Osprey: A Reference Framework for Online Grooming Detection via Neural Models and Conversation Features[*]

Hamed Waezi
University of Windsor, ON., Canada
waezi@uwindsor.ca

Reza Barzgar
University of Windsor, ON., Canada
barzgar@uwindsor.ca

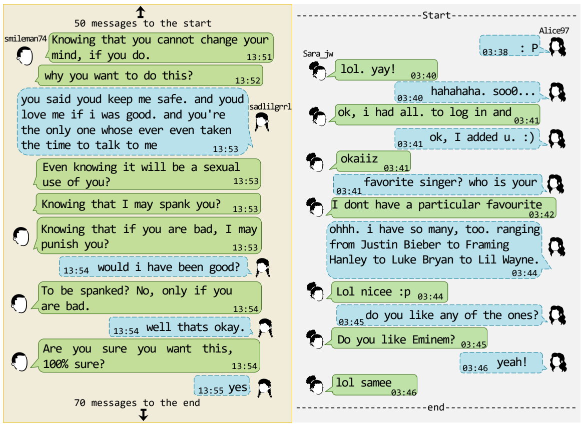
Hossein Fani
University of Windsor, ON., Canada
hosseinfani@gmail.com

ABSTRACT

Online grooming is the process of an adult initiating a sexual relationship with a minor through online conversation platforms. While neural networks are developed to detect such incidents, their practical implications in real-world settings remain moot for their closed, irreproducible, and poor evaluation methodologies under the sparse distribution of grooming conversations in the training datasets, like undermining `recall` over `precision`. Furthermore, proposed neural models overlook characteristic features of grooming in online conversations, including the number of participants, message exchange patterns, and temporal signals, such as the elapsed times between messages. In this paper, we foremost contribute `Osprey`, an open-source benchmark library to support a standard pipeline and experimental details, incorporating canonical neural models and a variety of vector representation learning for conversations while accommodating new models and training datasets. Further, we incorporate conversation features into the models to substantially improve `recall` while maintaining `precision`. Our experiments across neural baselines and vector representations of conversations demonstrated that recurrent neural models, particularly `gru`, on the sequence of pretrained transformer-based embeddings of messages in a conversation along with conversation features obtain state-of-the-art performance, winning the best `recall` while maintaining competitive `precision`. `Osprey` is available at `https://github.com/fani-lab/Osprey`.

1 INTRODUCTION

With the prevalence of more technology, minors access it before they are of legal age and with little cognitive development [18], facing an alarming problem of engaging with predators in online grooming [36]. Through grooming, the sexual predator tries to form an emotional relationship with a minor to get her trust and make her engage in sexual activities afterwards [14, 35, 37]. Recent stats show that 57% of girls and 48% of boys have experienced at least one online grooming, with some regions like north america and western europe being even higher [30]. Meanwhile, crimes involving minors are underreported for lack of awareness, support, or trust in authorities, fear of retaliation from the predator or legal repercussions, and distress of being judged or blamed, among others [12, 38]. Such proceedings have stressed the development of computational models that plug into an online chat environment and help warn minors, parents or police of such incidents while preserving the minor's privacy [8].



50 messages to the start
smileman74: Knowing that you cannot change your mind, if you do. 13:51
why you want to do this? 13:52
sadlilgrrl: you said youd keep me safe. and youd love me if i was good. and you're the only one whose ever even taken the time to talk to me 13:53
Even knowing it will be a sexual use of you? 13:53
Knowing that I may spank you? 13:53
Knowing that if you are bad, I may punish you? 13:53
13:54 would i have been good?
To be spanked? No, only if you are bad. 13:54
13:54 well thats okay.
Are you sure you want this, 100% sure? 13:54
13:55 yes
70 messages to the end

--------------------Start--------------------
Alice97: 03:38 : P
Sara_jw: lol. yay! 03:40
03:40 hahahaha. soo0...
ok, i had all. to log in and 03:41
03:41 ok, I added u. :)
okaiiz 03:41
03:41 favorite singer? who is your
I dont have a particular favourite 03:42
ohhh. i have so many, too. ranging from Justin Bieber to Framing Hanley to Luke Bryan to Lil Wayne. 03:44
Lol nicee :p 03:44
03:45 do you like any of the ones?
Do you like Eminem? 03:45
03:46 yeah!
lol samee 03:46
--------------------end--------------------

Figure 1: A predatory conversation between a 34-year-old predator and a 13-year-old victim (left) vs. a normal one.

