Enhancing Online Grooming Detection via Backtranslation Augmentation*

Hamed Waezi and Hossein Fani✉
School of Computer Science, University of Windsor, Windsor, ON, Canada

Abstract

Grooming minors for sexual exploitation become an increasingly significant concern in online conversation platforms. For a safer online experience for minors, machine learning models have been proposed to tap into explicit *textual* remarks and automate detecting predatory conversations. Such models, however, fall short of real-world applications for the sparse distribution of predatory conversations. In this paper, we propose backtranslation augmentation to augment training datasets with more predatory conversations. Through our experiments on 8 languages from 4 language families using 3 neural translators, we demonstrate that backtranslation augmentation improves models' performance with less number of training epochs for better classification efficacy. Our code and experimental results are available at github.com/fani-lab/Osprey/tree/coling25.

1 Introduction

An alarming problem in online conversation platforms is the presence of minors before legal age with little cognitive development and the prevalence of online grooming, where an adult sexual predator initiates a sexual relationship with a minor (victim) (Susi et al., 2019; Georgia M. Winters and Jeglic, 2017). Further, online grooming is 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 (Taylor and Gassner, 2010).

For a safer online experience, researchers have proposed neural models, including feedforward (Villatoro-Tello et al., 2012; Escalante et al., 2013; Cheong et al., 2015), convolutional (Ebrahimi et al., 2016) and recurrent networks (Kim et al., 2020; Ngejane et al., 2021b), and transformers (Vogt et al., 2021; Agarwal et al., 2021), to learn from explicit *textual* remarks of predators for online grooming detection and help warn minors, parents or police of such incidents while preserving minors' privacy. Such models, however, suffer from low `recall` due to the sparse distribution of predatory conversations; e.g., in `pan` (Inches and Crestani, 2012) benchmark dataset, merely `2.3%` of conversations are predatory.

In this paper, we proposed to bridge the gap by natural language backtranslation augmentation to enrich training datasets with more predatory conversations. Specifically, we translate original predatory conversations from their original language, e.g., `english`, to a target language, e.g., `french`, and then translate them back to the original language using an off-the-shelf neural translator, e.g., meta's `nllb` (Team et al., 2022), to generate new synthetic predatory conversations. While languages share underlying commonalities, they carry differences on the surface (Friederici, 2017), especially in an informal context like in online conversations, that can be leveraged via backtranslation to generate diverse paraphrases of a predatory conversation while withholding its predatory intent.

From Table 1, backtranslation can uncover latent terms in a predatory conversation as they may *not* be commonly known in a target language and, hence, should be explicitly generated through translation, like when *'having it with minor'* is translated to `french` as *'l'avoir avec mineur'* followed by a backtranslation to `english`, it brings up *'having sex'*. Moreover, backtranslation can augment *context-aware* synonymous terms from a target language to the original predatory conversation, as opposed to simple synonym replacement by a thesaurus (Shiri, 2004). For instance, when *'hooked up'* is translated to `chinese` as '交过', followed by a backtranslation to `english` as *'to have sex'*, it augments *'sex'* as opposed to other semantics like *'to plug in'* in electrical nomenclature. Finally, back-

*__Warning:__ This paper discusses online grooming that may be offensive or upsetting.

Table 1: Backtranslation examples of predatory messages from pan (Inches and Crestani, 2012).

original message	(language) translation	backtranslation
'having it with minor'	(french) *'l'avoir avec mineur'*	*'having sex with a minor'*
'i feel little aroused'	(deutsch) *'ich fühle mich ein wenig erregt'*	*'i'm feeling a little turned on'*
'you ever hooked up with anybody...?'	(chinese) '你有没有和网上的人交过?'	*'have you ever had sex with ...?'*
'like two guys doing each other?'	(deutsch) *'wie zwei typen, die es miteinander treiben?'*	*'like two guys having sex?'*

translation can disambiguate polysemous collocations, like translating an ambiguous message *'doing each other'* to deutsch *'miteinander treiben'*, and backtranslating to english, maps the term *'each other'* to *'sex'*;

For similar reasons, backtranslation has been employed in review analysis and opinion mining (Fei et al., 2021; Liesting et al., 2021; Hemmatizadeh et al., 2023) and other natural language processing tasks like text summarization (Fabbri et al., 2021) and question-answering (Bhaisaheb et al., 2023), and machine translation (Guo et al., 2021; Sennrich et al., 2016). Furthermore, the open-source accessibility to neural translators (Team et al., 2022), capable of delivering high-quality translations between many languages, as well as their seamless integration into any pipeline with few lines of code, have already set off a surge of interest. Nonetheless, other augmentation techniques such as rule-based (Wei and Zou, 2019), synonym replacement (Kolomiyets et al., 2011), and structure-based (Min et al., 2020) fall short in online grooming detection due to the short, noisy and informal messages.

2 Problem Definition

A conversation c in a language l is a sequence of $|c|$ timestamped messages $m_i^c; 1 < i < |c|$, each message of which includes id, text, author, and timestamp. Furthermore, as opposed to an online post or comment, an online conversation should have at least two different authors, each of whom has at least one message, i.e., $\exists m_i^c, m_j^c, i \neq j$ such that m_i^c.author $\neq m_j^c$.author. Let $\mathcal{C} = \{c\}$ be the set of conversations, our task is to learn $f_\theta : \mathcal{C} \rightarrow \{0 : normal, 1 : predatory\}$, 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.

3 Backtranslation Augmentation

We learn the mapping function f_θ from a set of conversations $\mathcal{C}^+ = \mathcal{C} \cup \mathcal{C}^l$ that is augmented by backtranslated versions of predatory conversations via a language l. Let $\mathcal{L}$ be the set of natural languages, τ be a two-way translator, and c be a predatory conversation. We *forward* translate each message of the predatory conversation m_i^c to a target language l and translate it *back* to the source language using the translator τ, resulting in the backtranslated version of each message, denoted by $m_i^{c \leftarrow l}$. We collect the backtranslated messages and form a new predatory conversation $c^{\leftarrow l}$ as the backtranslated version of c, withholding the same values in other attributes like timestamp and author. Finally, we augment the dataset with backtranslated versions of existing conversations.

