A Weakly Supervised Deep Model for Cyberbullying Detection

Elaheh Raisi and Bert Huang
Department of Computer Science, Virginia Tech

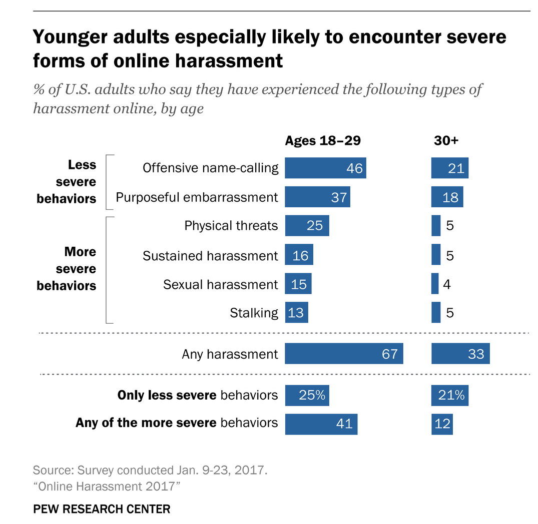
Cyberbullying

"willful and repeated harm inflicted through the use of computers, cell phones, and other electronic devices"
Patchin & Hinduja, 2006

- Forms of cyberbullying:
  - Offensive and negative comments
  - Name calling
  - Rumor spreading
  - Public shaming
  - Threads

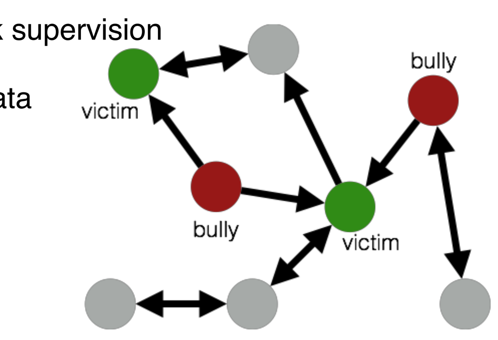
Dangers of Cyberbullying

- Linked to mental health issues:
  - decreasing academic performance
  - depression
  - anxiety
  - suicide
- Cyberbullying
  - persistent
  - public
  - not bounded by location or time
  - can be anonymously



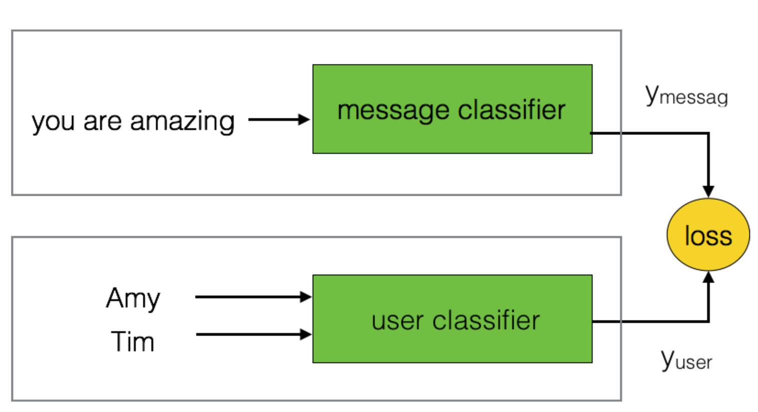
Machine Learning Challenges

- Cyberbullying involves rapidly evolving vocabulary and behavioral patterns
- Labeled examples of bullying require costly human expertise
- We must be able to learn with only weak supervision
- Need scalable algorithms for massive data
- Social structure is important



Co-Trained Ensemble Framework

- Two types of classifiers for harassment detection:
  - Message classifier (f : M → R):
    - Input: message
    - Output: classification score for whether the message is an example of harassment
  - User-relationship classifiers (g : U2 → R):
    - Input: pair of users
    - Output: score indicating whether one user is harassing the other user

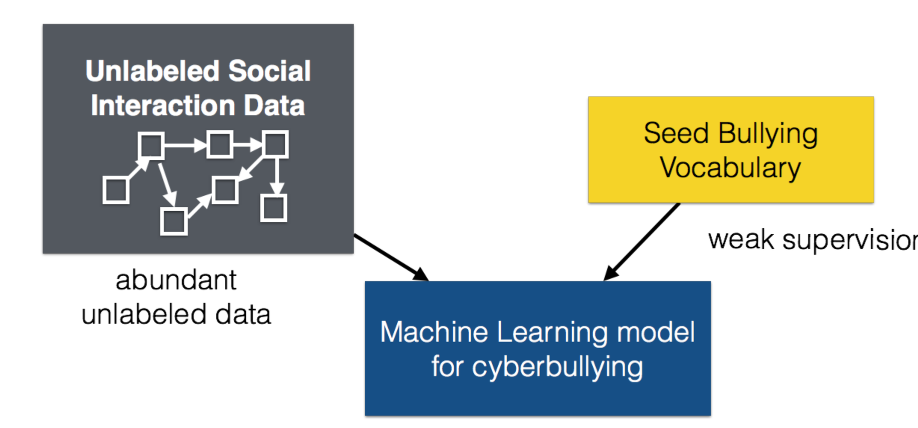


Training Objective

- *Consistency loss*: penalizes the disagreement between the message classifier and the user classifier
- *Weak supervision loss:* over message learner

