Not so *Cute* but *Fuzzy*: Estimating Risk of Sexual Predation in Online Conversations.

Tatiana Ringenberg*, ***Kanishka Misra****, Julia Taylor Rayz

Applied Knowledge Representation and Natural Language Understanding Lab
Purdue University, USA

Grooming Stages (O’Connell, 2003)

- Process is motivation-driven
- Non-Linear Stages
    - Differ in length and order
    - Repetitive
- Varies based on desired outcome
- Desired outcome may change

Pacing of Conversations

- First 20% introduces multiple stages (Black, et al. 2015)
- Taboo topics gradually introduced
- Escalation and deescalation based on response
- Stages often overlap

Perverted Justice Corpus

- Vigilante organization which helps law enforcement perform sting operations
- Website stores conversations between offenders and decoys
- Decoys pretend to be a minor for Law Enforcement
- 2004 to present
- 623 chats
- Variety of motivations of offenders

Automatic Detection of Grooming Lines

- Researchers have identified lines corresponding to offender conversations (Cano, et al. 2014):
    - Grooming
    - Approach
    - Trust
- Others identified features specific to grooming (Michalopoulos & Mavridis 2011):
    - Sexual affair
    - Gaining Access
    - Deceptive relationship
- The majority have focused on differentiating offender versus non-offender (McGhee et al. 2011; Parapar et al. 2012; Ebrahimi et al. 2016)

Labeling Risk

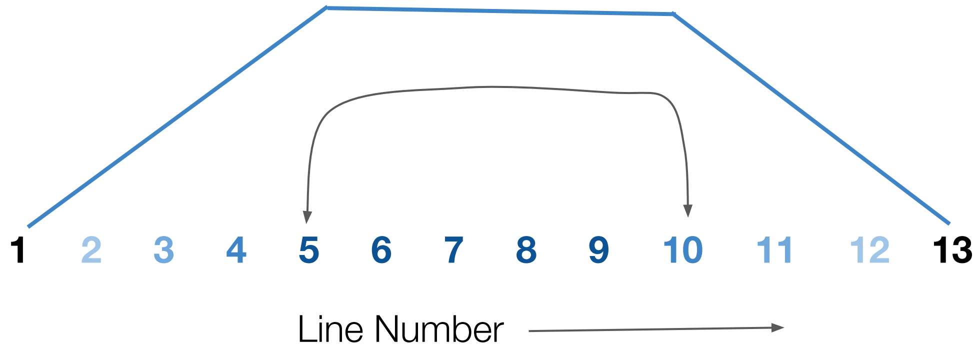
Low	Typical, non-sexual chat	Friendship Forming, Relationship Forming, Non-Sexual Risk Assessment
Medium	Affection, physical compliments, secrecy, guilt, implicit sexual undertones	Exclusivity
High	Explicit sexual content, references to digital to physical transition	Sexual stage, Meeting

Labeling Perverted Justice Corpus

- 13,648 labeled lines in total
- Labeled by researcher in field
- Labeled as chunks
- ±3 lines chosen for transition (3 before, 3 after)

Labeling Perverted Justice Corpus

- 13,648 labeled lines in total
- Labeled by researcher in field
- ±3 lines chosen for transition (3 before, 3 after)



Low Risk Example

Solicitor: hey
Decoy: hey. ur in jasper?
Solicitor: yes
Decoy: kool wats u doin
Solicitor: nothing
Solicitor: i’m just laying in bed

Medium Risk Example

Solicitor: look at you just a (***)
Solicitor: lol
Decoy: thanks :p
Solicitor: i think my fav is you in the (***)
Solicitor: well i like them all actually
Decoy: thanks yeah it shows the most of me
Solicitor: yeah a lil bit of your (***)
Solicitor: lol
Decoy: lol yeah i bet you like that ¿:)
Solicitor: yeah i do

High Risk Example

Solicitor: i'm soo bored ..i'm coming to get u
Solicitor: jk
Solicitor: ouch ..good move
Decoy: ohhh ur jk?lol
Solicitor: unless u want me to ;)