In many cases of online grooming, as shown in Figure 1, the predators mix explicit textual remarks to give the minor a feeling of endearment and to lure her into their trap while going through several stages [14]: *i*) targeting emotionally vulnerable minors who are facing problems in their life (e.g., userid *'sadlilgrrl'*); *ii*) gaining trust by making the minor feel safe (e.g., *'why you want to do this?'* or *'are you sure you want this, 100% sure?'*); *iii*) filling a void by being affectionate; *iv*) isolating the minor by giving more attention and forming a bond; *v*) sexualizing the relationship by asking for inappropriate photographs in situations where the minor is naked like swimming trips; and *vi*) maintaining control and domination. Such remarks can be extracted by natural language processing techniques and tapped into machine learning models to detect predators or predatory conversations [22]. To this end, researchers have formulated online grooming detection problem as a Boolean classification task and have tackled it at two granularity levels of *i*) message or *ii*) entire conversation through machine learning models including support vector machine (svm) [4, 7, 9, 11, 40], logistic regression [6, 7, 25], knn [6, 31], naïve Bayes [2, 25, 39], decision trees [7, 23, 25], and recently neural models, including feedforward [7, 11, 40], convolutional [9], and recurrent [19, 29] neural networks, and transformers [41]. In predatory message classification, a message in a conversation is mapped into a sparse vector representation like an occurrence vector of tokens [7, 9–11, 27, 40] or a dense low-dimensional vector using distributional vector representation learning (embedding) techniques [3, 9, 28] like `glove` [32] and `word2vec` [26], or contextual pretrained vectors from large language models like `bert` [6, 19, 41] which is further classified into a Boolean value of *'predatory'* or *'normal'*. The proposed works for predatory message classification, however, overlooked the conversation context for an input message such as the messages before and after. To fill the gap, Kim et al. [19] fed messages of a conversation in sequence to `lstm` followed by a final classifier and obtained state-of-the-art performance.

[*] **Warning:** This paper discusses online grooming that may be offensive or upsetting.

Successful as they are, proposed neural methods forego distinctive features of a predatory conversation, including *i*) the number of participants in a conversation; little to no predatory conversation has more than two participants, *ii*) the message exchange pattern; predatory conversations lack a fair exchange of messages through turn-taking, and *iii*) the elapsed times of the message exchanges. From `pan` benchmark dataset [17, 18], as seen in Table 1, all predatory conversations are one-on-one; that is, only two participants are involved. Further, the average number of consecutive messages from the same participant, referred to as a *segment*, for predatory conversations and normal ones are `2.12±0.64` vs. `2.00±2.50`, respectively, and the average time elapsed for message exchange between a predator and a minor is `1.43±8.61` compared to `0.91±1.41` in normal conversations.

Moreover, to the best of our knowledge, no proposed method is publicly available regarding implementation and experimental setup except that of Vogt et al. [41][1], which includes only their proposed method falling short of being a benchmark library. With regard to the evaluation methodology, existing work reported `accuracy`, which is misleading in online grooming detection where training datasets are highly skewed toward normal conversations, and predatory conversations are very scarce; e.g., in `pan`, merely `2.3%` of conversations are predatory. Regarding other metrics, studies forego the critical precedence of `recall` and the dire consequences of misclassifying even `1` online grooming in the real world.

In this paper, foremost, (1) we contribute `Osprey`, an open-source reference benchmark library to foster reproducibility, and sound evaluation methodology, which literature of this field lacks. `Osprey` sets forth a variety of sparse and dense vector representations of conversations at both levels of the message and the entire conversation along with neural and non-neural models with a one-click pipeline that orchestrates the standard flow of machine learning benchmark with no human in the loop. The pipeline accepts a set of models and brings it through a *stratified* cross-fold train-validation stage, followed by test and evaluation on an unseen test set with a special focus on `recall`, and receiver operating characteristic (`roc`) and `precision-recall` curves as a recognized evaluation strategy in the presence of class imbalances. Leveraging an object-oriented structure, `Osprey` readily accommodates the addition of new training datasets, vector representation techniques, and classifiers. (2) We then propose to incorporate features from the conversation context, including message timestamps and the number of participants to improve the efficiency of neural models during training while improving inference effectiveness. (3) Moreover, through a cross-study of a variety of vector representations at levels of a message or the entire conversation with conversation features and lack thereof on neural and non-neural models, we systematically addressed research questions that are yet to be answered in online grooming detection, including *i*) conversation features show synergistic impact across all models and all varieties of vector representations, and more importantly, they help models prioritize `recall` while maintaining `precision`, *ii*) recurrent models outperform other baselines across all varieties of vector representations; specifically, `gru` outperforms `lstm` [19] and yields state-of-the-art performance, and *iii*) the sequence of pretrained *contextual* vectors of enclosing messages yields the best results.

