This research is about online **grooming** that may be offensive or upsetting.

Warning

Enhancing Online Grooming Detection via Backtranslation Augmentation

Online Conversation (Chats)

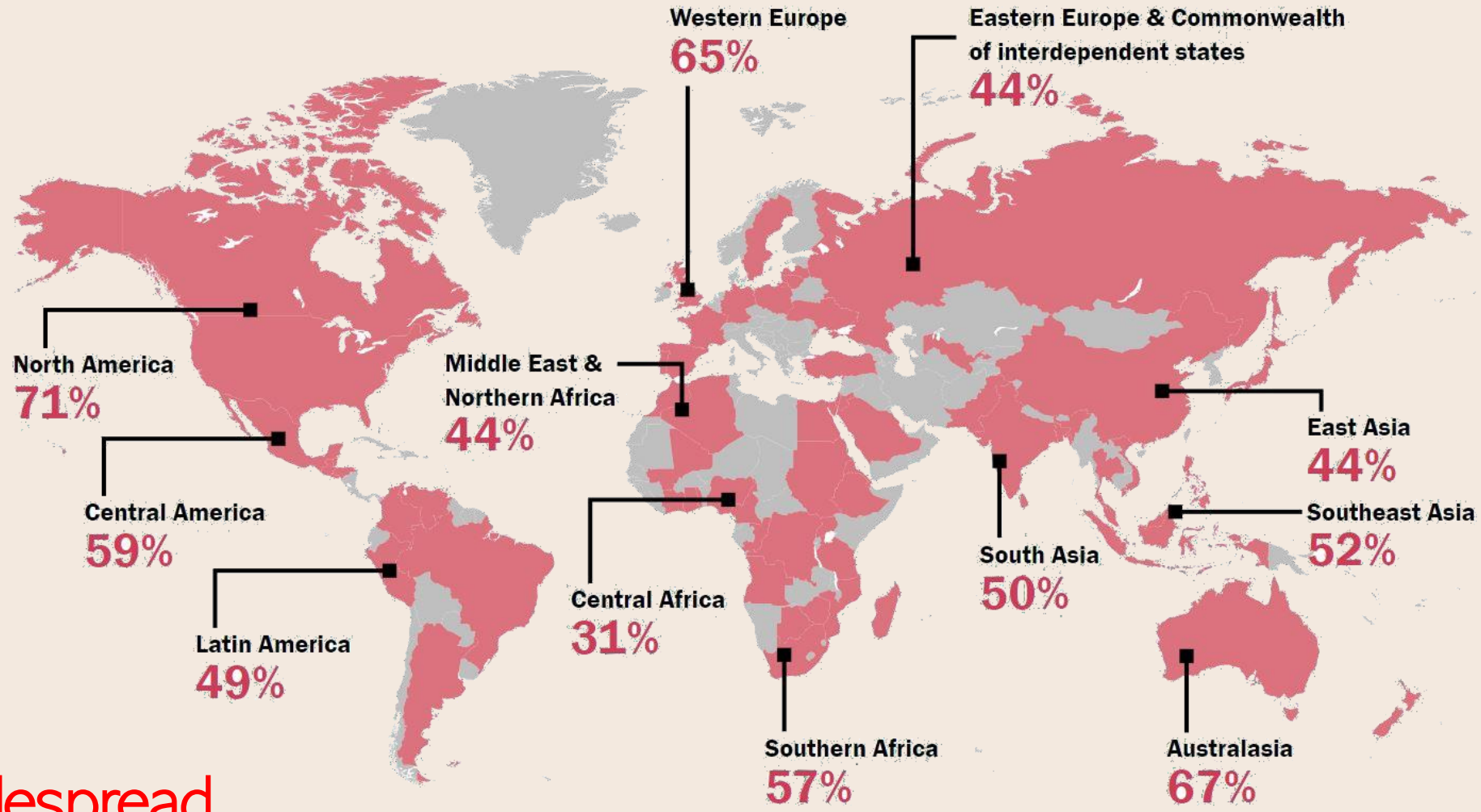
- Social Media
- Messaging Apps
- Dating Sites
- Online Games
- Group Chatrooms

Minors (Kids) w/ little cognitive development

Easy **Preys**

Predators (Abusers)

https://www.childline.org.uk/info-advice/bullying-abuse-safety/online-mobile-safety/online-grooming
https://rumbletalk.com/blog/index.php/knowledge-base/what-are-the-5-chat-types/



Widespread ...

Respondents experienced at least once

Setter et al. Global Threat Assessment Report 2021

Online Grooming Detection

Let $\mathcal{C} = \{c\}$, a mapping function f_θ of parameters θ is desired such that: $f_\theta: \mathcal{C} \rightarrow$ {0: normal, 1: predatory}
Considering g_φ a mapping function from string to a vector of real values $g_\varphi: c \rightarrow \mathcal{R}^d$

$$f_\theta(c) \approx f_\theta(g_\varphi(c))$$

	f_θ	g_φ	
Villatoro-Tello et al. ; A Two-step Approach for Effective Detection of Misbehaving Users in Chats; [Clef 2012]	SVM	Bag-of-words; Text only	First work on PAN-12
Ebrahimi et al.; Detecting predator conversations in social media by deep Convolutional Neural Networks; [Digital Investigation 2016]	SVM, CNN	Concatenated 1-hot vectors; word2vec; glove; Text only	Keeping the order of words; poor distributional vector results
Escalante et al.; Early detection of deception and aggressiveness using profile-based representations [Expert Syst. Appl. 2017]	Naïve Bayes	Profile Specific Representation (PSR)	Early detection
J. Kim et al.; Analysis of Online Conversations to Detect Cyberpredators Using Recurrent Neural Networks [STOC@LREC 2020]	LSTM	Text only	Labelling each message, and classifying them as a whole
Vogt et al.; Early Detection of Sexual Predators in Chats; [ACL/IJCNLP 2021]	Variants of BERT	BERT internal tokenizer; text-only	Segmentation of chat
Nguyen et al.; Fine-Tuning Llama 2 Large Language Models for Detecting Online Sexual Predatory Chats and Abusive Texts; [CoRR 2023]	LLaMA	Text only	Finetuning open-source LLMs
Milon-Flores and Cordeiro; How to take advantage of behavioral features for the early detection of grooming in online conversations [Knowl. Based Syst. 2022]	MLP; KNN; RF; XGB; GBM	BF-PSR (some behavioral features)	Early detection
Chehbouni et al.; Early detection of sexual predators with federated learning [FL-NeurIPS 2022]	LR, KNN	Pre-trained BERT	Early detection
Risch and Krestel; Aggression Identification Using Deep Learning and Data Augmentation [TRAC@COLING 2018]	GRU	Character TF-IDF	Backtranslation Augmentation; Aggression Detection
Parisa Rezaee Borj. Online Grooming Detection on Social Media Platforms." Doctoral Thesis, NTNU (2023)	SVM; NB; RF; etc.	TF-IDF; bag-of-words; Glove; TC	Feature-level Perturbation Augmentation; etc.
Hemmatizadeh et al.; Latent Aspect Detection via Backtranslation Augmentation [CIKM 2023]	many	BF-PSR (some behavioral features)	Backtranslation Augmentation; Aspect Detection
Rajaei et al.; No Query Left Behind: Query Refinement via Backtranslation [CIKM 2024]	T5	Profile Specific Representation (PSR);	Backtranslation Augmentation; Query Refinement