Fuzzification of Expert Labels

Crisp Labels of risk → Trapezoidal Membership Function

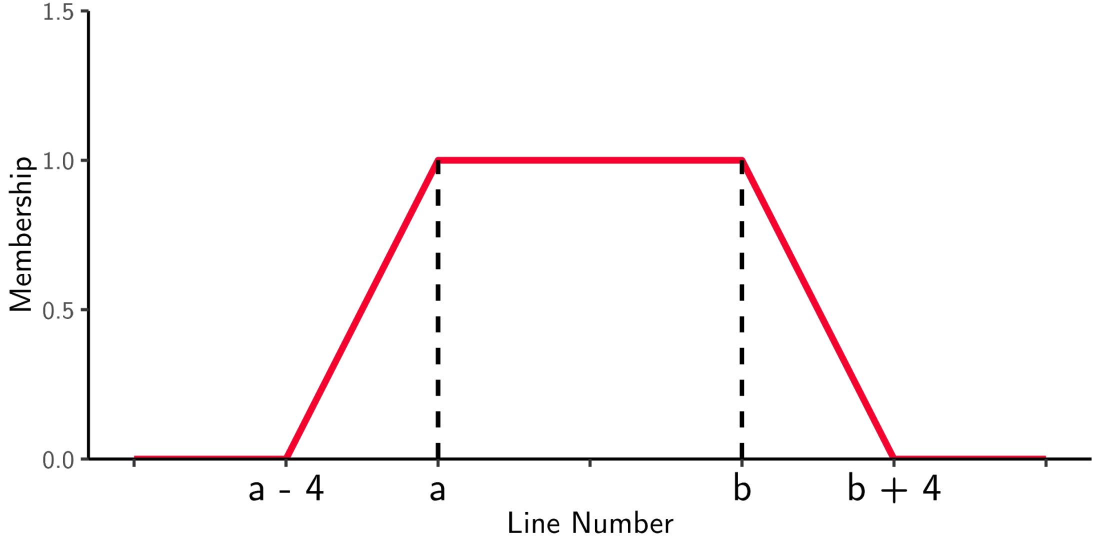
Uniformly increase membership for the 3 preceding lines, decrease for 3 succeeding.

$$\mu_C(l) = \begin{cases} \frac{l-a}{4} & \text{if } a-4 \leq l < a \\ 1 & \text{if } a \leq l \leq b \\ \frac{b+4-l}{4} & \text{if } b < l \leq b+4 \\ 0 & \text{otherwise} \end{cases}$$

Fuzzification of Expert Labels

Crisp Labels of risk → Trapezoidal Membership Function

Uniformly increase membership for the 3 preceding lines, decrease for 3 succeeding.



Fuzzification of Expert Labels

Crisp Labels of risk → Trapezoidal Membership Function

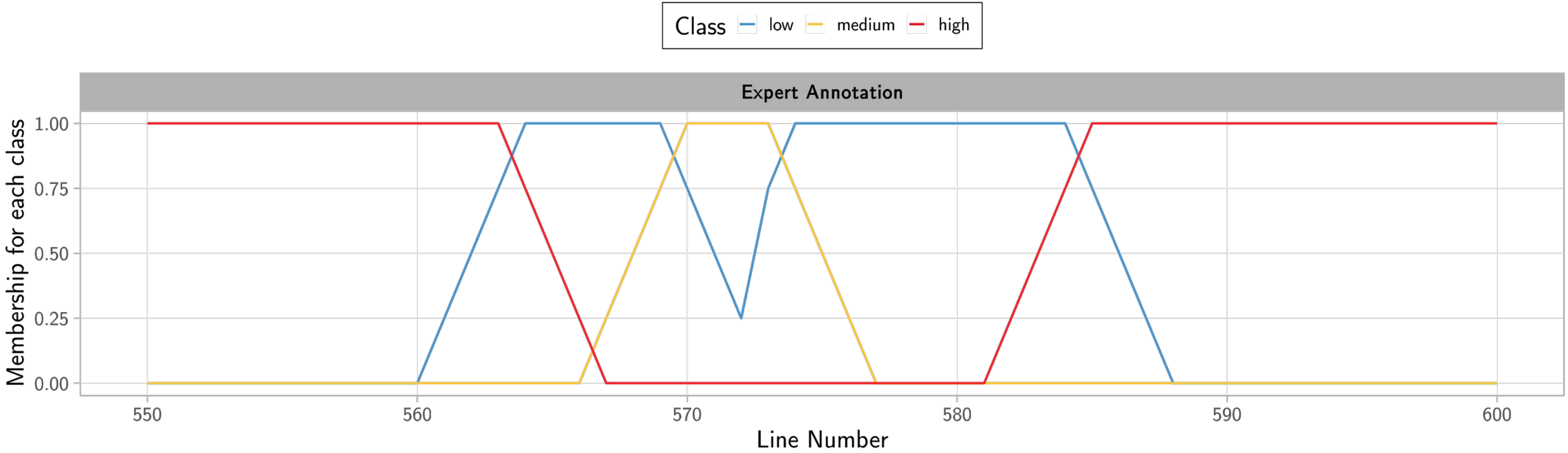
Uniformly increase membership for the 3 preceding lines, decrease for 3 succeeding.