Table 1: Statistics of `pan` [17] dataset.

	raw		filtered	
	train	**test**	**train**	**test**
#conversations	66,927	155,128	16,529	38,246
#***predatory*** conversations	2,016	3,737	957	1,698
#conversations w/ 1 participant	12,773	29,561	**0**	**0**
#conversations w/ 2 participants	45,741	105,862	9,474	21,722
#binary ***predatory*** conversations	2,016	3,737	957	1,698
#non-binary ***predatory*** conversations	**0**	**0**	**0**	**0**
avg #msgs in a ***predatory*** conversation	**60.73**	**90.07**	**80.68**	**71.48**
avg #msgs in a normal conversation	**12.74**	**12.86**	**41.73**	**41.78**

2 PROBLEM DEFINITION

A conversation c is a sequence of $|c|$ messages $[m_i]$; $1 < i \leq |c|$, such that each message includes `id`, `text`, `participant`, and `timestamp`. Furthermore, as opposed to an online post or comment, an online conversation should have at least two different participants, each of whom has at least one message, i.e., $\exists m_i, m_j \in c, i \neq j$ such that m_i.`participant` $\neq m_j$.`participant`. Given $C = \{c\}$ be the set of conversations, online grooming detection is to learn $f_\theta : C \rightarrow \{0 : \text{normal}, 1 : \text{predatory}\}$, a mapping function of parameters θ from the conversation set to the Boolean set, such that $f_\theta(c) = 1$ if c is predatory and `0` otherwise. We estimate f_θ based on different vector representations of a conversation in a d-dimensional space through a function $g_\phi : c \rightarrow \mathcal{R}^d$ such that $f_\theta(c) \approx f_\theta(g_\phi(c))$.

3 CONVERSATION FEATURES

Predatory conversations embody distinctive characteristics; almost *all* are one-on-one conversations, lack a fair distribution of messages between participants via turn-taking, and have long elapsed times in message exchanges, the capture of which presumably improve the performance of online grooming detection methods. In this paper, we propose incorporating the timestamp of a message and the number of participants to improve upon neural and non-neural estimators for f_θ. Specifically, we cross-examine (1) neural models trained on (2) a variety of conversation vector representation methods g_ϕ(c), (3) fused with conversation features, including message timestamp and the number of participants via (4) `Osprey`'s unified framework followed by (5) the *stratified* k-fold cross-validation training and test evaluation methodology, considering the highly skewed distributions of classes.

To estimate f_θ, we integrated vanilla (`rnn`) and gated recurrent neural models, including `lstm` and `gru` as the cutting-edge class of approaches to learn from the sequence of messages within the context of a conversation followed by a final classification. We also used feedforward and non-neural classifiers to classify the entire conversation at once. We vectorize conversations at two levels:

Message-level Embeddings. Given a conversation $c = [m_i]_{i=1}^{|c|}$, we map each message m_i onto a vector using bag-of-word, distributional, and contextual embeddings, as our g_ϕ, such that $g_\phi(c) \approx [g_\phi(m_i)]_{i=1}^{|c|}$, that is, a conversation becomes a list of vector representations of its constituents messages. Distributional and contextual embeddings can be transferred from a pretrained model on an external corpus, finetuned, or trained from scratch on an online grooming dataset. For distributional embeddings, we used `word2vec` [26, 32] to map the message's words onto embeddings

[1] `gitlab.com/early-sexual-predator-detection/eSPD-lab`

Table 2: **Comparative results of 3-fold train-valid models with conversation features and lack thereof on pan's test set. Bold and underlined numbers indicate the *column-wise* highest and second highest, respectively.**