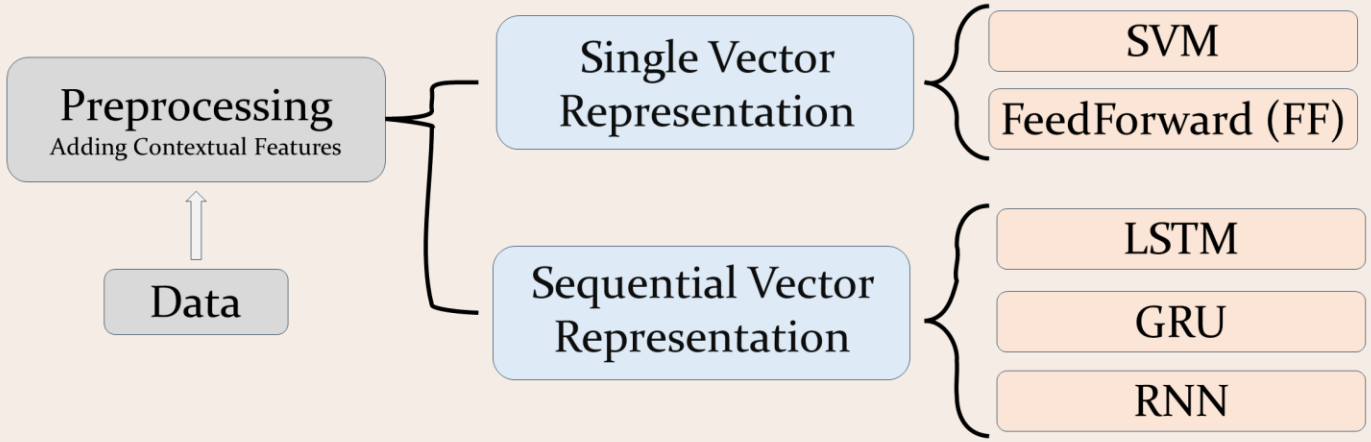
	f_θ	g_φ	
Villatoro-Tello et al. ; A Two-step Approach for Effective Detection of Misbehaving Users in Chats; [Clef 2012]	SVM	Bag-of-words; Text only	First work on PAN-12
Ebrahimi et al.; Detecting predator conversations in social media by deep Convolutional Neural Networks; [Digital Investigation 2016]	SVM, CNN	Concatenated 1-hot vectors; word2vec; glove; Text only	Keeping the order of words; poor distributional vector results
Escalante et al.; Early detection of deception and aggressiveness using profile-based representations [Expert Syst. Appl. 2017]	Naïve Bayes	Profile Specific Representation (PSR)	Early detection
Chehbouni et al.; Early detection of sexual predators with federated learning [FL-NeurIPS 2022]	LR, KNN	Pre-trained BERT	Early detection
Risch and Krestel; Aggression Identification Using Deep Learning and Data Augmentation [TRAC@COLING 2018]	GRU	Character TF-IDF	Backtranslation Augmentation; Aggression Detection
Parisa Rezaee Borj. Online Grooming Detection on Social Media Platforms." Doctoral Thesis, NTNU (2023)	SVM; NB; RF; etc.	TF-IDF; bag-of-words; Glove; TC	Feature-level Perturbation Augmentation; etc.
Hemmatizadeh et al.; Latent Aspect Detection via Backtranslation Augmentation [CIKM 2023]	many	BF-PSR (some behavioral features)	Backtranslation Augmentation; Aspect Detection
Rajaei et al.; No Query Left Behind: Query Refinement via Backtranslation [CIKM 2024]	T5	Profile Specific Representation (PSR);	Backtranslation Augmentation; Query refinement

Limits of Literature

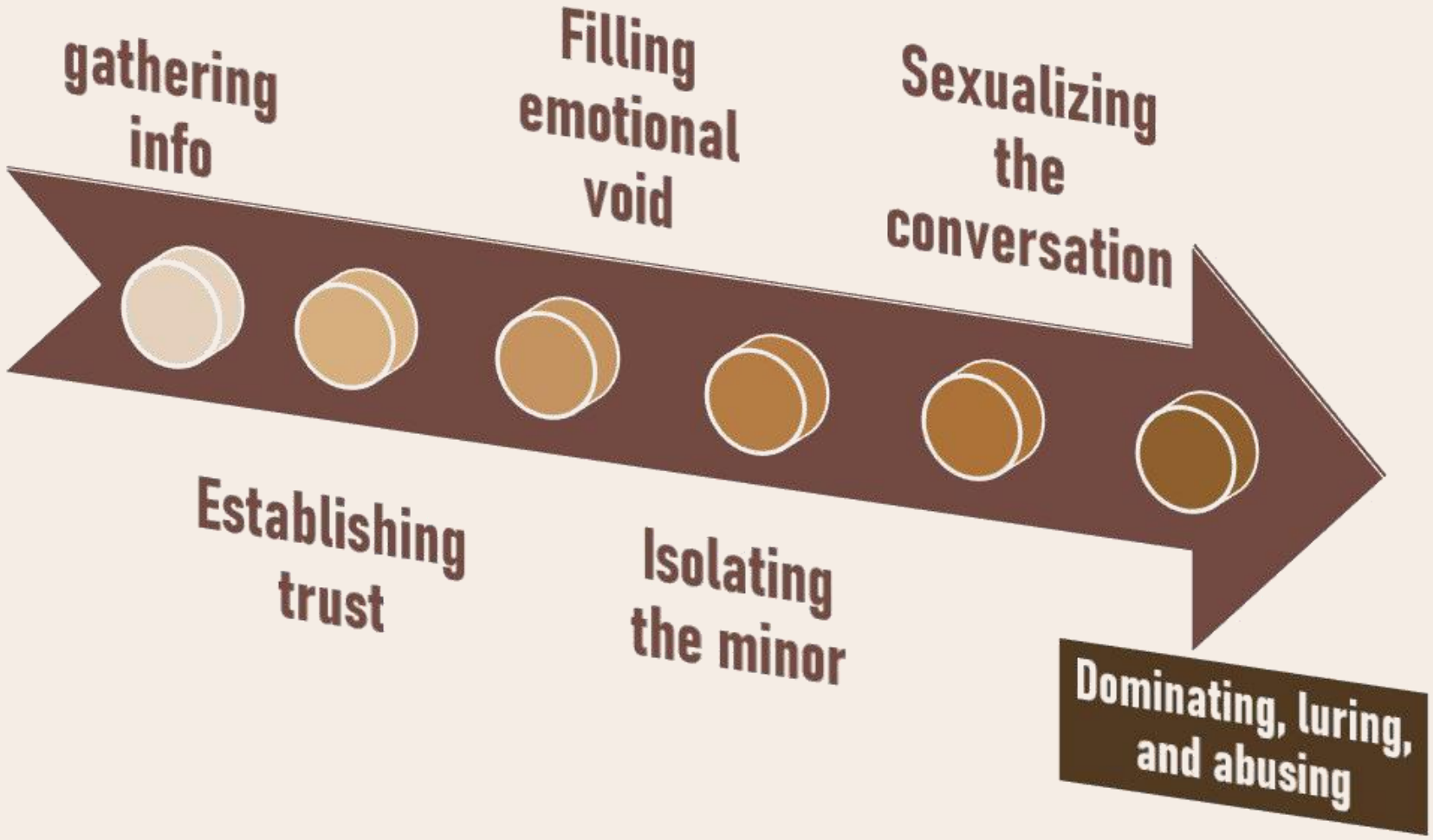
- Foregoing **conversational** features
- Prioritizing **Recall** over **Precision**
- **Irreproducibility!**
- Almost **no augmentation**