Chat Message	**Crisp Label**	**Fuzzy Representation**
Message 1	**medium**	[**0.0**, 1.0, **0.5**]
Message 2	**medium**	[**0.0**, 1.0, **0.75**]
Message 3	**high**	[**0.0**, 0.5, **1.0**]
Message 4	**high**	[**0.0**, 0.75, **1.0**]
Message 5	**medium**	[**0.0**, 1.0, **0.75**]

Fuzzification of Expert Labels

Crisp Labels of risk → Trapezoidal Membership Function

Uniformly increase membership for the 3 preceding lines, decrease for 3 succeeding.



Fuzzy Risk Detection Task

Given a chat line l and its fuzzy representation of risk level,

$$\mu(l) = [\mu_{low}(l), \mu_{medium}(l), \mu_{high}(l)]$$

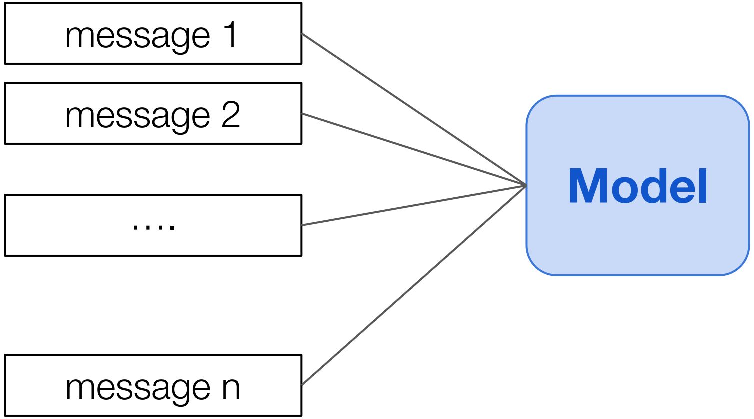
Learn a model m that can accurately estimate $\mu(l)$.

Fuzzy Risk Detection Task

Given a chat line *l* and its fuzzy representation of risk level,

$$\mu(l) = [\mu_{low}(l), \mu_{medium}(l), \mu_{high}(l)]$$

Learn a model *m* that can accurately estimate $\mu(l)$.

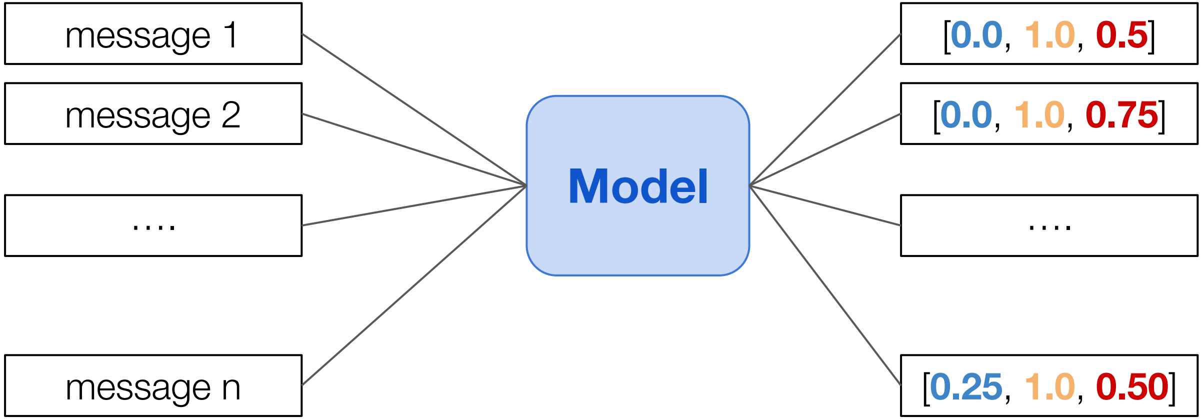


Fuzzy Risk Detection Task

Given a chat line *l* and its fuzzy representation of risk level,

$$\mu(l) = [\mu_{low}(l), \mu_{medium}(l), \mu_{high}(l)]$$

Learn a model *m* that can accurately estimate $\mu(l)$.



Experimental Setup

1. **Data**
    a. **13648 lines** comprising of **8** online conversations.
    b. Split in terms of separate chats
        i. **11900** lines (6 Conversations) as **Train**.
        ii. **977** lines (1 Conversation) as **Validation**.
        iii. **771** lines (1 Conversation) as **Test**.
2. **Training**
    a. Train 2 simple models as baselines to estimate the fuzzified risk level of each message.
    b. Select best model with highest metric on validation set.
3. **Evaluating**
    a. Evaluate on test set (A full conversation) using metric.

Baseline Models

Competitive Baselines in NLP literature for short sentence classification tasks.