	aucroc			f_2 (favours recall)			$f_{0.5}$ (favours precision)		
	−ctx	+ctx	Δ	−ctx	+ctx	Δ	−ctx	+ctx	Δ
conversation as a sequence of message embeddings ($c = [g_\phi(m_i)]_{i=1}^{\lvert c\rvert}$)									
bow-rnn	54.04±01.16	58.40±02.56	+04.36	10.65±00.92	12.38±02.22	+01.73	03.30±00.09	05.01±01.51	+01.71
word2vec-rnn	54.65±04.25	57.62±01.92	+02.97	10.98±01.71	11.73±00.37	+00.75	10.37±09.49	18.23±10.38	+07.87
word2vec†-rnn	57.97±01.46	65.13±06.27	+07.16	12.91±00.60	14.98±03.90	+02.06	11.59±10.55	11.30±08.82	−00.28
roberta-rnn	59.11±02.79	67.01±05.83	+07.90	11.69±00.41	12.87±01.89	+01.18	8.06±06.53	10.72±09.79	+02.66
roberta†-rnn	54.75±00.41	55.79±00.44	+01.04	10.80±00.30	11.22±00.22	+00.42	11.48±05.58	22.85±00.15	+11.37
bow-lstm	94.36±01.89	96.24±00.16	+01.88	62.44±03.36	60.69±04.75	−01.75	62.48±05.05	60.35±05.23	−02.13
word2vec-lstm	91.79±03.37	90.51±03.42	−01.28	53.38±00.96	54.13±02.65	+00.74	59.54±03.32	55.76±07.37	−03.77
word2vec†-lstm	90.96±00.59	93.02±01.52	+02.05	47.20±02.66	51.69±01.63	+04.50	34.23±06.31	45.91±09.76	+11.68
roberta-lstm [19]	96.08±00.28	97.98±00.48	+01.90	64.07±04.28	70.74±01.89	+06.68	63.69±04.38	64.43±03.14	+00.74
roberta†-lstm	81.95±13.88	93.87±02.53	+11.91	39.23±19.27	49.22±02.42	+09.99	37.43±18.14	29.14±13.06	−08.29
bow-gru	96.69±00.54	97.09±00.74	+00.40	67.95±01.38	62.75±02.55	−05.21	48.84±02.42	67.20±03.90	+18.36
word2vec-gru	95.38±00.85	96.00±00.45	+00.62	55.63±02.13	54.66±02.17	−00.97	51.35±06.13	50.63±10.51	−00.72
word2vec†-gru	92.18±00.58	92.98±00.39	+00.80	48.49±01.49	47.91±01.49	−00.58	50.71±03.21	45.52±13.83	−05.19
roberta-gru	97.04±00.82	**98.29**±00.22	+01.25	67.30±00.97	**74.00**±01.25	+06.71	59.04±06.97	54.87±03.90	−04.17
roberta†-gru	95.53±01.18	97.09±00.39	+01.56	45.66±17.29	63.80±04.22	+18.15	23.16±12.93	38.89±07.76	+15.73
conversation as a single embedding ($c = g_\phi(m_c^*)$)									
bow-svm [40]	49.97±00.01	50.10±00.06	+00.13	00.03±00.00	00.28±00.15	+00.24	00.12±00.00	01.06±00.58	+00.94
word2vec-svm	50.23±00.02	82.26±00.31	+32.02	00.59±00.04	61.77±00.17	+61.18	02.28±00.15	51.42±01.43	+49.13
word2vec†-svm	50.02±00.01	50.04±00.02	+00.03	00.04±00.02	00.11±00.04	+00.07	00.18±00.06	00.44±00.16	+00.26
roberta-svm	82.49±00.26	82.78±00.42	+00.29	65.92±00.24	66.83±00.55	+00.91	**66.31**±00.91	**68.51**±00.60	+02.20
roberta†-svm	83.85±00.81	55.48±02.07	−28.36	62.18±00.23	13.18±04.57	−49.00	46.75±02.16	24.33±04.56	−22.43
bow-ff [9]	48.50±00.43	96.17±00.21	+47.67	04.00±00.42	68.39±00.39	+64.38	01.88±00.16	39.73±00.65	+37.85
word2vec-ff	48.69±06.92	95.89±00.49	+47.20	11.03±01.04	59.10±00.51	+48.08	03.04±00.32	29.80±00.76	+26.77
word2vec†-ff	91.46±00.21	92.52±00.79	+01.06	46.55±03.12	48.70±01.88	+02.16	21.33±02.82	22.54±01.88	+01.21
roberta-ff	**97.84**±00.16	98.25±00.00	+00.41	**68.20**±01.00	70.33±00.69	+02.13	44.20±02.76	47.29±02.84	+03.09
roberta†-ff	93.77±00.44	94.88±00.45	+01.11	56.73±06.46	58.99±00.62	+02.27	32.78±11.55	30.39±00.53	−02.39

whose average yields the embedding for the message. For contextual embeddings, we used roberta [20] to map the message onto an embedding. Our choices of distributional and contextual embedding methods are without loss of generality to other recent and advanced transformer-based language models. Yet, our goal is to study the effect of different embedding types using pioneering methods, distributional vs. contextualized embeddings, toward grooming detection; a better model most likely shows similar findings. The sequence of message-level embeddings for a conversation, each of which is concatenated with their timestamp and the number of participants to signal the unique features of predatory conversations, is fed to the recurrent neural models in order to classify the entire conversation as predatory or else.

Conversation-level Embeddings. To represent a conversation c as an embedding, we concatenate all its constituent messages into a long synthetic message, that is, $c \approx m_c^* = m_1 : \ldots : m_i : \ldots : m_{|c|}$, and apply one of our message-level embedding methods such that $g_\phi(c) \approx g_\phi(m_c^*)$. Herein, we add only the number of participants since the entire conversation becomes a single synthetic message. Such embeddings are fed to feedforward or non-neural classifiers.

4 EXPERIMENTS

In this section, we lay out the details of our experiments and findings aimed at addressing the following research questions:

RQ1: Does the inclusion of conversation features improve the efficacy of online grooming detection across models and different vector representations of conversations?

RQ2: Does representing conversations as sequences of message-level embeddings yield better model performance than single embedding for the entire conversation?

RQ3: Do methods of online grooming detection prioritize recall while maintaining precision?

RQ4: Amongst the recurrent neural model, which gating strategy yields the best performance?

4.1 Setup

4.1.1 Dataset. Despite the significant societal benefit of computational models for online grooming detection, access to training sets of online conversations has remained challenging due to privacy and legal concerns. All few datasets in prior work, including collection of conversations from an online multiplayer game for minors [7], chat-coder [23] and pan-chat-coder [41] are *in*accessible, to the best of our and other researcher's effort, except that of pan [17], which has become the sole benchmark dataset in the literature [1, 2, 5, 9, 15, 23, 24]. In pan, cases of online grooming were obtained from trained volunteers (decoys) posing as minors in public chatrooms to catch and convict predators[13] and the normal conversations are from omegle[21] online chatrooms. For our experiments, we filtered out the conversations with 1 participant or an exchange of fewer than 6 messages between participants. Table 1 shows the dataset's statistics. As seen, the number of predatory conversations with more than two participants is 0, signifying predators approach their victims in private. Also, predatory conversations are notably longer, extending to around 80 messages on average. Such characteristics can be leveraged by a model to improve the performance of online grooming detection.