Osprey

A Reference Framework for Detection of Online Grooming
+ Conversation Features
+ Backtranslation Augmentation



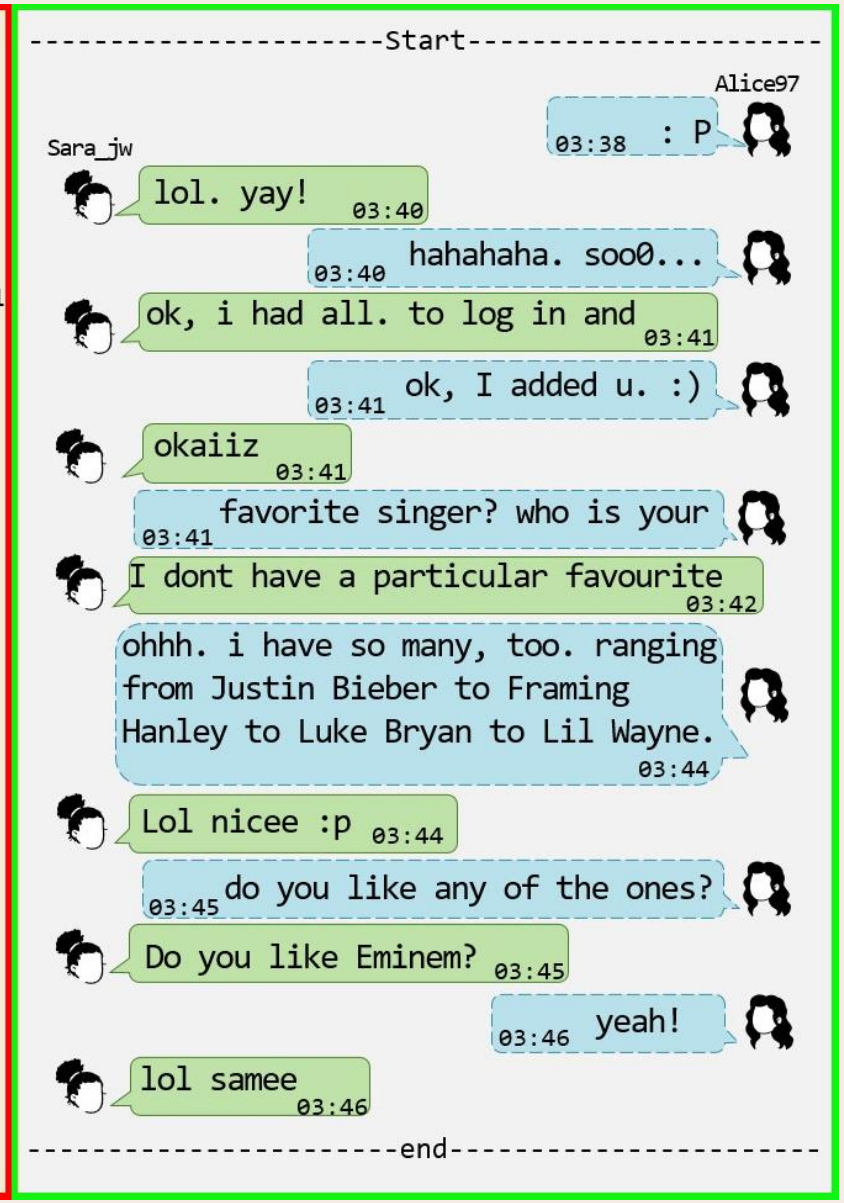
Phase 0: https://github.com/fani-lab/osprey



Kaylor et al. 2017. Sexual offenders contacting children online: an examination of transcripts of sexual grooming. Journal of Sexual Aggression