1. Deep Averaging Network (Iyyer et al. 2015):
    a. Sentence representation is composed of an average of each of the individual word vectors.
    b. A FeedForward Layer on top of the sentence representation can help establish a very simple baseline.
2. Convolutional Neural Networks for Sentence Classification (Kim 2014):
    a. Sentence representation composed of running multiple width convolutions over the word vectors and max-pooling.
    b. Typically uses 2 channels, one with pre-trained representations (frozen) and one without (to be trained).

Baseline Models - Word Vector Initialization

Used **fasttext** embedding (Bojanowski et al. 2016) as the input to the models.

promice = <pr + pro + rom + omi + mic + ice + ce> + <pro + prom + romi + omic + mice + ice> + <prom + promi + romic + omice + mice> + <promi + promic + romice + omice>

promise = <pr + pro + rom + omi + mis + ise + se> + <pro + prom + romi + omis + mise + ise> + <prom + promi + romis + omise + mise> + <promi + promis + romise + omise> + promice

Baseline Models - Word Vector Initialization

Used **fasttext** embedding (Bojanowski et al. 2016) as the input to the models.

promice = **<pr** + **pro** + **rom** + **omi** + mic + ice + ce> + **<pro** + **prom** + **romi** + omic + mice + ice> + **<prom** + **promi** + romic + omice + mice> + **<promi** + promic + romice + omice>

promise = **<pr** + **pro** + **rom** + **omi** + mis + ise + se> + **<pro** + **prom** + **romi** + omis + mise + ise> + **<prom** + **promi** + romis + omise + mise> + **<promi** + promis + romise + omise>

Deep Averaging Network (Iyyer et al. 2015)

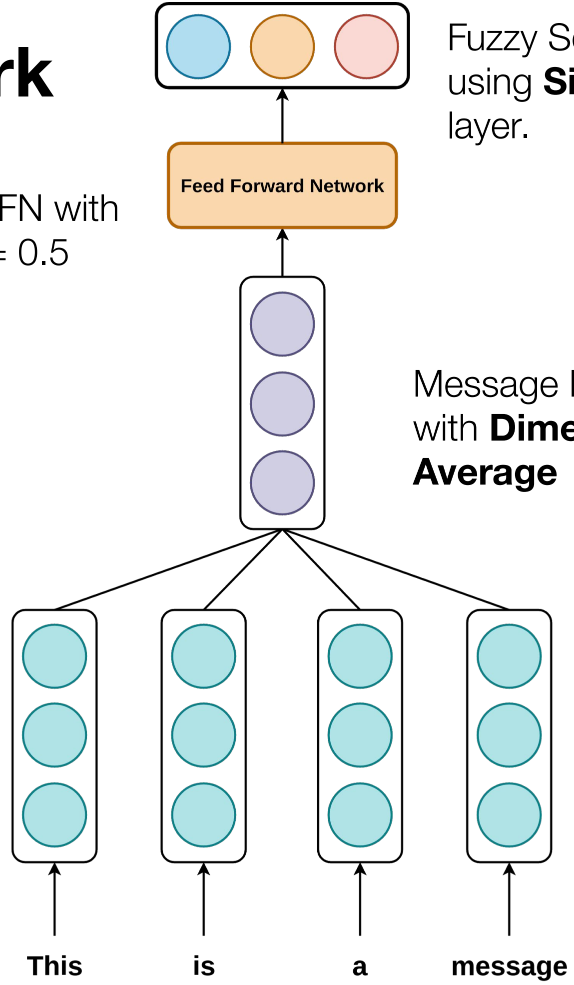
Rivals LSTMs as a strong baseline, especially for smaller datasets.

Fuzzy Set Estimation using **Sigmoid** in the last layer.