$$\min_{\Theta} \underbrace{\frac{1}{2|M|}\sum_{m\in M}\left(f(m;\Theta) - g(s(m), r(m);\Theta)\right)^2}_{\text{consistency loss}} + \underbrace{\frac{1}{|M|}\sum_{m\in M}\ell(f(m;\Theta))}_{\text{weak supervision loss}},$$

for all messages; message learner score; user learner score; all parameters; sender; receiver



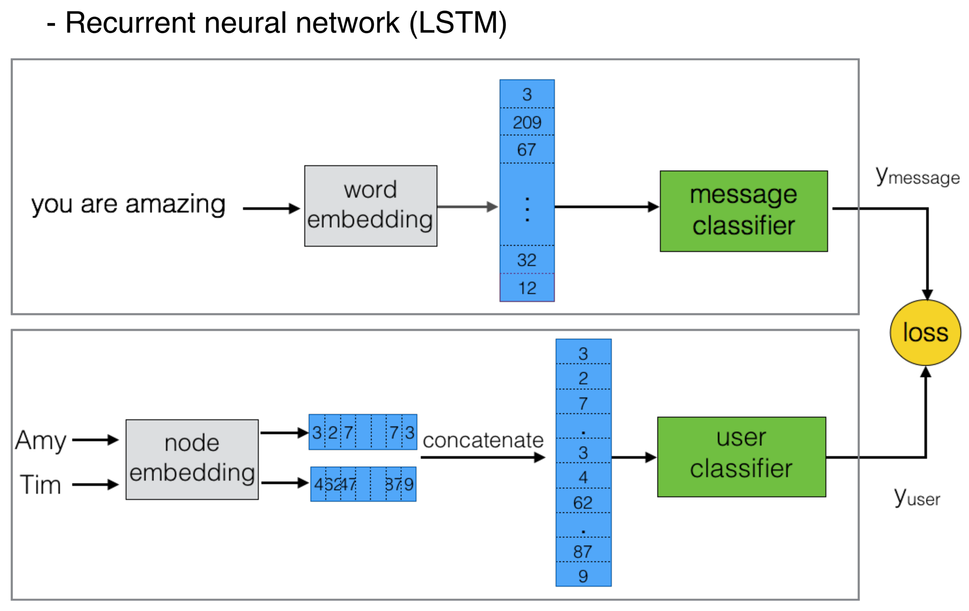
- Weak supervision loss on message learner:
  - Lower bound: Harassment indicator e.g. curse words, slurs, etc.
  - Upper bound: Harassment counter-indicator e.g. 'thanks'

$$\underbrace{\frac{n^+(m)}{n(m)}}_{\text{Lower Bound}} < y_m < \underbrace{1 - \frac{n^-(m)}{n(m)}}_{\text{Upper Bound}}$$

$$\ell(y_m) = -\log\left(\min\left\{1, 1 + \left(1 - \frac{n^-(m)}{n(m)}\right) - y_m\right\}\right) - \log\left(\min\left\{1, 1 + y_m - \frac{n^+(m)}{n(m)}\right\}\right).$$

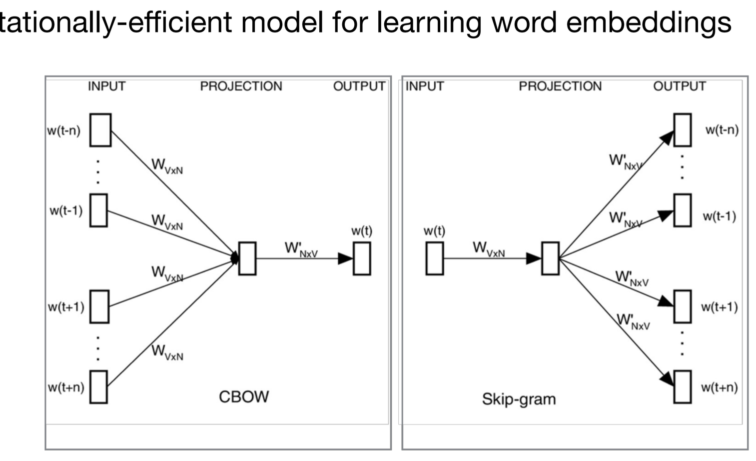
Models

- Message learner:
  - BOW
  - Pre-trained doc2vec
  - Custom-trained embedding
  - Recurrent neural network (LSTM)
- User learner:
  - Pre-trained node2vec
  - Custom-trained embedding
  - None



