Data Mining & Machine Learning

Arizona State University

Hierarchical Attention Networks for Cyberbullying Detection on the Instagram Social Network

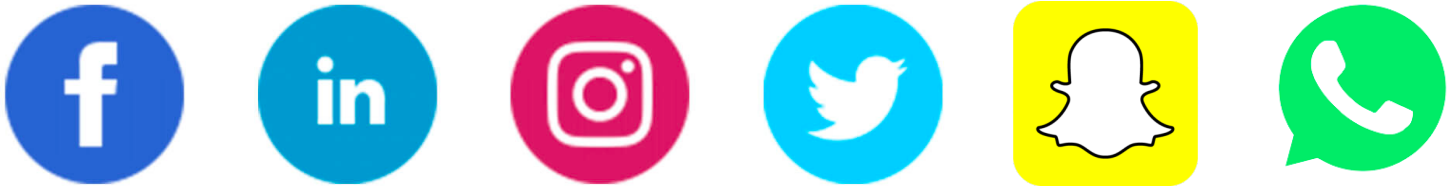
Lu Cheng, Ruocheng Guo, Yasin Silva, Deborah Hall, Huan Liu

Arizona State University

http://www.public.asu.edu/~lcheng35/

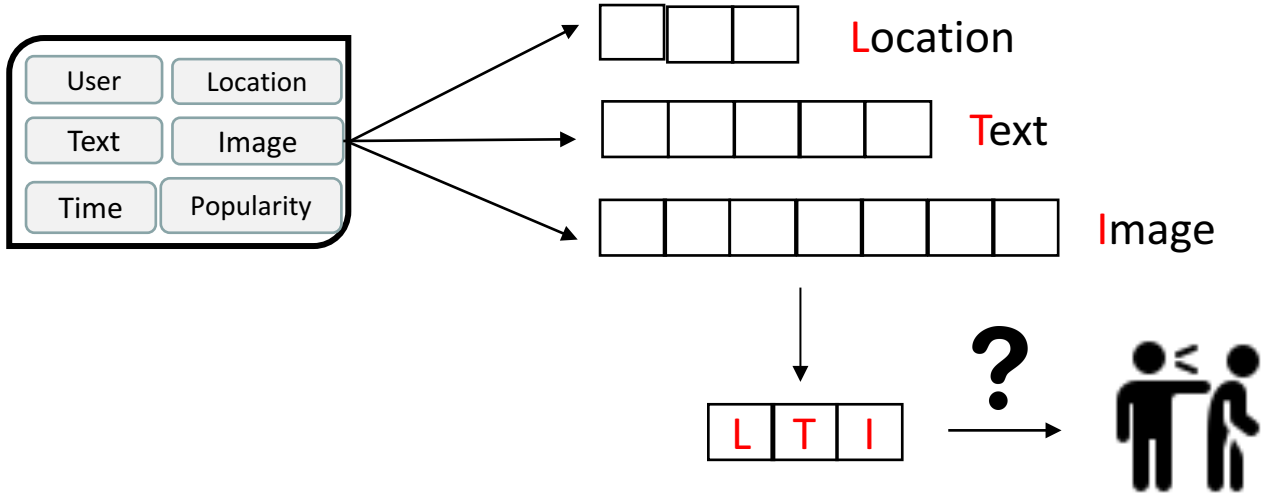
Cyberbullying in Social Media

- Traditional Definition
  - Cyberbullying is commonly defined as the electronic transmission of insulting or embarrassing comments, photos, or videos.
- Studies show that over half of adolescents and teens are faced with cyberbullying
  - Canada: ~20%, US: ~43%, Mainland China: ~57%