4 Experiments

4.1 Dataset

Access to training sets of online conversations remains challenging due to privacy and legal concerns. Previous datasets such as conversations from an online game for minors (Cheong et al., 2015), chat-coder (McGhee et al., 2011), and pan-chat-coder (Vogt et al., 2021), are inaccessible to researchers. The sole accessible benchmark dataset in the literature is pan (Inches and Crestani, 2012) (Appendix B), which extensively used in prior studies (McGhee et al., 2011; Bogdanova et al., 2012; Ebrahimi et al., 2016; Cardei and Rebedea, 2017; Aragón and López-Monroy, 2018). In our experiments, we removed conversations with only 1 participant or those with fewer than 6 message exchanges.

4.2 Backtranslation

We chose french, deutsch, icelandic, and catalan from indo-european, farsi and pashto from iranic, and chinese and myanmarese from sino-tibetan, among which icelandic, catalan, pashto, and myanmarese are low-resource languages. For translation and backtranslation, we utilized three two-way neural translators: meta's nllb (Team et al., 2022) and m2m100 (Fan et al., 2021), and google's translator. These translators can perform translations to and from over 100 languages with a single model, enabling us to con-

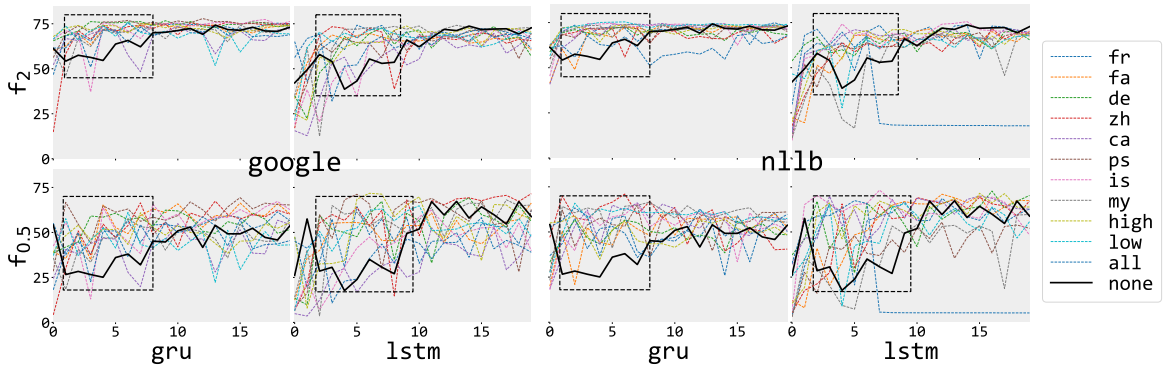

Figure 1: Training efficiency vs. inference efficacy. Baselines converge faster in the first 10 epochs on the augmented dataset (colored lines) for better f-measures on the test set compared to the lack thereof (black line). As seen in Appendix D, m2m100 follows the same trend.

duct a comprehensive study on a wide variety of languages. All three translators are based on transformers. However, while meta's translators are open-sourced, google's translator is closed, yet it is a well-known commercial translator. In terms of translation quality, nllb is the state-of-the-art on benchmark translation datasets (Team et al., 2022).

4.3 Baselines

We trained state-of-the-art recurrent model by Waezi et al. (2024)(gru), and the strong competitor by Kim et al. (2020)(lstm), to estimate f_θ for online grooming detection on the backtranslated augmented dataset and lack thereof. Both models have a single layer with 512 units, utilizing the tanh activation function and the Adam optimizer. Each conversation was vectorized as a sequence of its message embeddings using pretrained 768-dimensional vectors of distilroberta (Sanh et al., 2019).

4.4 Evaluation Methodology

We performed 3-fold cross-validation. For each fold, we conducted two separate training sessions for a baseline model: one using the original fold and one using the augmented one. We evaluated the performance of the trained models on the same test set using f-measures with $\beta = 2.0$ to favour recall over precision vs. $\beta = 0.5$ vice versa, and $\beta = 1.0$ for equal importance. Finally, we compared the average results over the folds. To study how backtranslation augmentation improves models' efficiency during training, we reported the models' performance on the test set at each training epoch.

4.5 Results

From Figure 1, we observe that baselines converge faster during less number of training epochs when the training set is augmented with backtranslations across different languages compared to the lack thereof in terms of f-measures. In terms of efficacy, Table 2 shows the performance delta before and after backtranslation augmentation of the training set for baselines after 20 epochs. As seen, backtranslation augmentation helps with the models' efficacy overall. However, the performance gain depends on the language, translator, and baseline model.

Regarding the effects of each language, language families, and their combinations on the baselines' efficacy, we observe that low-resource languages individually have shown an overall better performance like catalan (ca), pashto (pa), and myanmarese (my), which can be attributed to their better paraphrasing (Appendix C). Low-resource languages have shown relatively higher semantic similarities for the relatively low bleu scores. In contrast, chinese, a high-resource language, yielded a lower performance in Table 2 due to its poor backtranslations with the lowest bleu *and* semantic similarities. From the results of integrating the backtranslations from languages of the same family, we observe that *not* all language families show synergy. While integrating backtranslations of french (fr) + catalan (ca) from the western romance family boost the baselines' performance, chinese (zh) + myanmarese (my) from the sino-tibetan have a discounting effect. However, when we integrate more languages based on their high or low-resource richness, or integrating all languages, backtranslation augmentation shows positive impacts in general.