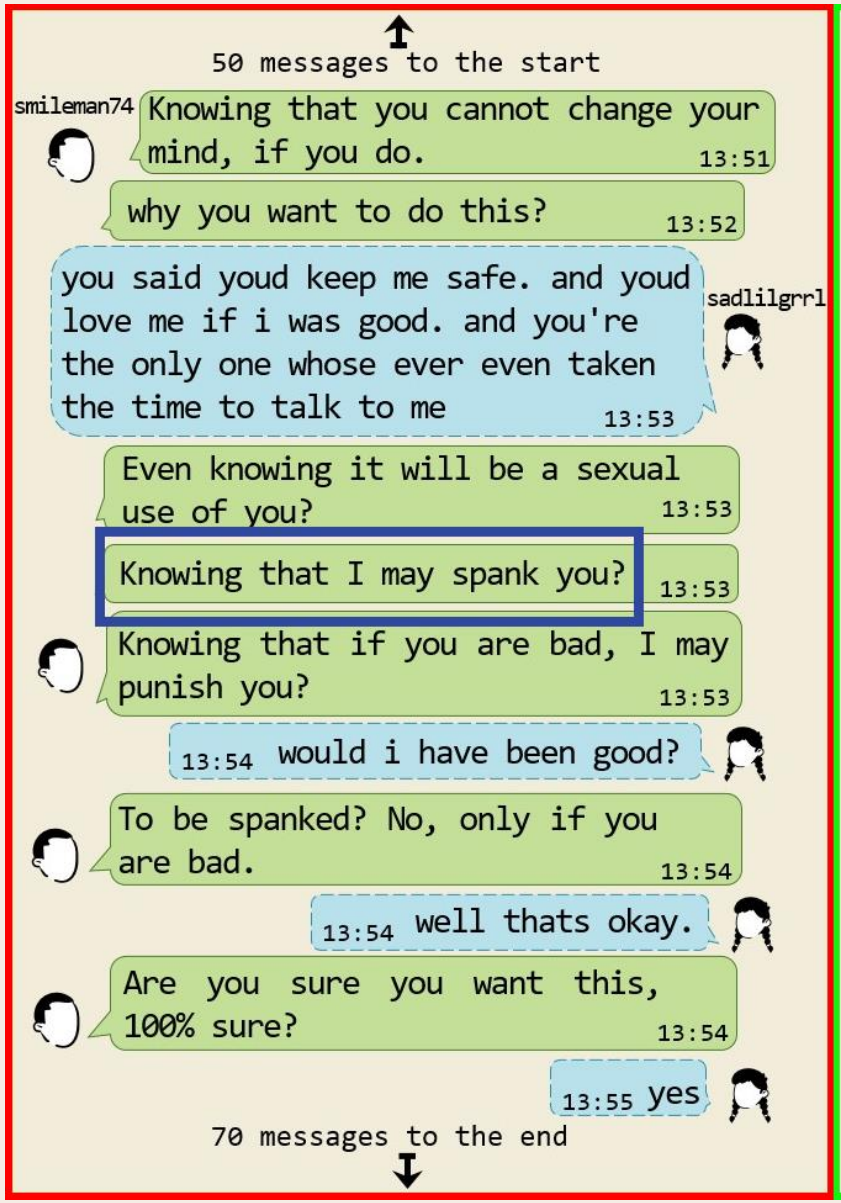
Long …

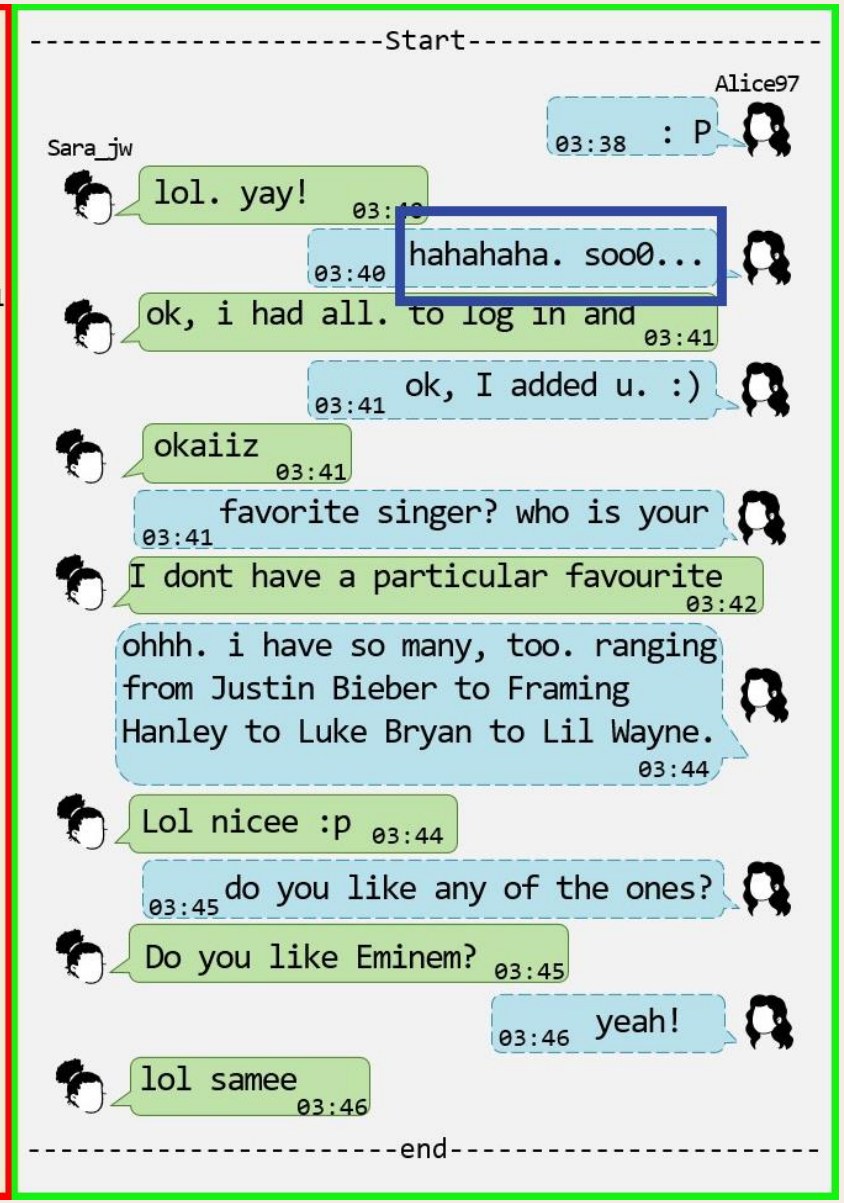

50 messages to the start
smileman74
Knowing that you cannot change your mind, if you do. 13:51
why you want to do this? 13:52
sadlilgrrl
you said youd keep me safe. and youd love me if i was good. and you're the only one whose ever even taken the time to talk to me 13:53
Even knowing it will be a sexual 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----------

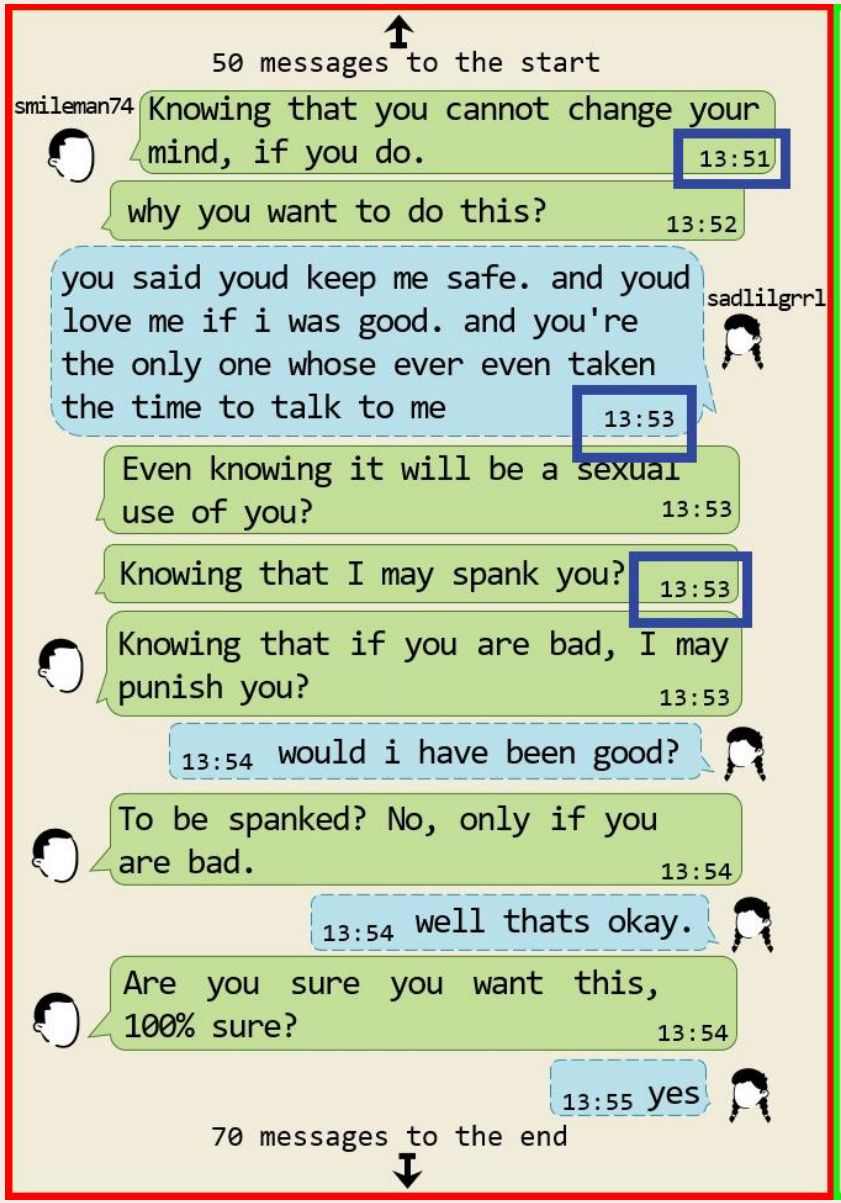
Short.