2 Layer FFN with dropout = 0.5

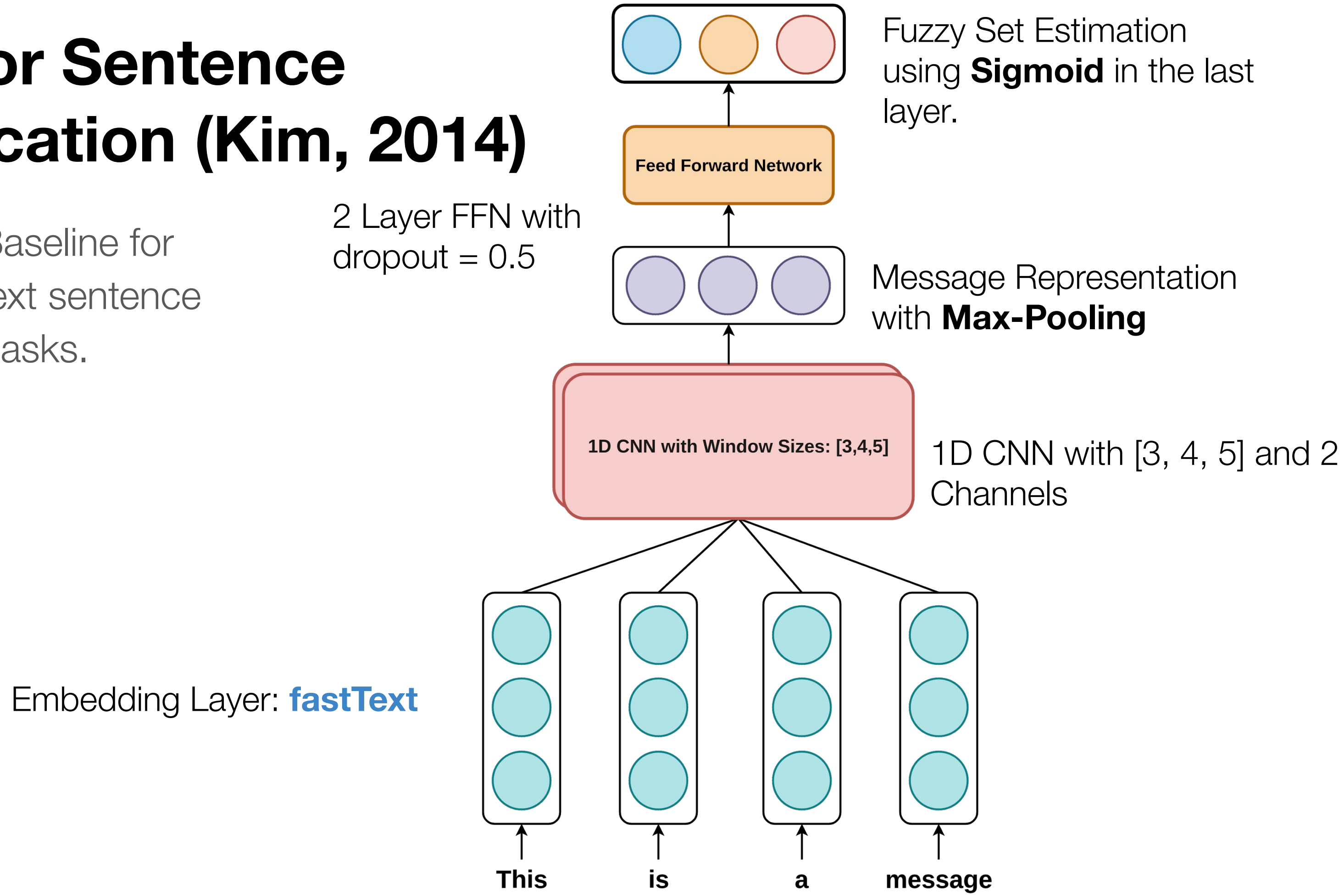
Message Representation with **Dimension wise Average**

Embedding Layer: **fastText**



CNNs for Sentence Classification (Kim, 2014)

Competitive Baseline for Small, short text sentence classification tasks.