Table 2: Average 3-fold cross-validation results of baselines for 20 training epochs using backtranslation augmentations and lack thereof (none) on the same test set based on the performance delta (Δ). Best viewed in color. Actual values of the metrics are in Appendix D, also in our codebase.

			Δf0.5			Δf1			Δf2		
			google	m2m100	nllb	google	m2m100	nllb	google	m2m100	nllb
gru	none		0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	+fr		+8.76	+2.61	+2.65	+7.08	+2.84	+2.63	+5.80	+2.46	+3.35
	+fa		+1.68	+3.50	+4.05	+1.88	+3.64	+3.39	+2.43	+5.42	+3.04
	+de		+4.10	+3.81	+9.00	+3.76	+2.97	+6.39	+4.22	+2.92	+2.27
	+zh		-8.97	+1.55	+5.63	-5.87	+2.14	+4.27	-3.92	+4.89	+2.61
	+ca		+7.29	**+7.44**	+11.03	+6.73	+5.40	+8.04	+5.01	+2.67	+4.70
	+ps		9.57	-5.32	+13.60	+6.48	-3.79	+7.58	+3.04	+1.56	+0.52
	+is		+3.73	-2.24	+10.14	+4.04	-0.68	+6.98	+6.27	+2.25	+3.24
	+my		**+12.51**	-3.13	+9.34	+9.13	-1.66	+6.06	+4.63	+1.46	+2.00
	+fr+ca	western romance	+8.07	-4.40	**+15.29**	+7.02	-2.35	**+9.94**	+6.76	-0.81	+3.52
	+fa+ps	iranic	-3.31	-6.71	+2.91	-1.88	-4.75	+2.62	+0.34	+0.78	+3.60
	+de+is	west germanic	+11.84	+5.94	9.43	**+9.39**	**+5.60**	+7.65	**+6.98**	**+6.63**	**+6.16**
	+zh+my	sino-tibetan	+8.83	-8.76	-3.15	+6.80	-6.33	-1.21	+4.55	-7.05	+0.66
	+fr+fa+de+zh	high-resource	+5.88	-1.36	10.14	+4.62	-0.40	+7.63	+3.94	+3.01	+4.97
	+ca+ps+is+my	low-resource	+8.10	-1.38	+10.41	+6.61	-0.52	+5.26	+5.63	+2.64	-1.00
	all		+4.25	-0.32	+6.51	+4.30	+0.06	+5.20	+6.02	+2.16	+4.64
lstm	none		0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	+fr		+8.42	-2.78	+5.30	+1.79	-2.10	+2.55	-6.71	-2.64	-1.43
	+fa		+1.99	+4.06	-6.82	-0.03	+0.65	-5.59	-3.53	-4.19	-7.95
	+de		+2.14	+4.16	-0.58	+1.78	-0.17	-0.70	+1.21	-5.75	-3.90
	+zh		+5.80	+2.04	-6.30	-0.02	-0.40	-2.66	-7.02	-4.87	-13.18
	+ca		-2.09	+2.83	+8.03	-0.74	-0.08	+1.97	+0.69	-5.70	-5.46
	+ps		-9.05	+4.32	+1.68	-7.05	+2.06	-1.89	-6.36	-0.82	-9.89
	+is		+1.83	-10.86	+6.64	+1.66	-8.07	+1.10	-2.03	-3.81	-6.26
	+my		-3.94	+0.28	+8.86	-3.15	-2.63	+4.52	-7.74	-6.91	-0.87
	+fr+ca	western romance	0.08	-12.39	+0.96	-0.67	-8.89	-1.16	-2.02	-4.14	-4.55
	+fa+ps	iranic	+8.68	-11.79	+2.90	+2.38	-8.87	-0.33	-5.59	-4.43	-6.13
	+de+is	western germanic	-13.23	-0.62	+1.70	-9.32	-0.97	-0.18	-8.14	-2.33	-3.62
	+zh+my	sino-tibetan	-1.32	+4.27	+0.57	-0.04	+0.52	-2.88	+0.91	-4.26	-9.11
	+fr+fa+de+zh	high-resource	+1.44	-8.80	+9.38	-1.60	-6.98	+2.91	-5.55	-8.82	-5.00
	+ca+ps+is+my	low-resource	-11.90	+0.56	+7.01	-8.57	-1.17	+1.67	-4.89	-3.87	-5.78
	all		+3.24	+3.28	+4.54	+0.69	-0.17	+1.51	-2.85	-4.60	-3.39

For the quality of neural translators on the performance gain, from Table 2, we see that the translation by nllb and google have resulted in the best and runner-up performance improvements, respectively, while m2m100 has shown less effectiveness. Specifically, in low-resource languages, m2m100's backtranslations have shown subpar performance compared to nllb and google. Our results are also aligned with translation benchmarks, and the fact that nllb has been developed with low-resource languages in mind (Team et al., 2022).

To see whether backtranslation augmentation consistently benefits the performance of the baseline models, we clearly observe that gru's performance improvement has been positive overall across different languages and metrics. Surprisingly, lstm's performance is *not* following a similar trend; while lstm's f0.5 has been improved across high-resource languages, its performance drops in other languages for f1 and f2. Our results are in line with Waezi et al. (2024)'s work where gru outperformed lstm due to its better gating strategy to retain dependencies from earlier messages in long conversations as in predatory conversations.

5 Concluding Remarks

In this paper, we proposed backtranslation augmentation of predatory conversations for online grooming detection. We showed that (1) backtranslation augmentation improves models' performance with less number of training epochs for better classification efficacy; (2) low-resource languages have shown better performance; (3) higher quality neural translators yield more performance gain; and (4) finally, the underlying model architecture matters where gru consistently improves upon backtranslation augmentation across all languages while lstm improves only across high-resource languages.