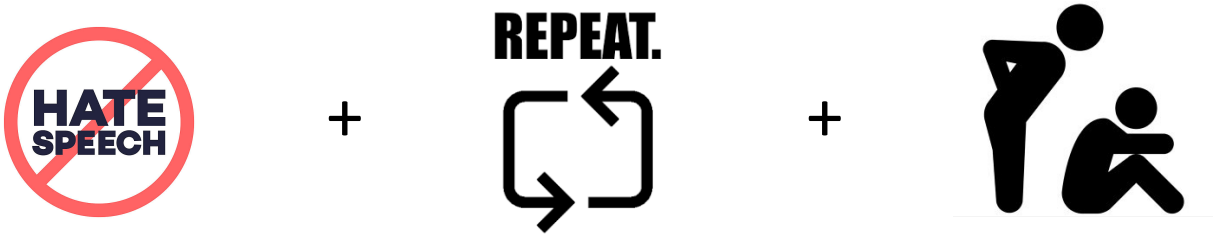
Simple Binary Classification Task?

- Existing Work
  - Feature engineering
  - Off-the-shelf classification models



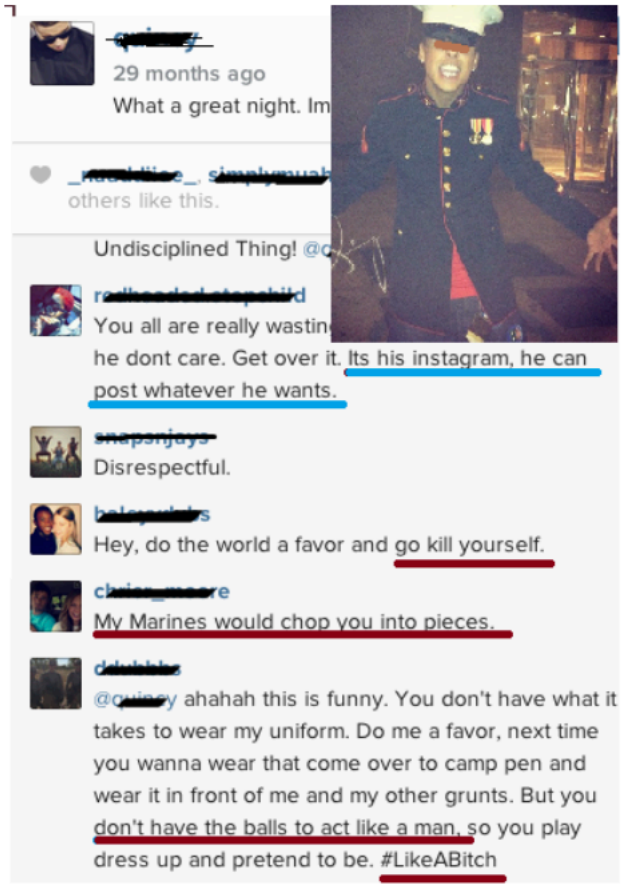
Simple Binary Classification Task?

- The Emerging Definition
  - "An *aggressive, intentional act* carried out by a group or individual using electronic forms of contact, *repeatedly or over time* against a victim that *cannot easily defend him or herself*"