4.1.2 Baselines. In this paper, we benchmarked Osprey using svm as a non-neural and a feedforward neural classifier (ff) when an entire conversation has been vectorized as a single embedding ($c = g_\phi(m_c^*)$), and vanilla (rnn), lstm [16], and gru recurrent neural models when a conversation is vectorized as a sequence of its

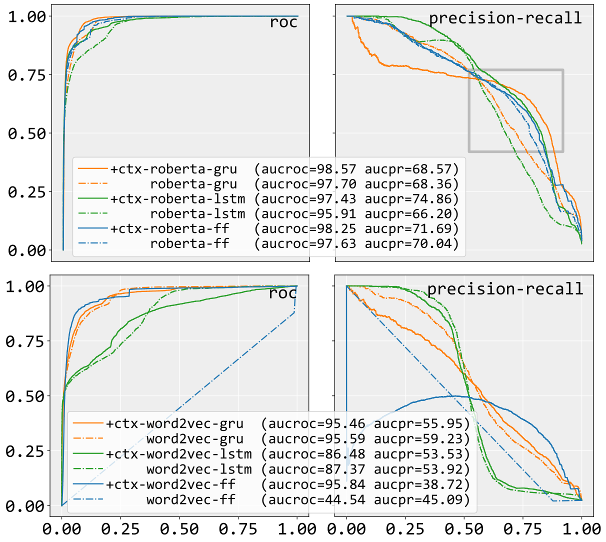


Figure 2: Efficacy at different classification thresholds.

messages' embeddings ($c = [g_\phi(m_i)]_{i=1}^{|c|}$). We used the rbf kernel for svm from scikit[34] with default values for hyperparameters. The feedforward neural model has a single layer of size 256 with relu activation function. The recurrent neural models have a single layer of size 512 with tanh as the activation function. Adam is the optimizer for all our neural models. To vectorize our conversations, we used bag-of-word (bow), 300-dimensional embeddings from pretrained word2vec on google news as the distributional vector representations, and 768-dimensional vectors from pretrained roberta-base on openwebtext [33] (roberta) as the contextual vector representations. We also trained word2vec (word2vec†) and finetuned roberta-base (roberta†) on pan. In total, we compare {5 models} × {5 embeddings} = 25 baselines.

4.1.3 Evaluation. Via 3-fold cross-validation procedure, we trained our baselines for 30 epochs and validated after each epoch for early stopping, which results in one trained model per each fold for a baseline. We finally evaluated the per-fold trained models of a baseline on the pan's test set and reported the recognized metrics under highly imbalanced class distributions, including roc and precision-recall curves along with f-measure with $\beta = 2.0$ to favour recall over precision vs. $\beta = 0.5$ vice versa.

4.2 Results

RQ1: Improvement upon inclusion of conversation features. From Table 2, we observe the general positive effects (+Δ) of including conversation features (+ctx) in all models in terms of aucroc, f_2, and $f_{0.5}$ compared to the lack thereof (-ctx). Specifically, substantial performance gains were obtained when conversation features were in tandem with the conversation as a single embedding. On a per-model basis, from a row-wise view, while recurrent baselines *-lstm and *-gru show the best performance overall, as expected, their gains from conversation features are marginal, which can be attributed to their ability to capture the conversation features through sequence processing of message embeddings. It is worth noting that even marginal improvement in online grooming detection, particularly in terms of f_2, leads to major positive societal contributions. On a per-metric basis, comparing f_2 and $f_{0.5}$, we see the overall positive impact of conversation features on both recall and precision and the few negatively impacted $f_{0.5}$ are negligible in favour of recall in online grooming detection.

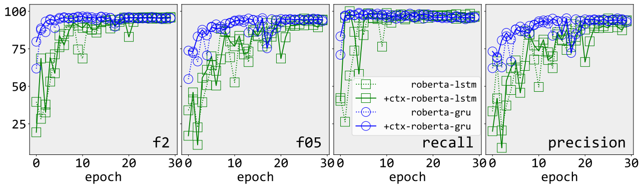


Figure 3: Efficacy vs. training efficiency of gru vs. lstm.

RQ2: Comparing message- vs. conversation-level embeddings. From Table 2, it is evident that conversations as sequences of embeddings for their constituent messages are consistently better representations than single embeddings for the entire conversations, enabling recurrent models to capture the message exchange patterns of online grooming and yield superior performance across metrics. Not unexpectedly, dense embeddings (roberta-* and word2vec-*) are better representations vs sparse bag-of-word vectors (bow-*), and contextual embeddings (roberta-*) are of higher quality compared to distributional embeddings (word2vec-*). In terms of pretrained vs. trained or finetuned embeddings, *oddly*, pretrained embeddings on external corpora yield better performance compared to training or finetuning on pan, which calls for more investigation.