6 Limitations

The main limitation of this study lies in the benchmark dataset, pan, which is solely in English. This restricts the generalizability of our findings to other languages. We acknowledge that online grooming occurs across other languages, highlighting the critical need for non-English training datasets. Expanding research to include multilingual datasets would allow for a more comprehensive evaluation of online grooming detection techniques, including the effectiveness of our backtranslation augmentation. Additionally, the victims in pan are trained adult *decoys* rather than actual minors, which may affect the quality and reliability of results. Finally, while the ultimate objective of online grooming detection is to identify predators before they can harm potential victims, our study requires the *entire* conversation for classification. In future work, we plan to focus on the task of *early* detection, that is, identifying grooming behaviors at initial or early stages based on the first few messages before the conversation escalates into more serious exploitation or abuse.

7 Ethical Considerations

The researchers involved in this study were all adults who were warned and fully informed about the harmful content of predatory conversations in the benchmark dataset. Additionally, they underwent appropriate training to ensure they were prepared and would not experience any mental distress during the course of the research.

References

Nancy Agarwal, Tugçe Ünlü, Mudasir Ahmad Wani, and Patrick Bours. 2021. Predatory conversation detection using transfer learning approach. In *Machine Learning, Optimization, and Data Science - 7th International Conference, LOD 2021, Grasmere, UK, October 4-8, 2021, Revised Selected Papers, Part I*, volume 13163 of *Lecture Notes in Computer Science*, pages 488–499. Springer.

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*, volume 2125 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Segun Taofeek Aroyehun and Alexander F. Gelbukh. 2018. Aggression detection in social media: Using deep neural networks, data augmentation, and pseudo labeling. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying, TRAC@COLING 2018, Santa Fe, New Mexico, USA, August 25, 2018*, pages 90–97. Association for Computational Linguistics.

Markus Bayer, Marc-André Kaufhold, and Christian Reuter. 2023. A survey on data augmentation for text classification. *ACM Comput. Surv.*, 55(7):146:1–146:39.

Yonatan Belinkov and Yonatan Bisk. 2018. Synthetic and natural noise both break neural machine translation. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Shabbirhussain Bhaisaheb, Shubham Paliwal, Rajaswa Patil, Manasi Patwardhan, Lovekesh Vig, and Gautam Shroff. 2023. Program synthesis for complex QA on charts via probabilistic grammar based filtered iterative back-translation. In *Findings of the Association for Computational Linguistics: EACL 2023, Dubrovnik, Croatia, May 2-6, 2023*, pages 2456–2470. Association for Computational Linguistics.

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*, pages 110–118. The Association for Computer Linguistics.

Tolga Bolukbasi, Adam Pearce, Ann Yuan, Andy Coenen, Emily Reif, Fernanda B. Viégas, and Martin Wattenberg. 2021. An interpretability illusion for BERT. *CoRR*, abs/2104.07143.

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*, pages 1–6. IEEE.

Laura Jayne Broome, Cristina Izura, and Jason Davies. 2020. A psycho-linguistic profile of online grooming conversations: A comparative study of prison and police staff considerations. *Child Abuse Neglect*, 109:104647.

Rui Cao and Roy Ka-Wei Lee. 2020. Hategan: Adversarial generative-based data augmentation for hate speech detection. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 6327–6338. International Committee on Computational Linguistics.

Claudia Cardei and Traian Rebedea. 2017. Detecting sexual predators in chats using behavioral features and imbalanced learning. *Nat. Lang. Eng.*, 23(4):589–616.

Camilla Casula and Sara Tonelli. 2023. Generation-based data augmentation for offensive language detection: Is it worth it? In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2023, Dubrovnik, Croatia, May 2-6, 2023*, pages 3351–3369. Association for Computational Linguistics.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Brian Strope, and Ray Kurzweil. 2018. Universal sentence encoder for english. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018: System Demonstrations, Brussels, Belgium, October 31 - November 4, 2018*, pages 169–174. Association for Computational Linguistics.

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)*.

Jiaao Chen, Zichao Yang, and Diyi Yang. 2020. Mixtext: Linguistically-informed interpolation of hidden space for semi-supervised text classification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 2147–2157. Association for Computational Linguistics.

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):220–232.

Jonie Chiu and Ethel Quayle. 2022. Understanding online grooming: An interpretative phenomenological analysis of adolescents' offline meetings with adult perpetrators. *Child Abuse Neglect*, 128:105600.

Claude Coulombe. 2018. Text data augmentation made simple by leveraging NLP cloud apis. *CoRR*, abs/1812.04718.

Mohammad Reza Ebrahimi, Ching Y. Suen, and Olga Ormandjieva. 2016. Detecting predatory conversations in social media by deep convolutional neural networks. *Digit. Investig.*, 18:33–49.

Sergey Edunov, Myle Ott, Marc'Aurelio Ranzato, and Michael Auli. 2019. On the evaluation of machine translation systems trained with back-translation. *CoRR*, abs/1908.05204.

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*, pages 46–54. The Association for Computer Linguistics.

Alexander R. Fabbri, Simeng Han, Haoyuan Li, Haoran Li, Marjan Ghazvininejad, Shafiq R. Joty, Dragomir R. Radev, and Yashar Mehdad. 2021. Improving zero and few-shot abstractive summarization with intermediate fine-tuning and data augmentation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 704–717. Association for Computational Linguistics.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Michael Auli, and Armand Joulin. 2021. Beyond english-centric multilingual machine translation. *J. Mach. Learn. Res.*, 22:107:1–107:48.

Hao Fei, Yafeng Ren, Shengqiong Wu, Bobo Li, and Donghong Ji. 2021. Latent target-opinion as prior for document-level sentiment classification: A variational approach from fine-grained perspective. In *WWW '21: The Web Conference 2021, Virtual Event / Ljubljana, Slovenia, April 19-23, 2021*, pages 553–564. ACM / IW3C2.

Angela D Friederici. 2017. *Language in our brain: The origins of a uniquely human capacity*. MIT Press.

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):62–76.