  - Temporal dynamics

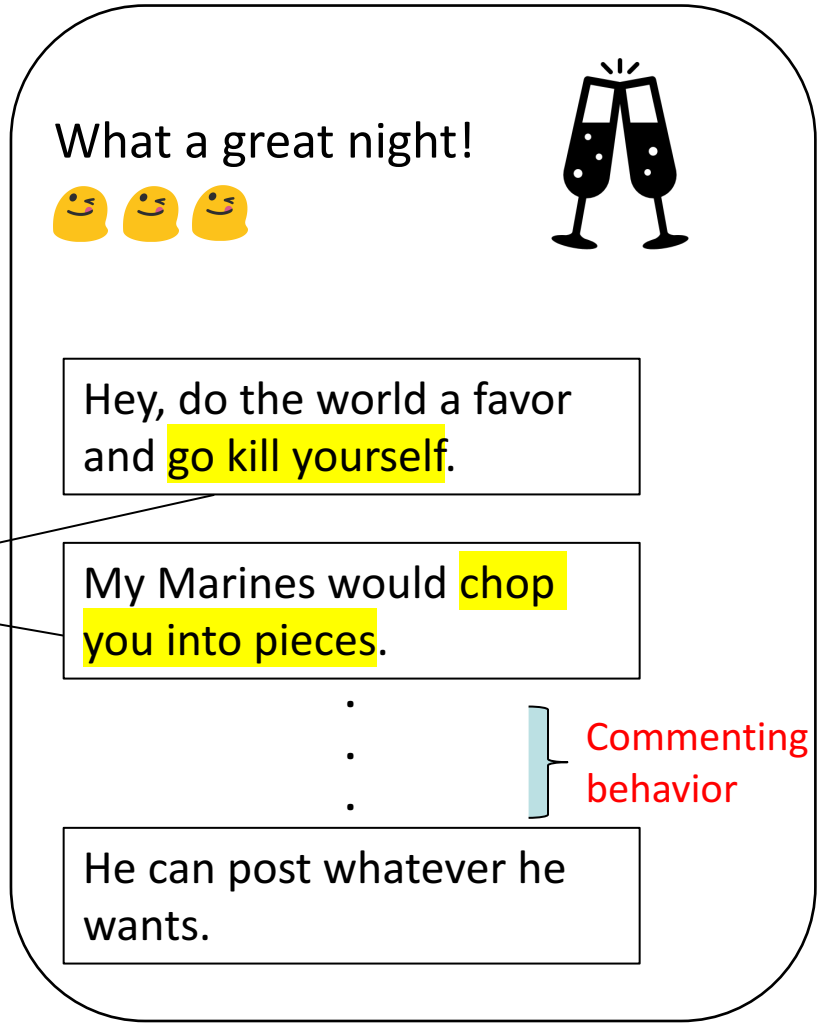
A Cyberbullying Example

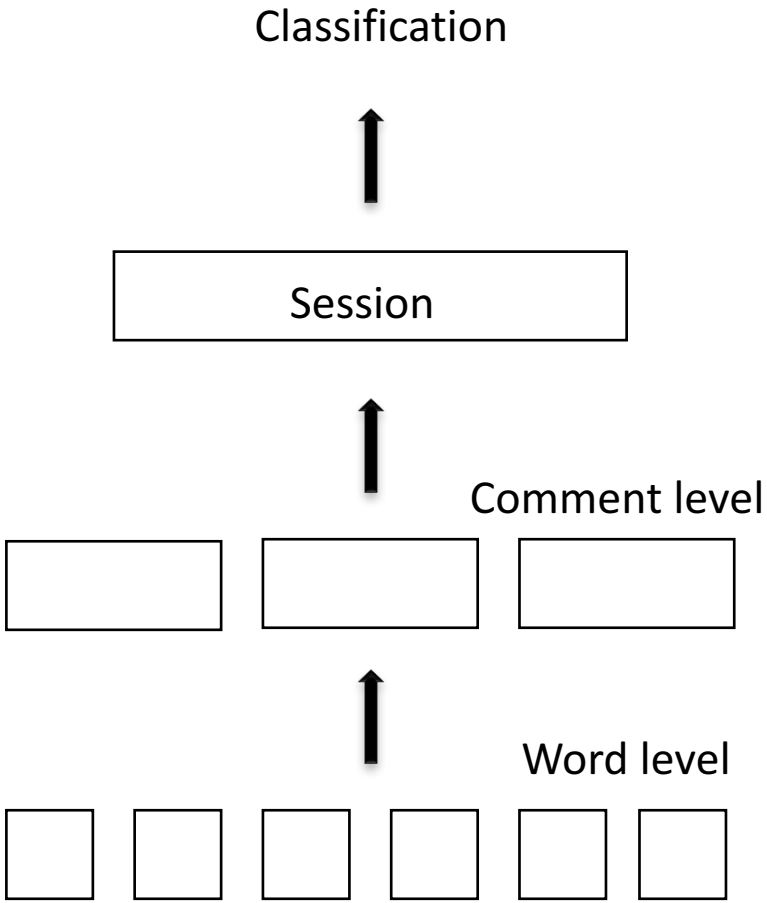


- Social Media Session
  - Image
  - Caption
  - Comments
  - Social information
  - Time stamps

Hosseinmardi, H., Mattson, S. A., Rafiq, R. I., Han, R., Lv, Q., & Mishra, S. (2015, December). Analyzing labeled cyberbullying incidents on the instagram social network. In *International conference on social informatics* (pp. 49-66). Springer, Cham.