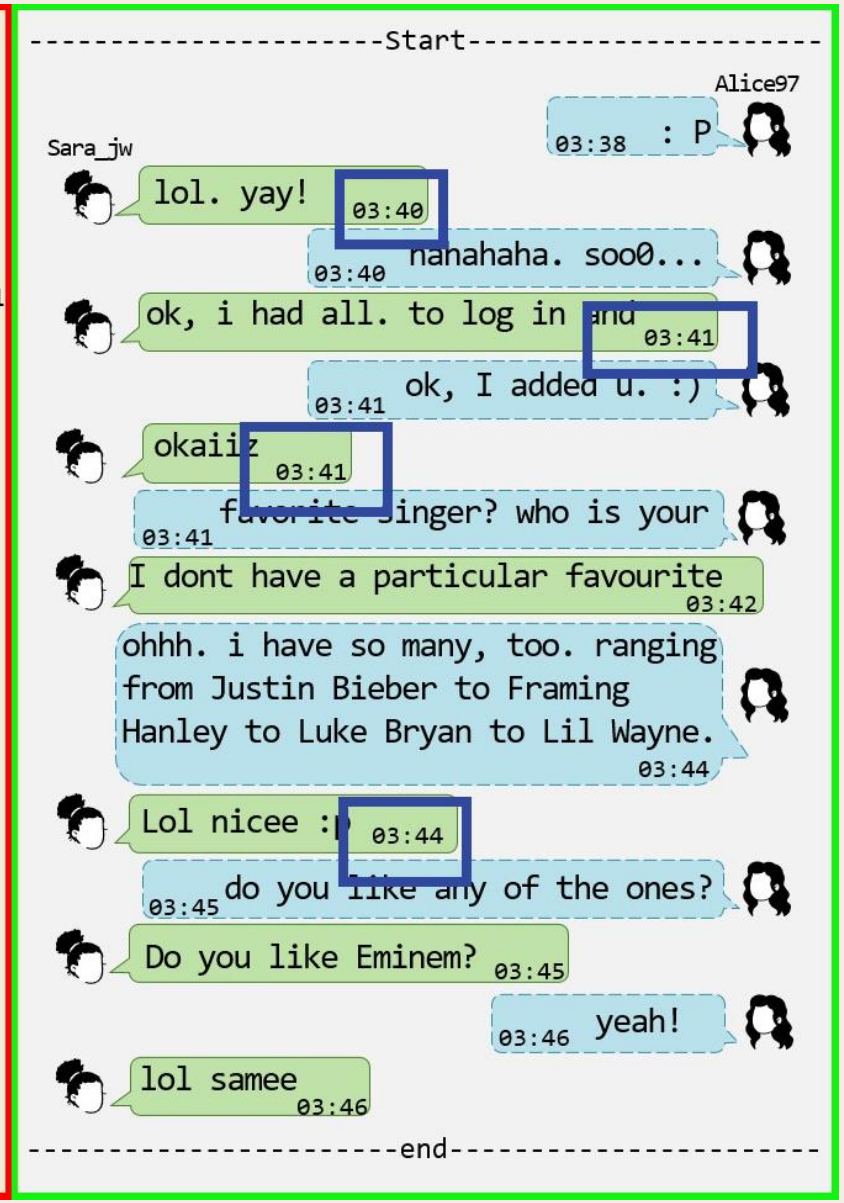
Language Pattern





Turn Taking Pattern





- RQ1: **Improved Precision and Recall**
- RQ2: **Recurrent** models **generally better** irrespective of feature representations
- RQ3: **GRU**'s better Recall compared to LSTM
- **RQ4: GRU + DistilRoBERTa + Conversation Features → Higher Recall while maintaining Precision**

Phase 1: Findings: Conversational Feature: **time** + **n_authors**

Low Precision vs. Low Recall

High false positives
Arrest of innocent person
Cleared by further investigation

High false negatives
Unable to capture a predator
Abuse of more kids

Online Grooming Detection

Phase 0: A Reference Framework
Phase 1: Sequence of messages as sequence of embeddings + Conversation features
Phase 2: Backtranslation data augmentation

Polysemy

English	like two guys doing each other?
Germany	*wie zwei typen, die es miteinander treiben?*
Backtranslation	like two guys **having sex**?

Normalization

Original	u really dont mind that i'm 13 rite?
French	*Ça ne te dérange pas que j'aie 13 ans?*
Backtranslation	**Doesn't it bother** you that I'm 13?

Context-aware Synonyms

Original	i feel little aroused
Germany	*ich fühle mich ein wenig erregt*
Backtranslation	i'm feeling a little **turned on**

Latent Terms

Original	having it with minor
French	*l'avoir avec mineur*
Backtranslation	Having **sex** with a minor

Examples … though few quantitative experiments!

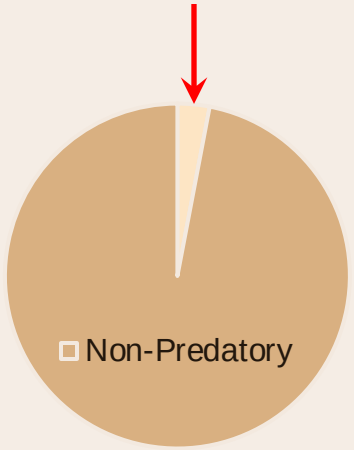
Setup

- Models: GRU and LSTM
- Message vectorizer: DistilRoBERTa + time + n_author (Conversation Features)
- Backtranslation augmentation but only for sparse (minority) predatory conversations
- RQ5: Augmentation improves Precision and Recall?
  - RQ6: Language (family) matters?
  - RQ7: Language richness matters?
  - RQ8: Translator matters?
- Dataset: PAN12

- PAN at CLEF 2012 (PAN12): **Kinda synthetically curated w/ controversial method!**
  - Predatory Conversations: Decoys from Perverted-Justice (Entrapment)
  - Non-predatory chats: Omegle (an online chatroom)
  - Only in English

Ethical Issues ...