Word2vec Embedding

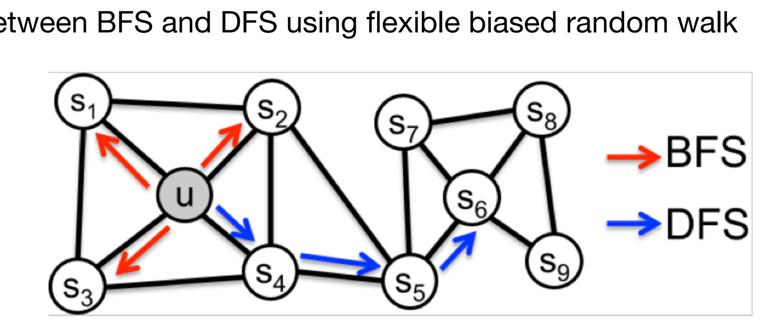
- Shallow, two-layer neural networks are trained
- Semantically similar words having similar vectors
- Computationally-efficient model for learning word embeddings



Mikolov, et al. 2013

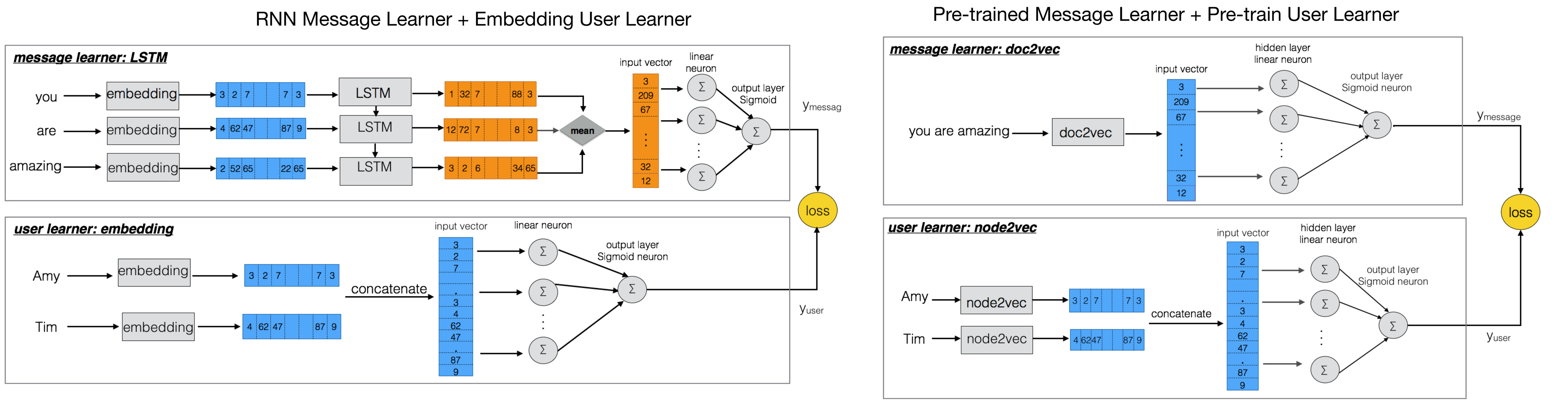
Node2vec Embedding

- Objective: maximizing the likelihood of preserving network neighborhoods of nodes
- Nodes neighborhood:
  - Communities the node belong to using BFS (u ~ s1)
  - Structural equivalence using DFS (u~s6)
- Interpolate between BFS and DFS using flexible biased random walk