RQ3: Prioritizing recall while maintaining precision. We draw the roc and precision-recall curves for strong baselines in Figure 2. Interestingly, we observe no baseline favours recall unless their inputs are augmented with conversation features. As indicated in Figure 2 (top), almost all neural models whose roberta embeddings are augmented with conversation features yield higher recall for competitive precision, which is the prime focus in online grooming detection. Further, from Figure 2, we see that while neural models on word2vec embeddings obtained higher recall, this came at the substantial cost of lower precision. Indeed, for *-word2vec-lstm, precision and recall are *inversely* correlated, rendering word2vec moot for conversation vectorization.

RQ4: Best gating strategy for recurrent neural models. Amongst recurrent models, from Table 2, we see that *-gru models are the best and *-lstm models [19] are runners up. Also, Figure 3 shows that *-roberta-gru outperforms *-roberta-lstm with much less training epochs. As already shown in the literature, vanilla recurrent models (*-rnn) are the poorest, even in comparison with feedforward and svm, which is attributed to their lack of gates and vanishing gradient problem; they fall short of retaining dependencies from earlier messages when processing long conversations as it is the case in predatory conversations.

5 CONCLUSIONS AND FUTURE WORK

In this paper, we addressed online grooming, a disturbing practice where adults initiate inappropriate relationships with minors on online chat platforms. We contributed Osprey, an open-source reproducible library that hosts conversation vectorizations along with conversation features and neural classifiers. We benchmarked distributional and contextual vector representations of conversations with conversation features and lack thereof via different neural models. Our results show that (1) conversation features have consistent synergistic effects across all baselines, (2) vectorizing a conversation through a sequence of message embeddings is of higher quality, (3) conversation features help models to prioritize `recall` while maintaining `precision`, and (4) `gru` is the best gating strategy for recurrent models in online grooming detection where the conversations are long and lack turn-taking. For future work, we are investigating the oddly low performance of finetuned embeddings as well as online grooming detection in other languages.

REFERENCES

[1] Mario Ezra Aragón and Adrián Pastor López-Monroy. 2018. A Straightforward Multimodal Approach for Author Profiling: Notebook for PAN at CLEF 2018. In *Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum, Avignon, France, September 10-14, 2018 (CEUR Workshop Proceedings, Vol. 2125)*, Linda Cappellato, Nicola Ferro, Jian-Yun Nie, and Laure Soulier (Eds.). CEUR-WS.org. https://ceur-ws.org/Vol-2125/paper_96.pdf

[2] Dasha Bogdanova, Paolo Rosso, and Thamar Solorio. 2012. On the Impact of Sentiment and Emotion Based Features in Detecting Online Sexual Predators. In *Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis, WASSA@ACL 2012, July 12, 2012, Jeju Island, Republic of Korea*, Alexandra Balahur, Andrés Montoyo, Patricio Martínez-Barco, and Ester Boldrini (Eds.). The Association for Computer Linguistics, 110–118. https://aclanthology.org/W12-3717/

[3] Parisa Rezaee Borj, Kiran B. Raja, and Patrick Bours. 2020. On Preprocessing the Data for Improving Sexual Predator Detection : Anonymous for review. In *15th International Workshop on Semantic and Social Media Adaptation and Personalization, SMAP 2020, Zakynthos, Greece, October 29-30, 2020*. IEEE, 1–6. https://doi.org/10.1109/SMAP49528.2020.9248461

[4] Patrick Bours and Halvor Kulsrud. 2019. Detection of Cyber Grooming in Online Conversation. In *IEEE International Workshop on Information Forensics and Security, WIFS 2019, Delft, The Netherlands, December 9-12, 2019*. IEEE, 1–6. https://doi.org/10.1109/WIFS47025.2019.9035090

[5] Claudia Cardei and Traian Rebedea. 2017. Detecting sexual predators in chats using behavioral features and imbalanced learning. *Nat. Lang. Eng.* 23, 4 (2017), 589–616. https://doi.org/10.1017/S1351324916000395

[6] Khaoula Chehbouni, Gilles Caporossi, Reihaneh Rabbany, Martine De Cock, and Golnoosh Farnadi. 2022. Early Detection of Sexual Predators with Federated Learning. In *Workshop on Federated Learning: Recent Advances and New Challenges (in Conjunction with NeurIPS 2022)*. https://openreview.net/forum?id=M84OnT0ZvDq

[7] Yun-Gyung Cheong, Alaina K. Jensen, Elin Rut Gudnadottir, Byung-Chull Bae, and Julian Togelius. 2015. Detecting Predatory Behavior in Game Chats. *IEEE Trans. Comput. Intell. AI Games* 7, 3 (2015), 220–232. https://doi.org/10.1109/TCIAIG.2015.2424932

[8] European Commission. 2022. *Fighting child sexual abuse: Commission proposes new rules to protect children.* https://ec.europa.eu/commission/presscorner/detail/en/ip_22_2976