John M. Giorgi, Osvald Nitski, Bo Wang, and Gary D. Bader. 2021. Declutr: Deep contrastive learning for unsupervised textual representations. 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*, pages 879–895. Association for Computational Linguistics.

Yinuo Guo, Hualei Zhu, Zeqi Lin, Bei Chen, Jian-Guang Lou, and Dongmei Zhang. 2021. Revisiting iterative back-translation from the perspective of compositional generalization. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 7601–7609. AAAI Press.

Farinam Hemmatizadeh, Christine Wong, Alice Yu, and Hossein Fani. 2023. Latent aspect detection via backtranslation augmentation. In *CIKM*, pages 3943–3947. ACM.

Mai Ibrahim, Marwan Torki, and Nagwa M. El-Makky. 2020. Alexu-backtranslation-tl at semeval-2020 task 12: Improving offensive language detection using

data augmentation and transfer learning. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation, SemEval@COLING 2020, Barcelona (online), December 12-13, 2020*, pages 1881–1890. International Committee for Computational Linguistics.

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*, volume 1178 of *CEUR Workshop Proceedings*. CEUR-WS.org.

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*, pages 15–20. European Language Resources Association.

Oleksandr Kolomiyets, Steven Bethard, and Marie-Francine Moens. 2011. Model-portability experiments for textual temporal analysis. In *The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference, 19-24 June, 2011, Portland, Oregon, USA - Short Papers*, pages 271–276. The Association for Computer Linguistics.

Anna Kruspe, Jens Kersten, Matti Wiegmann, Benno Stein, and Friederike Klan. 2018. Classification of incident-related tweets: Tackling imbalanced training data using hybrid cnns and translation-based data augmentation. *Notebook papers of TREC*.

Varun Kumar, Hadrien Glaude, Cyprien de Lichy, and William Campbell. 2019. A closer look at feature space data augmentation for few-shot intent classification. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP, DeepLo@EMNLP-IJCNLP 2019, Hong Kong, China, November 3, 2019*, pages 1–10. Association for Computational Linguistics.

Tomas Liesting, Flavius Frasincar, and Maria Mihaela Trusca. 2021. Data augmentation in a hybrid approach for aspect-based sentiment analysis. In *SAC '21: The 36th ACM/SIGAPP Symposium on Applied Computing, Virtual Event, Republic of Korea, March 22-26, 2021*, pages 828–835. ACM.

Sreekanth Madisetty and Maunendra Sankar Desarkar. 2018. Aggression detection in social media using deep neural networks. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying, TRAC@COLING 2018, Santa Fe, New Mexico, USA, August 25, 2018*, pages 120–127. Association for Computational Linguistics.

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):103–122.

Paul McNamee and Kevin Duh. 2023. An extensive exploration of back-translation in 60 languages. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 8166–8183. Association for Computational Linguistics.

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*.

Junghyun Min, R. Thomas McCoy, Dipanjan Das, Emily Pitler, and Tal Linzen. 2020. Syntactic data augmentation increases robustness to inference heuristics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 2339–2352. Association for Computational Linguistics.

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*, volume 1237 of *Advances in Intelligent Systems and Computing*, pages 11–20. Springer.

C.H. Ngejane, J.H.P. Eloff, T.J. Sefara, and V.N. Marivate. 2021a. Digital forensics supported by machine learning for the detection of online sexual predatory chats. *Forensic Science International: Digital Investigation*, 36:301109.

Cynthia H. Ngejane, Jan H. P. Eloff, Tshephisho J. Sefara, and Vukosi Ntsakisi Marivate. 2021b. Digital forensics supported by machine learning for the detection of online sexual predatory chats. *Digit. Investig.*, 36(Supplement):301109.

Sumant Patil and Patrick Davies. 2014. Use of google translate in medical communication: evaluation of accuracy. *BMJ*, 349.

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*, pages 1532–1543. ACL.

Siyuan Qiu, Binxia Xu, Jie Zhang, Yafang Wang, Xiaoyu Shen, Gerard de Melo, Chong Long, and Xiaolong Li. 2020. Easyaug: An automatic textual data augmentation platform for classification tasks. In *Companion of The 2020 Web Conference 2020, Taipei, Taiwan, April 20-24, 2020*, pages 249–252. ACM / IW3C2.

Yanru Qu, Dinghan Shen, Yelong Shen, Sandra Sajeev, Weizhu Chen, and Jiawei Han. 2021. Coda: Contrast-enhanced and diversity-promoting data augmentation

for natural language understanding. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Tatiana R. Ringenberg, Kathryn C. Seigfried-Spellar, Julia M. Rayz, and Marcus K. Rogers. 2022. A scoping review of child grooming strategies: pre- and post-internet. *Child Abuse Neglect*, 123:105392.

Julian Risch and Ralf Krestel. 2018. Aggression identification using deep learning and data augmentation. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying, TRAC@COLING 2018, Santa Fe, New Mexico, USA, August 25, 2018*, pages 150–158. Association for Computational Linguistics.

Georgios Rizos, Konstantin Hemker, and Björn W. Schuller. 2019. Augment to prevent: Short-text data augmentation in deep learning for hate-speech classification. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management, CIKM 2019, Beijing, China, November 3-7, 2019*, pages 991–1000. ACM.

Mehdi Sajjadi, Mehran Javanmardi, and Tolga Tasdizen. 2016. Regularization with stochastic transformations and perturbations for deep semi-supervised learning. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 1163–1171.

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.

Konstanze Schoeps, Montserrat Peris Hernández, María Teresa Garaigordobil Landazabal, Inmaculada Montoya Castilla, et al. 2020. Risk factors for being a victim of online grooming in adolescents. *Psicothema*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics.

Ali Asghar Shiri. 2004. End-user interaction with thesaurus-enhanced search interfaces: an evaluation of search term selection for query expansion. *SIGIR Forum*, 38(1):80.