Convolutional Neural Network Refresher

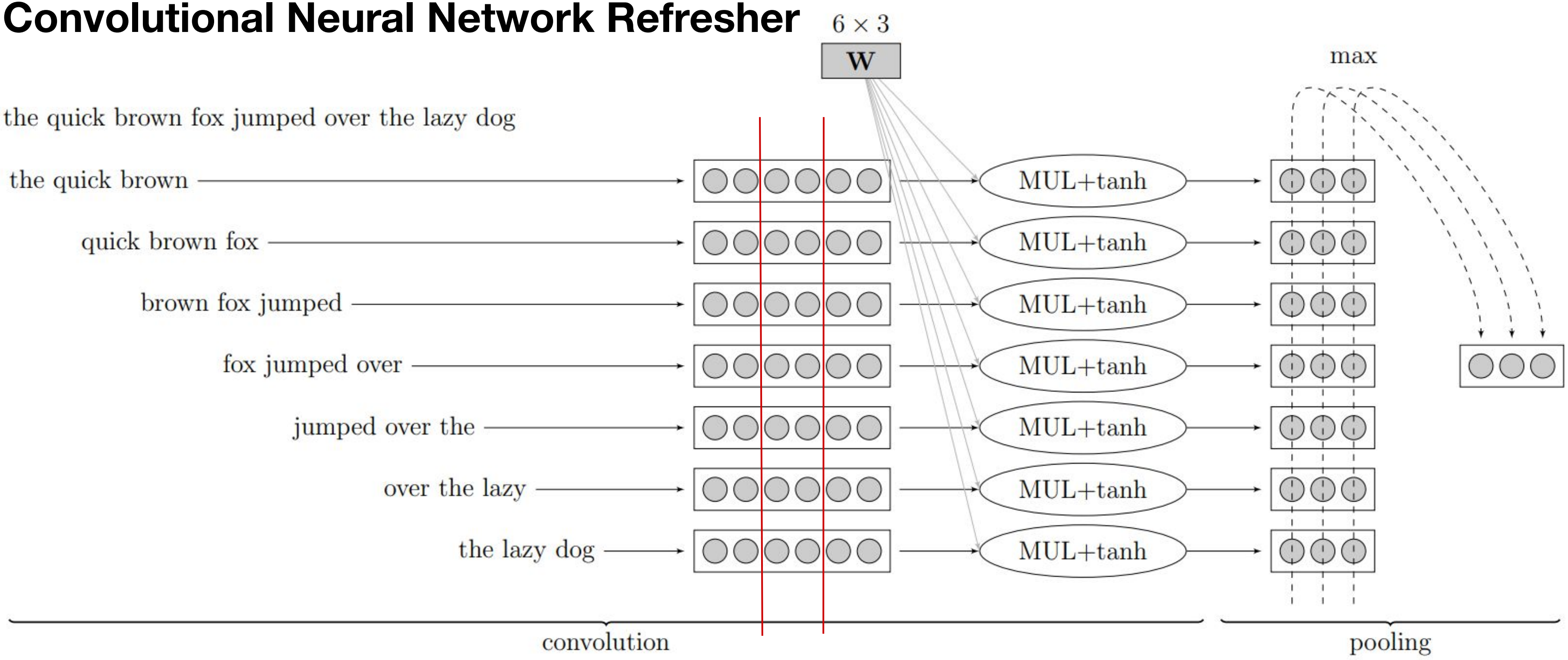


Figure Source: *A Primer on Neural Network Models for Natural Language Processing,* Yoav Goldberg

Baseline Models - Loss Function

L1 Loss along each position of [low, medium, high]

$$L = \sum_{i \in C} |\mu_i(l) - \hat{y}_i|$$

Baseline Models - Loss Function

L1 Loss along each position of [low, medium, high]

$$L = \sum_{i \in C} |\mu_i(l) - \hat{y}_i|$$

Truth = [0.00, 0.75, 1.00]

Predicted = [0.01, 0.45, 0.89]

L1 Loss = 0.01 + 0.30 + 0.11 = 0.42

Evaluation Metric - Fuzzy Jaccard Similarity

Jaccard Similarity = Similarity between two sets, A and B.

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

Evaluation Metric - Fuzzy Jaccard Similarity

Using Fuzzy versions of A∩B and A∪B, and the cardinality | A |, The fuzzy jaccard similarity is:

$$J_{Fuzzy}(A,B) = \frac{\sum_{i \in C} \min\{\mu_i(A), \mu_i(B)\}_{i \in C}}{\sum_{i \in C} \max\{\mu_i(A), \mu_i(B)\}_{i \in C}}$$

Evaluation Metric - Fuzzy Jaccard Similarity

Using Fuzzy versions of A∩B and A∪B, and the cardinality | A |, The fuzzy jaccard similarity is:

$$J_{Fuzzy}(A,B) = \frac{\sum_{i \in C} \min\{\mu_i(A), \mu_i(B)\}_{i \in C}}{\sum_{i \in C} \max\{\mu_i(A), \mu_i(B)\}_{i \in C}}$$

A = [0.00, 0.75, 1.00]

B = [0.21, 0.95, 0.89]