Motivation

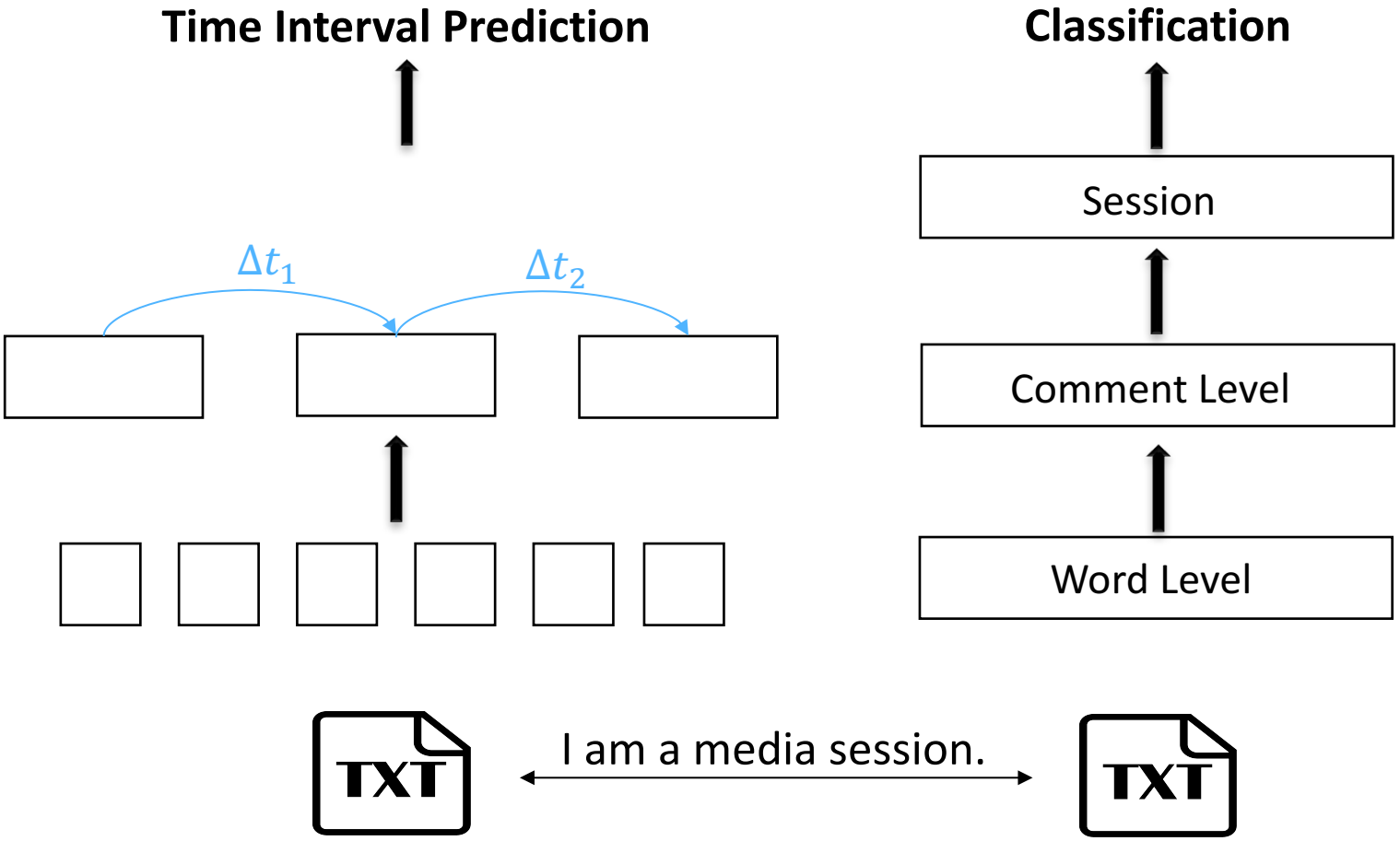

What a great night!
Hey, do the world a favor and go kill yourself.
Attention
My Marines would chop you into pieces.
Commenting behavior
He can post whatever he wants.