Grover et al. 2016

Two Models



Experiments

Data summary

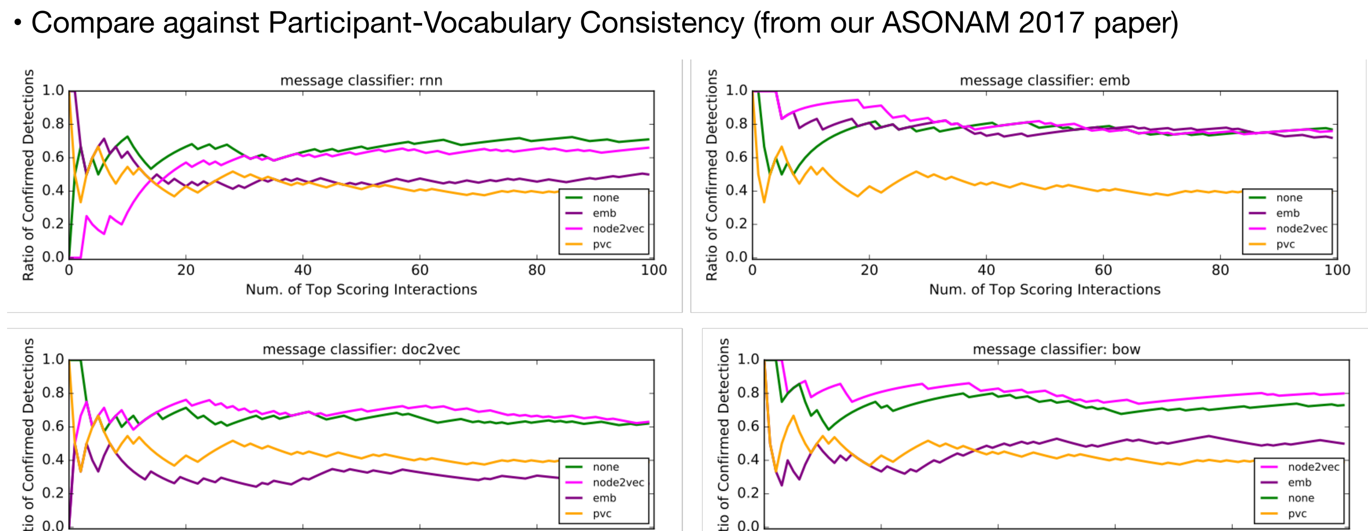
	# Users after preprocessing	# Messages after preprocessing
Twitter	180,355	296,308

noswearing.com	3,461 offensive unigrams and bigrams
positive opinion words (Hu et al., 2004)	2,005 positive words

BOW	1,000 hash functions
RNN	2 hidden layer of 100 dimensionality
Embedding	100 dimension

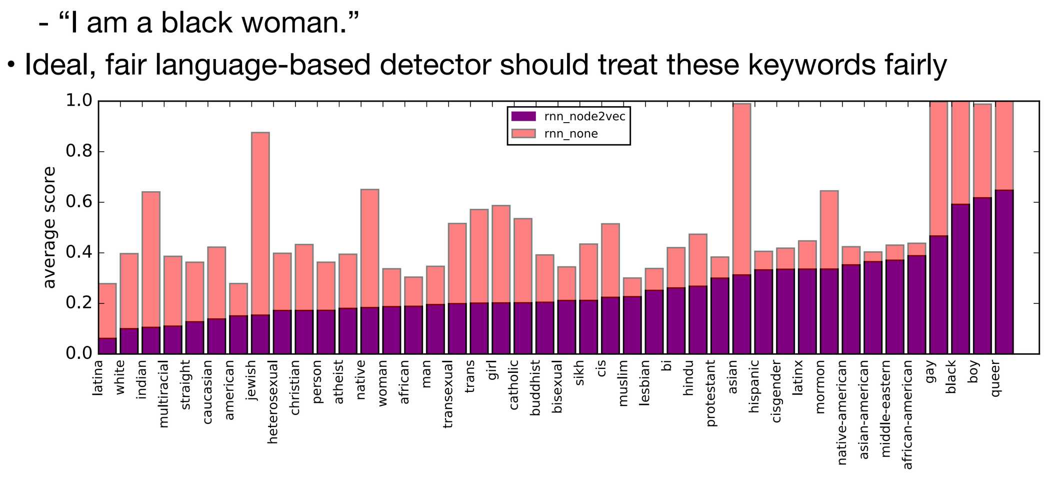
Precision@k

- For each method: extract 100 highest bullying-score conversations
- Five annotators rate as "yes", "no", or "uncertain"
- Compare against Participant-Vocabulary Consistency (from our ASONAM 2017 paper)



Identity Statement

- **Keyword score comparison**
- 42 sensitive keywords:
  - Sexual orientation, race, gender, and religion
- Create a corpus of sentences using the combination of sensitive keywords:
  - "I am a black woman."
- Ideal, fair language-based detector should treat these keywords fairly



- **Score-based Comparison**
- Using different combination of message and user learners
  - Compute the average score of sentences containing each keyword

method	average score
rnn_emb	0.147
rnn_node2vec	0.257
emb_none	0.381
doc2vec_none	0.497
doc2vec_emb	0.504
rnn_none	0.506
doc2vec_node2vec	0.511
bow_emb	0.515
emb_emb	0.518
bow_node2vec	0.536
bow_none	0.543
emb_node2vec	0.588