[9] Mohammad Reza Ebrahimi, Ching Y. Suen, and Olga Ormandjieva. 2016. Detecting predatory conversations in social media by deep Convolutional Neural Networks. *Digit. Investig.* 18 (2016), 33–49. https://doi.org/10.1016/j.diin.2016.07.001

[10] Gunnar Eriksson and Jussi Karlgren. 2012. Features for Modelling Characteristics of Conversations. In *CLEF 2012 Evaluation Labs and Workshop, Online Working Notes, Rome, Italy, September 17-20, 2012 (CEUR Workshop Proceedings, Vol. 1178)*, Pamela Forner, Jussi Karlgren, and Christa Womser-Hacker (Eds.). CEUR-WS.org. https://ceur-ws.org/Vol-1178/CLEF2012wn-PAN-ErikssonEt2012.pdf

[11] Hugo Jair Escalante, Esaú Villatoro-Tello, Antonio Juárez, Manuel Montes-y-Gómez, and Luis Villaseñor Pineda. 2013. Sexual predator detection in chats with chained classifiers. In *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, WASSA@NAACL-HLT 2013, 14 June 2013, Atlanta, Georgia, USA*, Alexandra Balahur, Erik Van der Goot, and Andrés Montoyo (Eds.). The Association for Computer Linguistics, 46–54. https://aclanthology.org/W13-1607/

[12] David Finkelhor, Janis Wolak, and Lucy Berliner. 2001. Police reporting and professional help seeking for child crime victims: A review. *Child maltreatment* 6, 1 (2001), 17–30.

[13] Perverted-Justice Foundation. [n. d.]. *Perverted Justice Website.* http://www.perverted-justice.com

[14] Leah E. Kaylor Georgia M. Winters and Elizabeth L. Jeglic. 2017. Sexual offenders contacting children online: an examination of transcripts of sexual grooming. *Journal of Sexual Aggression* 23, 1 (2017), 62–76. https://doi.org/10.1080/13552600.2016.1271146 arXiv:https://doi.org/10.1080/13552600.2016.1271146

[15] Aditi Gupta, Ponnurangam Kumaraguru, and Ashish Sureka. 2012. Characterizing Pedophile Conversations on the Internet using Online Grooming. *CoRR* abs/1208.4324 (2012). arXiv:1208.4324 http://arxiv.org/abs/1208.4324

[16] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Comput.* 9, 8 (1997), 1735–1780. https://doi.org/10.1162/NECO.1997.9.8.1735

[17] Giacomo Inches and Fabio Crestani. 2012. Overview of the International Sexual Predator Identification Competition at PAN-2012. In *CLEF 2012 Evaluation Labs and Workshop, Online Working Notes, Rome, Italy, September 17-20, 2012 (CEUR Workshop Proceedings, Vol. 1178)*, Pamela Forner, Jussi Karlgren, and Christa Womser-Hacker (Eds.). CEUR-WS.org. https://ceur-ws.org/Vol-1178/CLEF2012wn-PAN-InchesEt2012.pdf

[18] Vitaliy Kashuba, Natalia Nosova, and Yuri Kozlov. 2019. Theoretical and methodological foundations of the physical rehabilitation technology of children 5-6 years old, with functional disorders of the support-motional apparatus.

[19] Jinhwa Kim, Yoon Jo Kim, Mitra Behzadi, and Ian G. Harris. 2020. Analysis of Online Conversations to Detect Cyberpredators Using Recurrent Neural Networks. In *Proceedings for the First International Workshop on Social Threats in Online Conversations: Understanding and Management, STOC@LREC 2020, Marseille, France, May 2020*, Archna Bhatia and Samira Shaikh (Eds.). European Language Resources Association, 15–20. https://aclanthology.org/2020.stoc-1.3/

[20] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *CoRR* abs/1907.11692 (2019). arXiv:1907.11692 http://arxiv.org/abs/1907.11692

[21] Omegle.com LLC. [n. d.]. *Omegle Website.* https://www.omegle.com/

[22] Catherine D Marcum. 2007. Interpreting the intentions of Internet predators: An examination of online predatory behavior. *Journal of Child Sexual Abuse* 16, 4 (2007), 99–114.

[23] India McGhee, Jennifer Bayzick, April Kontostathis, Lynne Edwards, Alexandra McBride, and Emma Jakubowski. 2011. Learning to Identify Internet Sexual Predation. *Int. J. Electron. Commer.* 15, 3 (2011), 103–122. https://doi.org/10.2753/JEC1086-4415150305

[24] Maxime Meyer. 2015. *Machine learning to detect online grooming.* Master's thesis. Uppsala University, Department of Information Technology.

[25] Md. Waliur Rahman Miah, John Yearwood, and Siddhivinayak Kulkarni. 2011. Detection of child exploiting chats from a mixed chat dataset as a text classification task. In *Proceedings of the Australasian Language Technology Association Workshop 2011, Canberra, Australia, December 1-2, 2011*, Diego Mollá and David Martínez (Eds.). ACL, 157–165. https://aclanthology.org/U11-1020/

[26] Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. In *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*, Yoshua Bengio and Yann LeCun (Eds.). http://arxiv.org/abs/1301.3781

[27] Colin Morris. 2013. Identifying online sexual predators by svm classification with lexical and behavioral features. *Master of Science Thesis, University Of Toronto, Canada* (2013).