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.

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):240–255.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Y. Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loïc Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. *CoRR*, abs/2207.04672.

Ingrid Ravn Turkerud and Ole Jakob Mengshoel. 2021. Image captioning using deep learning: Text augmentation by paraphrasing via backtranslation. In *IEEE Symposium Series on Computational Intelligence, SSCI 2021, Orlando, FL, USA, December 5-7, 2021*, pages 1–10. IEEE.

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*, volume 1178 of *CEUR Workshop Proceedings*. CEUR-WS.org.

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*, pages 4985–4999. Association for Computational Linguistics.

Hamed Waezi, Reza Barzegar, and Hossein Fani. 2024. Osprey: A reference framework for online grooming detection via neural models and conversation features. In *CIKM*. ACM.

Jason W. Wei and Kai Zou. 2019. EDA: easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 6381–6387. Association for Computational Linguistics.

Qizhe Xie, Zihang Dai, Eduard H. Hovy, Thang Luong, and Quoc Le. 2020. Unsupervised data augmentation for consistency training. In *Advances in Neural*

Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual.

Adams Wei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc V. Le. 2018. Qanet: Combining local convolution with global self-attention for reading comprehension. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

A Related Works

The related works to this paper are largely centred around two areas: (1) online grooming detection and (2) data augmentation. We acknowledge research directions for online grooming from a non-computational perspective in psychology (Chiu and Quayle, 2022; Schoeps et al., 2020), behavioral studies (Ringenberg et al., 2022; Broome et al., 2020), and forensics (Ngejane et al., 2021a). While such studies help with developing computational models, we exclude them for being beyond the scope of this paper.

A.1 Online Grooming Detection

The primary means of online grooming is *textual* messages. Hence, natural language processing techniques have been widely used to detect online grooming through machine learning classifiers on vector representations of conversations, which can be categorized into *i*) sparse vector representation, *ii*) low-dimensional dense vector representations, and *iii*) hand-crafted feature representations from conversations. Initially, sparse vector representations for conversations have been widely used, like one-hot vectors for each word or message or bag-of-words representations (Villatoro-Tello et al., 2012; Escalante et al., 2013; Cheong et al., 2015; Ebrahimi et al., 2016). For instance, Villatoro-Tello et al. (Villatoro-Tello et al., 2012) used bag-of-words of raw messages without text preprocessing to capture the characteristics of online chats, including misspellings and emoticons. Ebrahimi et al. (Ebrahimi et al., 2016) employed concatenation of one-hot vectors for each message to preserve token order, resulting in marginal improvements over bag-of-words but at the cost of increased dimensionality in sparse vectors. Despite their simplicity, sparse representations are already known to suffer from out-of-vocabulary, loss of token order, and high dimensionality, to name a few. Next, pretrained word embeddings from `word2vec` (Mikolov et al., 2013) and `glove` (Pennington et al., 2014) have been employed, following their success in various nlp tasks like document classification (Ebrahimi et al., 2016; Muñoz et al., 2020). Such embeddings, however, performed poorly for being trained on corpora different from informal online chats. State-of-the-art methods use contextualized word embeddings for online grooming (Waezi et al., 2024; Vogt et al., 2021; Chehbouni et al., 2022). Chehbouni et al. (Chehbouni et al., 2022) used pretrained bert model to encode each message into embeddings for a logistic regression classifier. Kim et al. (Kim et al., 2020) employed universal sentence encoders (Cer et al., 2018) to encode each message as a single dense vector. More recently, Waezi et al. (Waezi et al., 2024) proposed to incorporate conversational features, such as the message's timestamp and the number of participants, to capture characteristic features of online grooming.

In terms of classifiers, earlier works such as Villatoro-Tello et al. (Villatoro-Tello et al., 2012) and others used support vector machines (Bours and Kulsrud, 2019; Villatoro-Tello et al., 2012; Escalante et al., 2013; Cheong et al., 2015) and logistic regression (Chehbouni et al., 2022; Cheong et al., 2015). Other classical machine learning models have also been employed, including k-nearest neighbors (Chehbouni et al., 2022), naive Bayes (Bogdanova et al., 2012), decision trees (McGhee et al., 2011; Cheong et al., 2015), and feedforward neural networks (Villatoro-Tello et al., 2012; Escalante et al., 2013; Cheong et al., 2015). Recent works have increasingly adopted neural models, including convolutional neural networks (Ebrahimi et al., 2016), recurrent neural networks (Waezi et al., 2024; Ngejane et al., 2021a), and transformer-based models (Vogt et al., 2021), which enable considering larger or even the entire context of a conversation for classification. For example, Kim et al. (Kim et al., 2020) used `lstm` to learn from conversations as sequences of words, in contrast to a single document and bag of words. Similarly, Waezi et al. (Waezi et al., 2024) processed conversations as sequences of messages but using `gru`, showing `gru` has a better gating strategy versus `lstm` for predatory conversations, which are often long.

Nonetheless, despite well-established data augmentation techniques in nlp, no work has been proposed to address the highly sparse distribution of predatory conversations in training datasets, to

Table 3: Details of neural translators.

	google	m2m100	nllb
#languages	133	101	196
model card	✗	✓	✓
#parameters	unknown	1.2 billion	3.3 billion
license	closed source	mit	cc-by-nc
owner	google	meta	meta
architecture	transformer +rnn	transformers	transformers

Table 4: Languages used in this paper.

resource	language	family
high	(en) english	west germanic
	(fr) french	western romance
	(fa) farsi	iranic
	(de) deutsch	west germanic
	(zh) chinese	sino-tibetan
low	(ca) catalan	western romance
	(ps) pashto	iranic
	(is) icelandic	west germanic
	(my) myanmarese	sino-tibetan

the best of our knowledge. In this paper, we are the first to bridge the gap and undertake a data augmentation method via backtranslation to enhance online grooming detection.