$\mathbf{J_{Fuzzy}}$ = (0.00 + 0.75 + 0.89)/(0.21 + 0.95 + 1.00) = 0.759

Results

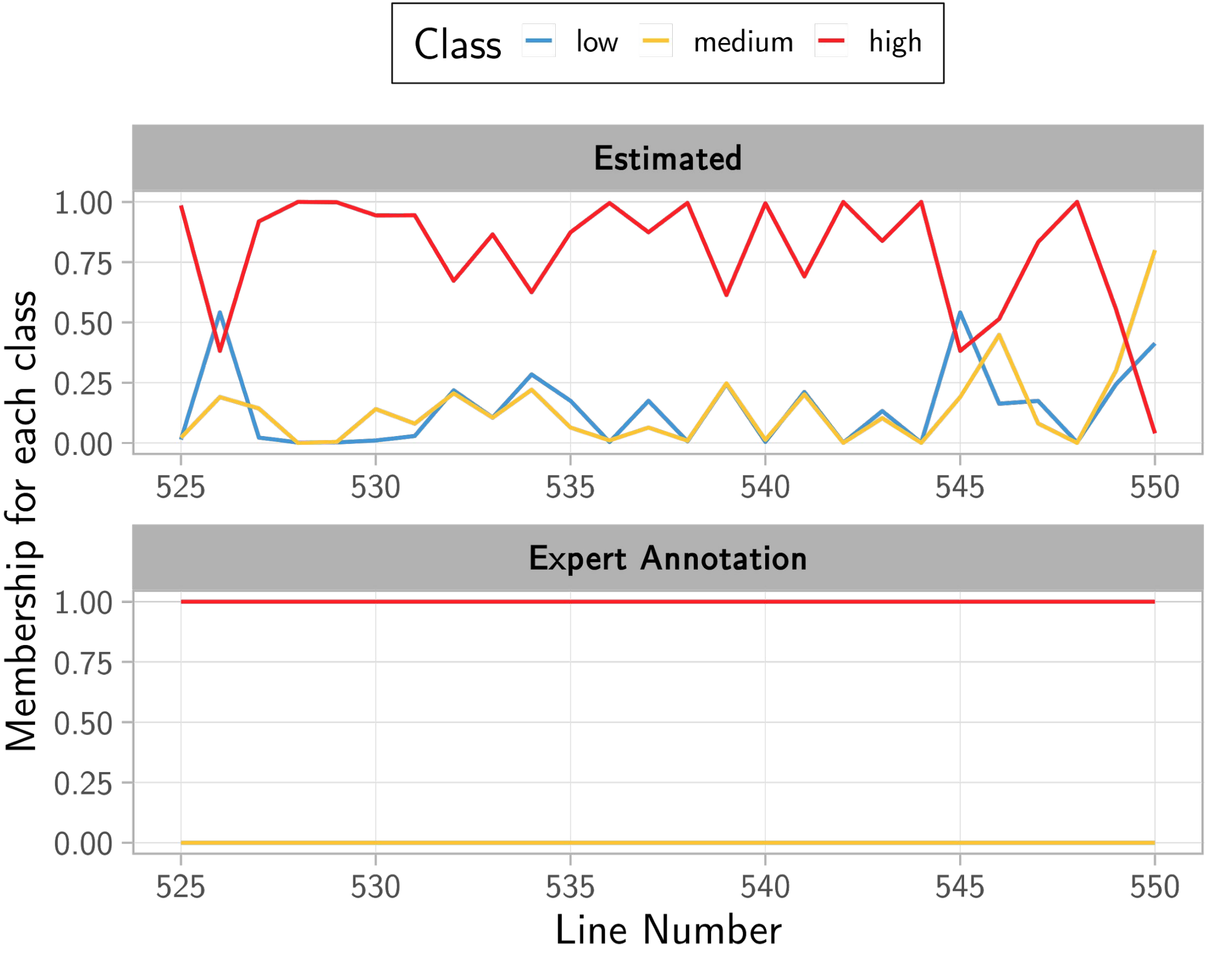
Model	**Epochs**	**Parameters**	J_{fuzzy}
DAN	1000	~30k	0.380
CNN	100	~1.4m	0.455

Results

Model	**Epochs**	**Parameters**	$\boldsymbol{J_{fuzzy}}$
DAN	1000	~30k	0.380
CNN	100	~1.4m	0.455
...	...	...	...
(Your Model)	...	?	?