- THE only accessible dataset
- Others are not accessible
  - PANC
  - ChatCoder2



PAN 2012

	Raw		Filtered (at least 2 author and 6 turns)	
	Train	Test	Train	Test
n_conversations	66,927	155,128	16,529	38,246
n_predatory_conversations	2,016	3,737	957	1,698
n_binary_predatory conversations	2,016	3,737	957	1,698
n_nonbinary_predatory conversations	0	0	0	0
Avg n_msgs in a predatory conversation	60.73	90.07	**80.68**	**71.48**
Avg n_msgs in a normal conversation	12.74	12.86	41.73	41.78

	Predatory	Non-Predatory
Avg time elapsed between each message	2.12±0.64	**2.00±2.50**
Avg number of consecutive messages from same participant	**1.43±8.61**	0.91±1.41

Languages ...

Resources	Language	Family
High	(en) English	West Germanic
	(fr) French	Western Romance
	(fa) Farsi	Iranic
	(de) Germany	West Germanic
	(zh) Chinese	Sino-Tibetan
Low	(ca) Catalan	Western Romance
	(ps) Pashto	Iranic
	(is) Icelandic	West Germanic
	(my) Myanmarese	Sino-Tibetan

Translators …

	google	m2m100*	nllb*
n_languages	133	101	196
Model Card	No!	Yes	Yes
n_parameters	unknown	1.2 Billions	3.3 Billions
License	Closed Source	MIT	CC BY-NC
Owner	Google	Meta	Meta
Architecture	Transformers+RNN	Transformers	Transformers

https://cloud.google.com/translate/docs/reference/rest
Fan, A., Bhosale, S., et al.: Beyond english-centric multilingual machine translation. J. Mach. Learn. Res. 22 (2021)
Team, N., Costa-jussà, M.R., et al.: No language left behind: Scaling human-centered machine translation. CoRR (2022)

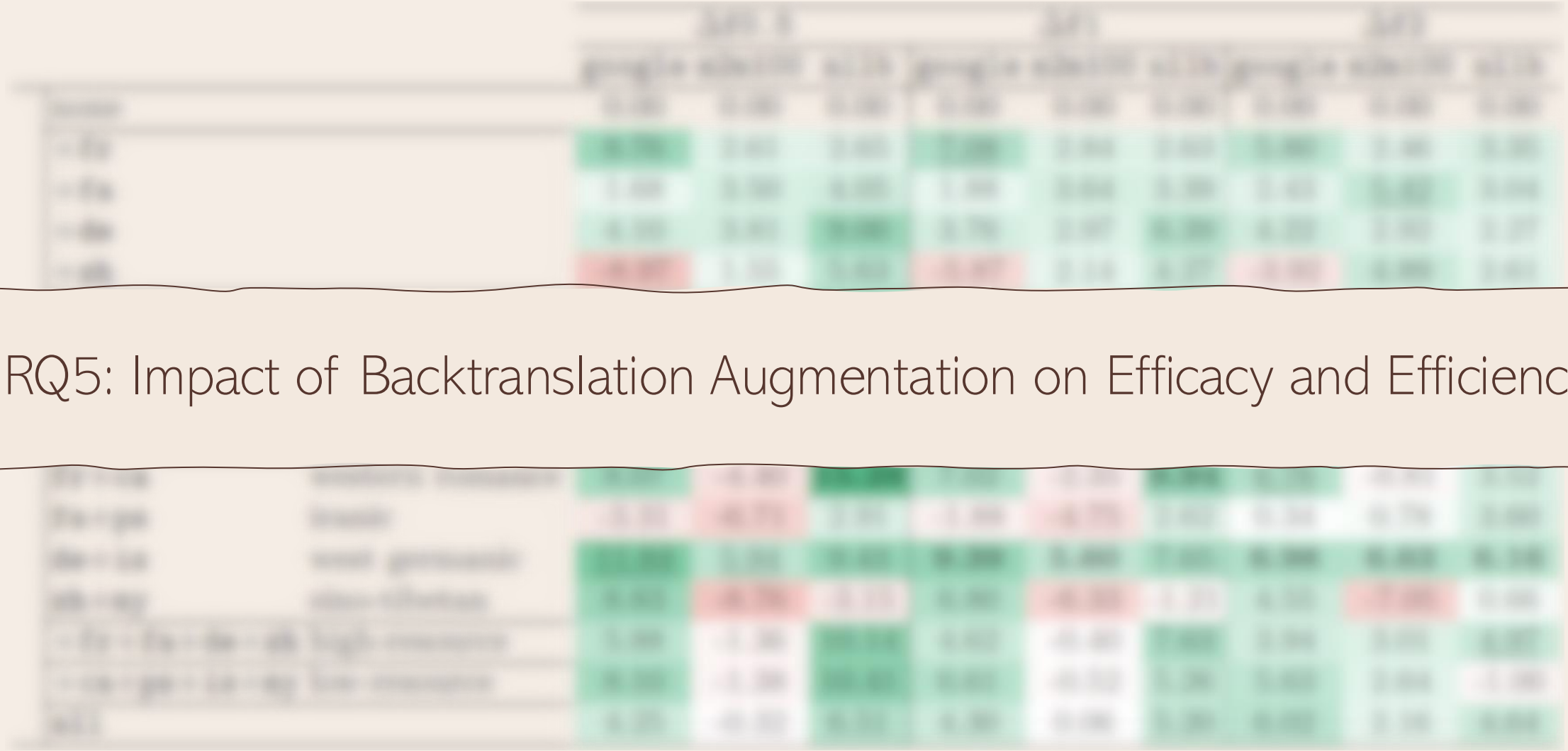
Low Precision vs. Low Recall

High false positives
Arrest of innocent person
Cleared by further investigation

High false negatives
Unable to capture a predator
Abuse of more kids

			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

RQ5: Impact of Backtranslation Augmentation on Efficacy and Efficiency



RQ5

			Δf0.5			Δf1			Δf2		
			google	m2m100	nllb	google	m2m100	nllb	google	m2m100	nllb
en	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
	+ny		**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+ny	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+ny	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