Classification
Session
Comment level
Word level

Research Questions

- How to model the temporal characteristics of the commenting behavior?
  - Extract more features, e.g. temporal features?
- How to model the hierarchical structure of a media session?
- How to differentiate the importance of words and comments?
  - e.g., 'You're a f**k gay!', 'I am a gay.'
  - Attention model

The HANCD Model



Some Math

- Encoder and Attention

$$w_{it} \rightarrow x_{it} : x_{it} = W_e w_{it}, \forall t \in [1, L_i], i \in [1, C].$$

$$\overrightarrow{s}_{it} = \overrightarrow{GRU}(x_{it}), \quad \forall t \in [1, L_i], i \in [1, C],$$

$$\overleftarrow{s}_{it} = \overleftarrow{GRU}(x_{it}), \quad \forall t \in [L_i, 1], i \in [1, C].$$

Word Encoder

$$h_{it} = \tanh(W_w s_{it} + b_w)$$

$$\alpha_{it} = \frac{\exp(h_{it}^\top u_w)}{\sum_t \exp(h_{it}^\top u_w)}.$$

Attention

$$c_i = \sum_t \alpha_{it} s_{it}.$$

- Loss Function

$$\ell_1 = -\sum_{n=1}^{N} \log p_n.$$

Classification Error

$$\ell_2 = \sum_{n=1}^{N} \sum_{i=1}^{C} \frac{1}{2} ||A_n s_i + q_n - \Delta t_i||^2,$$

Time Interval Prediction Error

Experiments

- Dataset
  - Instagram: 2,218 sessions, 678 are bullying sessions, 1,540 are the normal. The average #comments is ~70. The total #comments is 155,260
- Baselines
  - Standard classification models, e.g., XGBoost
  - Deep learning models for document classification, e.g., LSTM
  - Cyberbullying detection models

Experiments Cont'

Table 1: Performance comparisons of different models (F1 score).

Features	Count Vector	Word TF-IDF	N-gram TF-IDF	Char TF-IDF	LIWC	Embedding
KNN	0.476	0.521	0.501	0.479	0.559	0.236
Naive Bayesian	0.614	0.469	0.607	0.534	0.482	0.355
Logistic Regression	0.700	0.642	0.608	0.677	0.700	0.163
Random Forest	0.585	0.618	0.585	0.617	0.650	0.190
XGBoost	0.715	0.726	0.699	0.674	0.700	0.337
Deep Learning Models			Cyberbullying Detection Models			
LSTM	CNN	HAN	Xu et al.	Soni & Singh	HANCD	
0.613	0.613	0.708	0.502	0.740	**0.783**	

Table 2: Performance comparisons of different models (AUC score).

Features	Count Vector	Word TF-IDF	N-gram TF-IDF	Char TF-IDF	LIWC	Embedding
KNN	0.770	0.697	0.624	0.708	0.686	0.499
Naive Bayesian	0.706	0.815	0.797	0.786	0.622	0.525
Logistic Regression	0.812	0.825	0.827	0.830	0.776	0.629
Random Forest	0.788	0.804	0.788	0.781	0.743	0.544
XGBoost	0.838	0.828	0.831	0.840	0.772	0.621
Deep Learning Models			Cyberbullying Detection Models			
LSTM	CNN	HAN	Xu et al.	Soni & Singh	HANCD	
0.791	0.781	0.805	0.513	0.810	**0.851**	

Conclusions

- Cyberbullying detection from the temporal perspective
  - It is a *repetitive* act
  - Modeling temporal dynamics of commenting behavior is important
  - *Interdisciplinary study helps*
- Structure, context are important
  - Differentiate the importance of words and comments

Data Mining & Machine Learning

ARIZONA STATE UNIVERSITY

Q&A