[28] Fabián Muñoz, Gustavo A. Isaza, and Luis Castillo. 2020. SMARTSEC4COP: Smart Cyber-Grooming Detection Using Natural Language Processing and Convolutional Neural Networks. In *Distributed Computing and Artificial Intelligence, 17th International Conference, DCAI 2020, L'Aquila, Italy, 17-19 June 2020 (Advances in Intelligent Systems and Computing, Vol. 1237)*, Yucheng Dong, Enrique Herrera-Viedma, Kenji Matsui, Shigeru Omatsu, Alfonso González-Briones, and Sara Rodríguez-González (Eds.). Springer, 11–20. https://doi.org/10.1007/978-3-030-53036-5_2

[29] Cynthia H. Ngejane, Jan H. P. Eloff, Tshephisho J. Sefara, and Vukosi Ntsakisi Marivate. 2021. Digital forensics supported by machine learning for the detection of online sexual predatory chats. *Digit. Investig.* 36, Supplement (2021), 301109. https://doi.org/10.1016/J.FSIDI.2021.301109

[30] The Global Partnership and Fund to End Violence Against Children. 2021. *GLOBAL THREAT ASSESSMENT 2021 SHOWS DRAMATIC INCREASE IN ONLINE CHILD SEXUAL EXPLOITATION & ABUSE.* https://www.end-violence.org/articles/global-threat-assessment-2021-shows-dramatic-increase-online-child-sexual-exploitation

[31] Nick Pendar. 2007. Toward Spotting the Pedophile Telling victim from predator in text chats. In *Proceedings of the First IEEE International Conference on Semantic Computing (ICSC 2007), September 17-19, 2007, Irvine, California, USA*. IEEE Computer Society, 235–241. https://doi.org/10.1109/ICSC.2007.32

[32] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, Alessandro Moschitti, Bo Pang, and Walter Daelemans (Eds.). ACL, 1532–1543. https://doi.org/10.3115/v1/d14-1162

[33] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*

abs/1910.01108 (2019). arXiv:1910.01108 http://arxiv.org/abs/1910.01108

[34] scikit-learn Developers. [n. d.]. *Support Vector Machines.* https://scikit-learn.org/stable/modules/svm.html#svm-classification

[35] Michael C Seto. 2009. Pedophilia. *Annual Review of Clinical Psychology* 5 (2009), 391–407.

[36] StatCan. 2022. *Online child sexual exploitation and abuse in Canada, 2014 to 2020.* https://www150.statcan.gc.ca/n1/daily-quotidien/220512/dq220512a-eng.htm

[37] Tarja Susi, Niklas Torstensson, and Ulf Wilhelmsson. 2019. ”Can you send me a photo?” - A Game-Based Approach for Increasing Young Children's Risk Awareness to Prevent Online Sexual Grooming. In *Proceedings of the 2019 DiGRA International Conference: Game, Play and the Emerging Ludo-Mix, DiGRA 2019, Kyoto, Japan, August 6-10, 2019.* Digital Games Research Association. http://www.digra.org/digital-library/publications/can-you-send-me-a-photo-a-game-based-approach-for-increasing-young-childrens-risk-awareness-to-prevent-online-sexual-grooming/

[38] S. Caroline Taylor and Leigh Gassner. 2010. Stemming the flow: challenges for policing adult sexual assault with regard to attrition rates and under-reporting of sexual offences. *Police Practice and Research* 11, 3 (2010), 240–255. https://doi.org/10.1080/15614260902830153 arXiv:https://doi.org/10.1080/15614260902830153

[39] Anna Vartapetiance and Lee Gillam. 2014. ”Our Little Secret”: pinpointing potential predators. *Secur. Informatics* 3, 1 (2014), 3. https://doi.org/10.1186/s13388-014-0003-7

[40] Esaú Villatoro-Tello, Antonio Juárez-González, Hugo Jair Escalante, Manuel Montes-y-Gómez, and Luis Villaseñor Pineda. 2012. A Two-step Approach for Effective Detection of Misbehaving Users in Chats. In *CLEF 2012 Evaluation Labs and Workshop, Online Working Notes, Rome, Italy, September 17-20, 2012 (CEUR Workshop Proceedings, Vol. 1178)*, Pamela Forner, Jussi Karlgren, and Christa Womser-Hacker (Eds.). CEUR-WS.org. https://ceur-ws.org/Vol-1178/CLEF2012wn-PAN-VillatoroTelloEt2012b.pdf

[41] Matthias Vogt, Ulf Leser, and Alan Akbik. 2021. Early Detection of Sexual Predators in Chats. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, 4985–4999. https://doi.org/10.18653/v1/2021.acl-long.386