A.2 Data Augmentation

Augmentation techniques have helped models' robustness and generalization for out-of-vocabulary and out-of-distribution scenarios during inference on unseen text, which can be categorized based on where augmentation happens in the machine learning pipeline (Bayer et al., 2023): (1) data space, which involves augmenting pieces of text directly in levels of character, word, phrase, and sentence, and (2) feature space, where the vector representations (embeddings) of input texts in a latent space are used to augment new data by, e.g., introducing noise to a vector or interpolating new vectors from existing ones (Kumar et al., 2019; Chen et al., 2020). In contrast to feature space augmentation, where the augmented vectors are not interpretable for humans (Bolukbasi et al., 2021) and their generation is often computationally costly, data space augmentations are simpler yet more effective and include noise addition (Belinkov and Bisk, 2018), rule-based transformations (Coulombe, 2018), synonym replacement (Kolomiyets et al., 2011), structure-based manipulation (Min et al., 2020), machine-generated text (Qiu et al., 2020), and backtranslation (Hemmatizadeh et al., 2023; Risch and Krestel, 2018). For example, for the purpose of online text classification, Risch and Krestel (Risch and Krestel, 2018) utilized backtranslation to enhance the detection of online aggression and bullying, while Rizos et al. (Rizos et al., 2019) employed synonym substitution for hate speech detection. Additionally, Cao and Lee (Cao and Lee, 2020) and Casula and Tonelli (Casula and Tonelli, 2023) used machine-generated text to augment datasets of hate speech detection and offensive language detection, respectively.

Among data space augmentation methods, backtranslation has been notably used (Aroyehun and Gelbukh, 2018; Qu et al., 2021; Xie et al., 2020) due to its ability to create new paraphrases of an existing text with new vocabulary and structure while controlling the meaning and semantic context. Moreover, the open-source accessibility to two-way multilingual neural translators with high-quality translations between many languages, including low-resource ones, as well as their easy integration into any pipeline led to the growing interest in backtranslation augmentation. McNamee and Duh (McNamee and Duh, 2023) and others (Hemmatizadeh et al., 2023; Aroyehun and Gelbukh, 2018; Liesting et al., 2021) showed that backtranslation could significantly improve the translation task itself for languages with moderate and low resources. It also enhances the fluency (Edunov et al., 2019; Sennrich et al., 2016), reduces overfitting, and improves robustness (Sajjadi et al., 2016). For instance, Xie et al. (Xie et al., 2020) integrated backtranslation as part of *consistency training*[1], making translator models invariant to noise or minor changes, thus enhancing robustness (Xie et al., 2020). Backtranslation has also been used in squad benchmark (Yu et al., 2018), tweet classification (Kruspe et al., 2018), image captioning (Turkerud and Mengshoel, 2021), aspect-based sentiment analysis (Liesting et al., 2021; Hemmatizadeh et al., 2023), and in domains close but different from online grooming, like aggression (Aroyehun and Gelbukh, 2018) and offensive language (Ibrahim et al., 2020) detection.

However, in online grooming, where turn-taking conversations are involved as opposed to an online post or comment, the effect of data augmentation, in general, and backtranslation augmentation, in

[1]Consistency training involves perturbations to input text for semi-supervised learning where labeled data are scarce and robustness to adversarial attacks is required. (Sajjadi et al., 2016)

Table 5: Statistics of pan (Inches and Crestani, 2012) 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/ single participant	12,773	29,561	**0**	**0**
#**predatory** conversations w/ 2+ participants	**0**	**0**	**0**	**0**
avg #msgs in a **predatory** conversations	60.73	90.07	80.68	71.48
avg #msgs in a normal conversations	12.74	12.86	41.73	41.78
avg #words in a msg of a **predatory** conversations	4.47	4.63	4.38	4.51
avg #words in a msg of a normal conversations	6.39	6.77	6.91	7.16

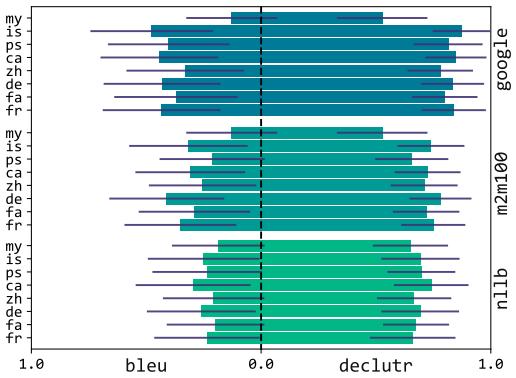

Figure 2: bleu and semantic similarity (declutr) of backtranslated messages against the original ones.

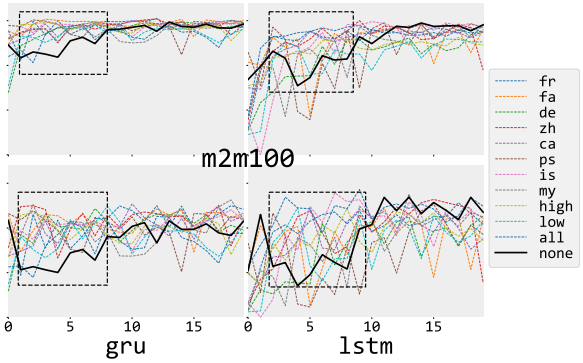

Figure 3: Training efficiency vs. inference efficacy for m2m100. Baselines converge faster in the first 10 epochs on the augmented dataset (colored lines) for better f-measures on the test set compared to the lack thereof (black line), which is a similar trend as in nllb and google.

particular, is yet to be studied. This paper is a pioneering effort to utilize backtranslation as a data augmentation strategy to improve online grooming detection, esp., when training datasets are inherently extremely imbalanced.