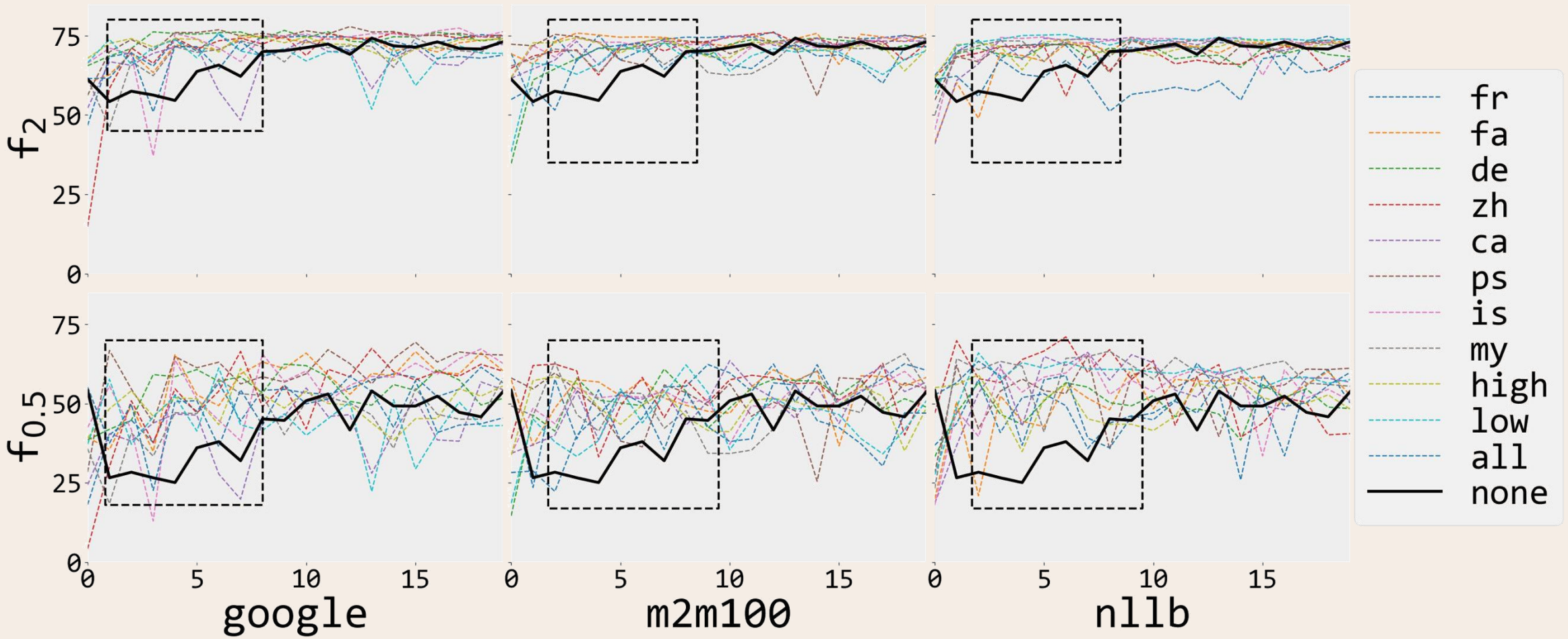
RQ5

		Δf0.5			Δf1			Δf2		
		google	m2m100	nllb	google	m2m100	nllb	google	m2m100	nllb
none		0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
en	+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
	+ny	**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+ny 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+ny 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

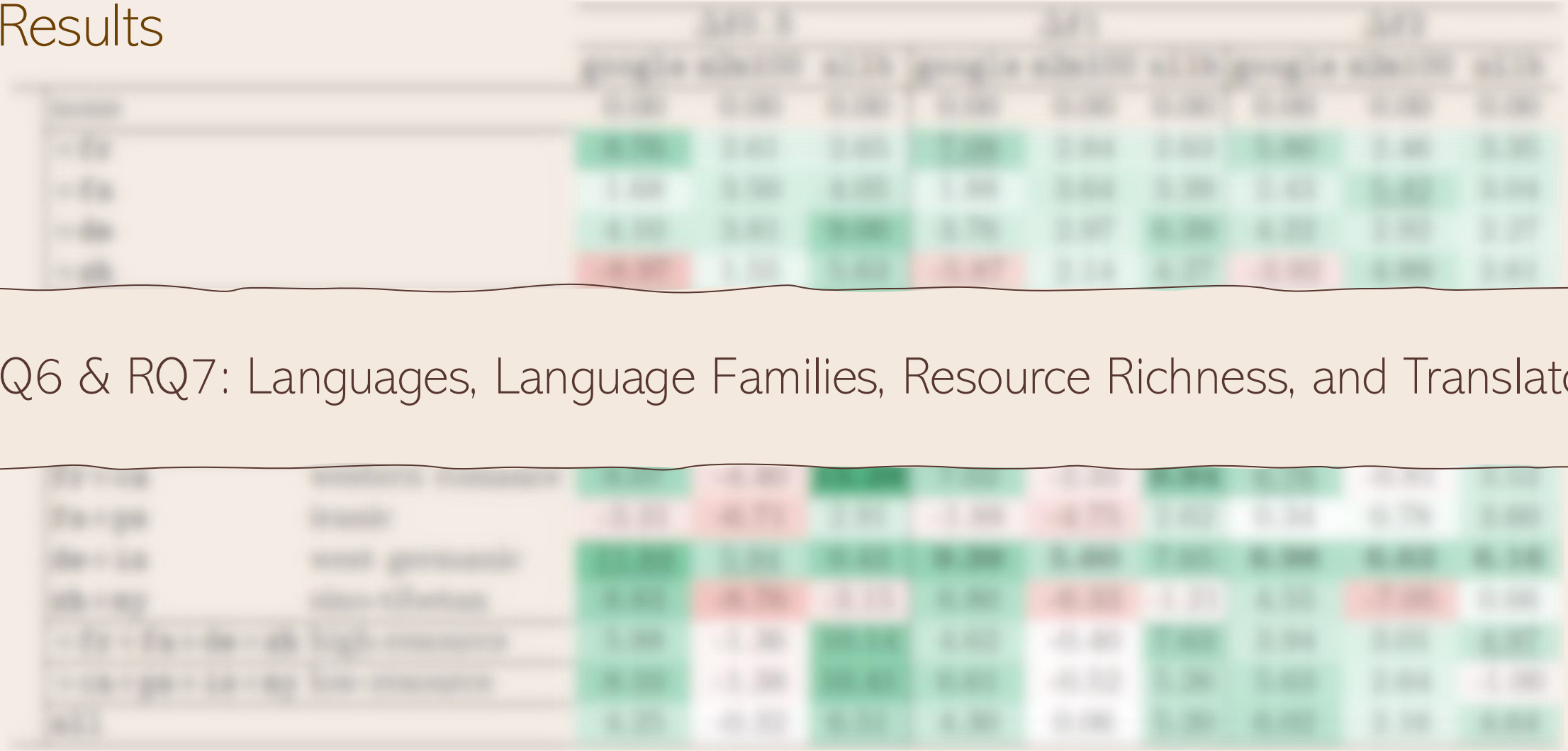
RQ5

			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

RQ5


f_2
$f_{0.5}$
0
25
50
75
0
5
10
15
google
m2m100
nllb
fr
fa
de
zh
ca
ps
is
my
high
low
all
none

