEER	1.9%	4.8%	3.0%	6.6%	10.0%
REP, HARM, POWER	2.4%	4.0%	10.2%	9.9%	6.8%
REP, PEER, POWER	0.9%	0.0%	4.5%	4.1%	0.0%
HARM, PEER, POWER	7.5%	16.8%	16.8%	16.3%	22.4%
AGGR, REP, HARM, PEER	1.9%	4.8%	3.0%	6.6%	10.0%
AGGR, REP, HARM, POWER	2.4%	4.0%	10.2%	9.9%	6.8%
AGGR, REP, PEER, POWER	0.9%	0.0%	4.5%	4.1%	0.0%
AGGR, HARM, PEER, POWER	8.2%	15.4%	16.0%	15.7%	20.6%
REP, HARM, PEER, POWER	0.0%	0.0%	3.9%	4.7%	0.0%
AGGR, REP, HARM, PEER, POWER	0.0%	0.0%	3.9%	4.7%	0.0%