B Dataset

The pan dataset includes cases of online grooming about 10 years obtained from trained volunteers (decoys) posing as minors in public conversation platforms to catch and convict predators. The normal conversations in this dataset are sourced from omegle online chatrooms[2] and internet relay chat logs[3]. It also includes conversations with a single participant and a small number of messages. We filter such conversations and those with less than 6 messages. Table 5 shows the statistics of the datasets before and after filtering. As seen, the dataset is extremely imbalanced against predatory conversations, which include only 2 participants and are generally longer.

[2] omegle.inportb.com

[3] irclog.org and krijnhoetmer.nl/irc-logs

C Effective Backtranslation for Augmentation

Table 3 summarizes the neural translators used in this paper, where google is closed-source yet a well-known commercial translator, widely used by the general public, industry, and academia (Patil and Davies, 2014; Yu et al., 2018; Madisetty and Desarkar, 2018), and m2m100 and nllb are open-source from meta. We presume backtranslation is effective for augmentation if it paraphrases the original predatory messages of a conversation into new wordings while withholding the semantic context of grooming. We opt for bleu to measure the n-gram overlap in wordings between the original and backtranslated messages. Meanwhile, we measure the semantic similarity between the original and backtranslated messages by declutr as the state-of-the-art model-based method (Giorgi et al., 2021), which calculates the semantic similarity of a pair of texts based on their cosine similarity in a vector space. From Figure 2, the semantic similarity of most backtranslations (paraphrases) to the original

Table 6: Average 3-fold cross-validation results of baselines for 20 training epochs using backtranslation augmentations and lack thereof (none) on the same test set.

			f0.5			f1			f2		
			google	m2m100	nllb	google	m2m100	nllb	google	m2m100	nllb
gru	none		50.95	50.95	50.95	59.03	59.03	59.03	68.15	68.15	68.15
	+fr		59.72	53.56	53.61	66.11	61.87	61.66	73.96	70.62	71.51
	+fa		52.63	54.45	55.01	60.91	62.68	62.42	70.59	73.58	71.20
	+de		55.05	54.77	59.96	62.79	62.00	65.42	72.38	71.07	70.42
	+zh		41.99	52.50	56.59	53.17	61.17	63.30	64.24	73.05	70.77
	+ca		58.24	58.40	61.99	65.76	64.44	67.07	73.17	70.83	72.85
	+ps		60.53	45.63	64.55	65.51	55.24	66.61	71.20	69.72	68.68
	+is		54.68	48.72	61.09	63.07	58.36	66.01	74.42	70.41	71.40
	+my		63.46	47.82	60.30	68.16	57.38	65.09	72.79	69.62	70.16
	+fr+ca	west romance	59.03	46.55	66.25	66.05	56.68	68.97	74.92	67.35	71.68
	+fa+ps	iranic	47.65	44.24	53.87	57.15	54.28	61.65	68.50	68.93	71.76
	+de+is	western germanic	62.80	56.90	60.38	68.43	64.64	66.68	75.14	74.79	74.32
	+zh+my	sino-tibetan	59.79	42.20	47.81	65.83	52.70	57.82	72.70	61.10	68.82
	+fr+fa+de+zh	high-resource	56.83	49.59	61.09	63.65	58.63	66.66	72.10	71.17	73.13
	+ca+ps+is+my	low-resource	59.05	49.58	61.37	65.64	58.51	64.29	73.79	70.80	67.16
	all		55.20	50.64	57.47	63.33	59.09	64.23	74.18	70.32	72.80
lstm	none		58.39	58.39	58.39	64.51	64.51	64.51	71.76	71.76	71.76
	+fr		66.82	55.62	63.70	66.31	62.42	67.07	65.06	69.12	70.33
	+fa		60.39	62.46	51.57	64.49	65.17	58.93	68.23	67.57	63.81
	+de		60.53	62.56	57.82	66.30	64.35	63.82	72.97	66.01	67.86
	+zh		64.19	60.44	52.10	64.50	64.11	61.86	64.74	66.89	58.59
	+ca		56.30	61.23	66.43	63.78	64.44	66.49	72.46	66.06	66.31
	+ps		49.35	62.72	60.08	57.47	66.58	62.63	65.40	70.94	61.88
	+is		60.23	47.54	65.03	66.18	56.45	65.61	69.74	67.95	65.50
	+my		54.46	58.68	67.25	61.36	61.89	69.03	64.02	64.86	70.89
	+fr+ca	west romance	58.48	46.00	59.36	63.85	55.63	63.36	69.74	67.62	67.21
	+fa+ps	iranic	67.08	46.61	61.29	66.90	55.65	64.19	66.17	67.34	65.63
	+de+is	western-germanic	45.17	57.78	60.10	55.19	63.55	64.34	63.62	69.43	68.14
	+zh+my	sino-tibetan	57.08	62.67	58.97	64.48	65.04	61.64	72.67	67.50	62.65
	+fr+fa+de+zh	high-resource	59.83	49.60	67.78	62.92	57.54	67.42	66.22	62.94	66.76
	+ca+ps+is+my	low-resource	46.50	58.96	65.41	55.95	63.35	66.19	66.87	67.89	65.98
	all		61.64	61.68	62.94	65.21	64.35	66.02	68.91	67.17	68.37

text typically falls between 40% and approximately 95%, indicating that, on average, the backtranslations retain the grooming intent of the conversations. Meanwhile, the bleu scores exhibit a lower range of values, indicating that word choices differ, which, together with semantic similarity, suggests a high-quality backtransaltion for augmentation. Conversely, a higher bleu implies that the original text and the paraphrase are very similar in terms of word usage and could even be identical, yielding poor backtranslation for augmentation.

D Complementary Results

Figure 3 shows the trade-offs between the training efficiency and inference efficacy for baselines when the dataset is augmented with backtranslations using m2m100. As seen, a similar trend is followed as in other translators, including nllb and google (Figure 1). Furthermore, Table 6 shows the values of metrics for the baselines whose delta (Δ) were presented in Table 2.