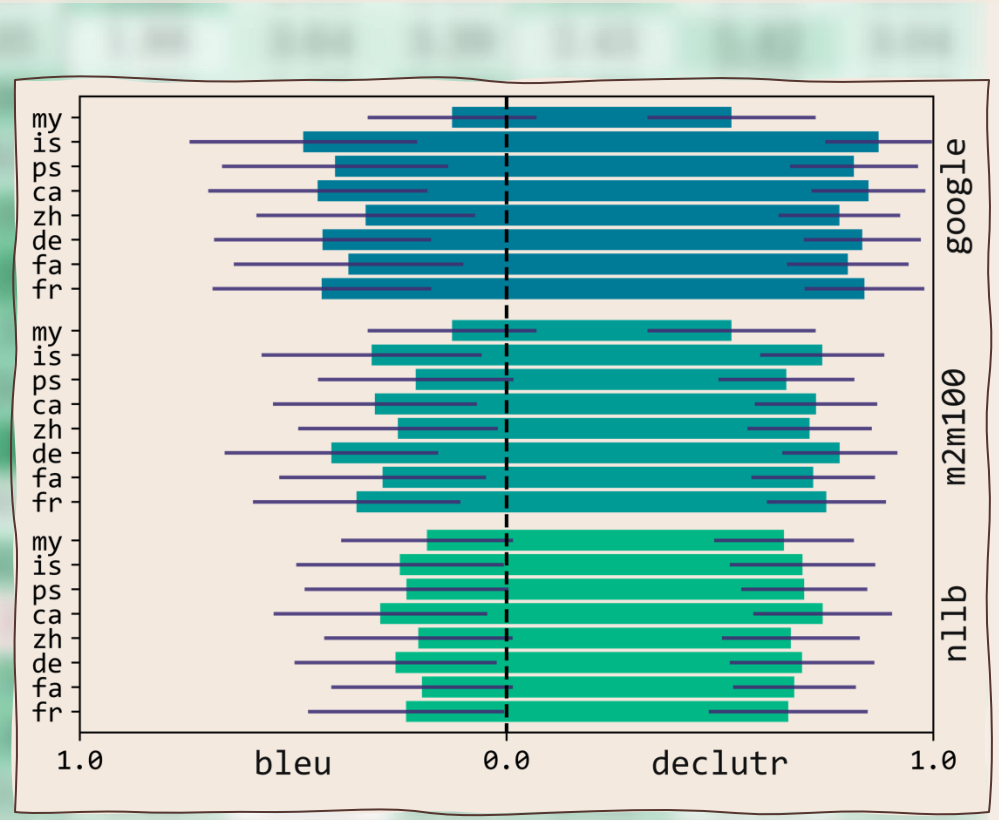
Results



RQ6 & RQ7: Languages, Language Families, Resource Richness, and Translators

A Good Backtranslation for Augmentation …

- Same Semantic (high declutre)
- Different Wording (low bleu)



Giorgi et al. 2021. Declutr: Deep contrastive learning for unsupervised textual representations. In ACL,

RQ6 & RQ7

+fa		1.68	3.50	4.05	1.88	3.64	3.39	2.43	5.42	3.04
+ps		9.57	-5.32	13.60	6.48	-3.79	7.58	3.04	1.56	0.52
fa+ps	iranic	-3.31	-6.71	2.91	-1.88	-4.75	2.62	0.34	0.78	3.60

RQ6 & RQ7

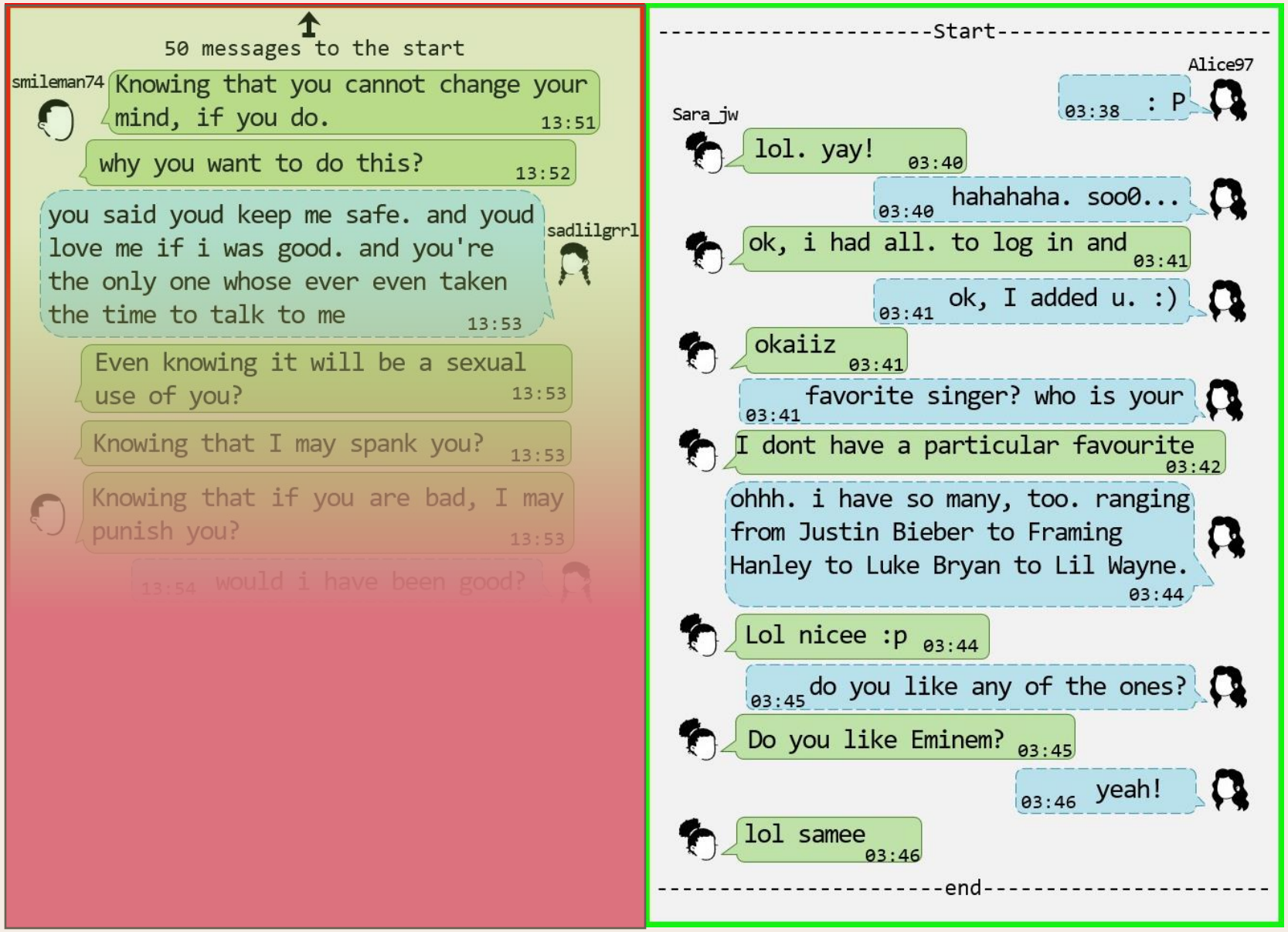
+de		4.10	3.81	9.00	3.76	2.97	6.39	4.22	2.92	2.27
+is		3.73	-2.24	10.14	4.04	-0.68	6.98	6.27	2.25	3.24
de+is	west germanic	11.84	5.94	9.43	**9.39**	**5.60**	7.65	**6.98**	**6.63**	**6.16**

RQ6 & RQ7

+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

Future Works

Early Detection



Fani's Lab!, School of Computer Science, University of Windsor, Canada


Hamed
Hossein