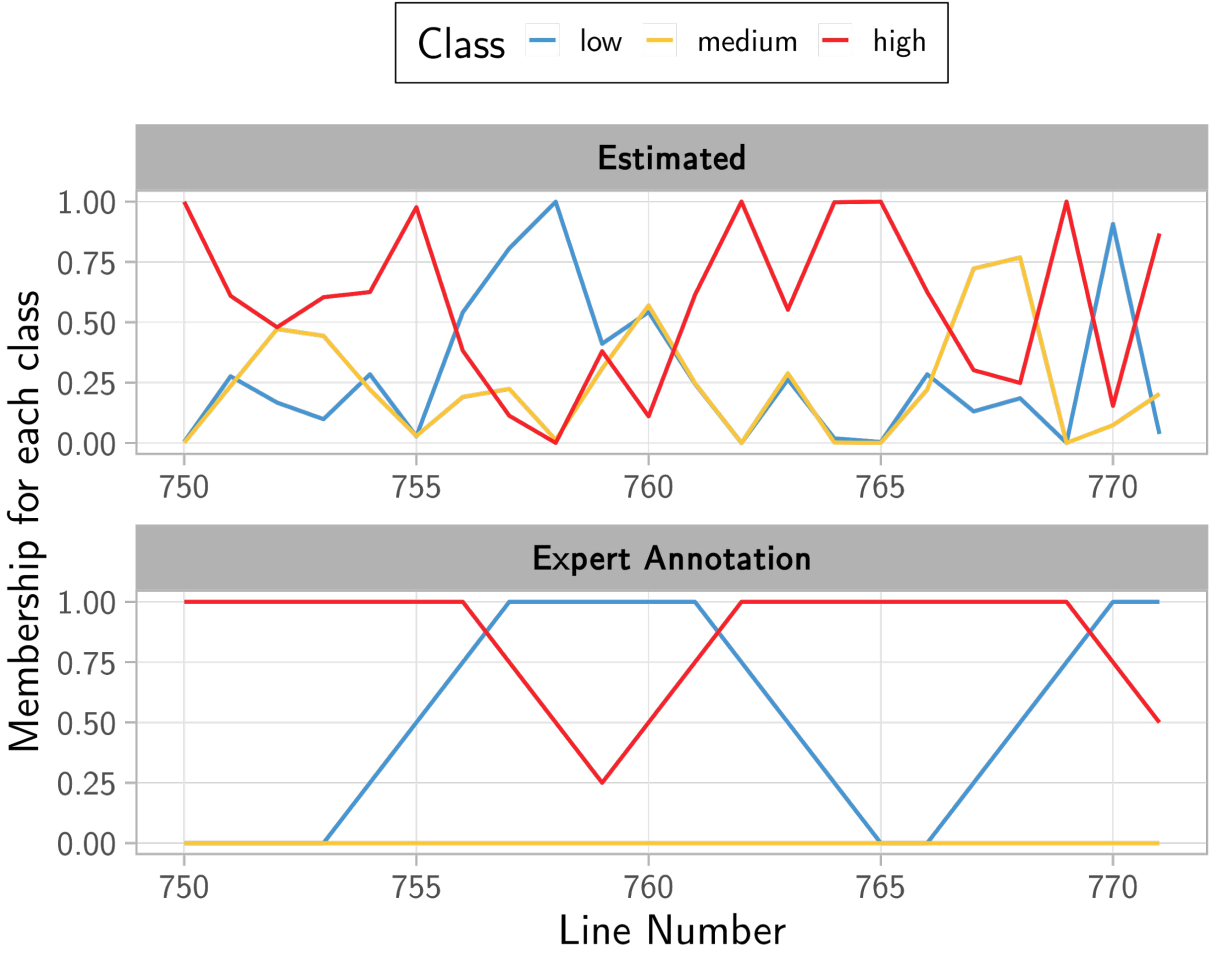
Results

The model learns some trivial properties, such as **continuous flow of highly risky messages**.



Results

The model also learns certain less trivial properties, such as **transitions between risk level**.



Conclusion

- Presented a methodology to quantify risk as a Fuzzy rather than Crisp phenomenon.
- Proposed simple baselines that provided modest performance (based on our evaluation metric).
- The models tend to capture many patterns that agree with the grooming literature.
    - It tends to capture continuous flow of risk level.
    - It tends to capture certain transitions between high and low risk.

Future Work

1. **Obvious:** Label more data to test more complex models such as Transformers, etc.
2. **Is low/medium/high enough?** Label for grooming events/strategies →WIP by Tatiana (First Author).
3. Test with other membership functions for **dynamic transition stages**.
4. Maintain overall discourse by ***remembering*** previous chat messages.
5. **Fuzzy loss functions**?

Thank You!

Questions?

Kanishka - @iamasharkskin

trigenb@purdue.edu
kmisra@purdue.edu
jtaylor1@purdue.edu

Coming soon..

References

[1] O'Connell, R. (2003). A typology of child cybersexploitation and online grooming practices. *Preston, UK: University of Central Lancashire*.

[2] Black, P. J., Wollis, M., Woodworth, M., & Hancock, J. T. (2015). A linguistic analysis of grooming strategies of online child sex offenders: Implications for our understanding of predatory sexual behavior in an increasingly computer-mediated world. *Child Abuse & Neglect*, *44*, 140-149.

[3] Cano, A. E., Fernandez, M., & Alani, H. (2014, November). Detecting child grooming behaviour patterns on social media. In *International conference on social informatics* (pp. 412-427). Springer, Cham.

[4] Michalopoulos, D., & Mavridis, I. (2011, June). Utilizing document classification for grooming attack recognition. In *2011 IEEE Symposium on Computers and Communications (ISCC)* (pp. 864-869). IEEE.

[5] McGhee, I., Bayzick, J., Kontostathis, A., Edwards, L., McBride, A., & Jakubowski, E. (2011). Learning to identify internet sexual predation. *International Journal of Electronic Commerce*, *15*(3), 103-122.

[6] Parapar, J., Losada, D. E., & Barreiro, A. (2012). A Learning-Based Approach for the Identification of Sexual Predators in Chat Logs. In *CLEF (Online Working Notes/Labs/Workshop)* (Vol. 1178).

[7] Ebrahimi, M., Suen, C. Y., & Ormandjieva, O. (2016). Detecting predatory conversations in social media by deep convolutional neural networks. *Digital Investigation*, *18*, 33-49.