Integrating linguistic knowledge into DNNs: Application to online grooming detection – Supplementary Materials

Anonymous Authors

1 Summary of the OG corpus's content

Our OG corpus contains chat logs from 308,335 users (623 groomer and 307,712 non-groomers). A user may be engaged into one or several *conversations* with other individuals. Some distinct conversations between the same users may be related, one being the continuation of another in a different chat session. This is not accounted for in our DNNs, and individual conversations are classified independently of others, even related ones. We refer to a group of related conversations as a *long-term digital relationship* between the users. Each conversation is made up of *messages*. The statistics of conversations, messages, and long-term digital relationships are summarised in Tables 1 to 4.

The effectiveness of WSR modification using selected variants is emphasised by the frequency of usage of these variants within the corpus: modifying the WSR space around frequently used words may have a larger impact on subsequent text analysis than adjusting it around rare words. The frequency of any word w_i is calculated as: $\mathcal{F}_{w_i} = \frac{\text{count}(w_i)}{N}$, where N is the total number of words in the corpus. Occurrence frequencies for the selected variants and other words are provided in Table 5. We see that the average frequency of selected variants is significantly larger than for other words in the corpus by two orders of magnitude. This may explain in part the effectiveness of our selective WSR normalisation.

2 OG Processes

There are total of seven different OG processes used in both base models. These can be characterised into 5 main categories as seen in Table 6. For each of these categories, we report the number of annotated collocations, and an example collocate (in italic text) in its context.

Table 1: Statistics of chat logs in both OG and non-OG classes.

	OG	Non-OG	Total
# Users	623	307,712	308,335
# Conversations	6,204	216,242	222,446
# Messages	648,463	3,433,824	4,082,287
# Words	27,388	134,075	161,463

Table 2: Statistics of conversations in the OG class.

Stats Name	Min / Max	Mean (STD)
# Messages	1 / 17,511	215.26 (688.83)
# Words	1 / 81,705	1,009.93 (3,231.16)

Table 3: Statistics of conversations in the non-OG class.

Stats Name	Min / Max	Mean (STD)
# Messages	1 / 1,023	12.70 (23.09)
# Words	1 / 122,763	93.84 (489.22)

3 Experiment Environment

3.1 Hardware/Software

- GPU: Nvidia GeForce RTX 2080 Ti
- Deep Learning Library: PyTorch 1.3.1
- Python: 3.7.5
- OS: CentOS 7

3.2 Model parameters

- Optimiser: RMSprop (using the default learning rate)
- Scheduler: Cyclic Learning Rate (RMSprop base, 5e-3 max)
- Early stopping (tracking validation loss metric)
  - Base model #1: 50 epochs
  - Base model #2: 100 epochs
- Batch-size:
  - Base model #1: 128
  - Base model #2: 8 (with gradient accumulation over 16 batches)
- Gradient clipping: ±0.5
- WSR Dimensionality:
  - Base model #1: 300

Table 4: Statistics of chat logs per individual groomer.

Stats Name	Min / Max	Mean (STD)
# Conversations	1 / 272	8.422 (20.13)
# Messages	8 / 27,025	2,032.49 (2,876.69)
# Words	1 / 64,841	5487.89 (8129.02)

Table 5: Comparison of occurrence frequencies for selected variants and all words in the corpus.

	Mean	Standard Deviation
Word Frequency	$3{,}922 \cdot 10^{-05}$	0,001
Variant Frequency	0,001	0,002

Table 6: Processes used by groomers in order to establish a connection with a child.

OG Process		# Coll.	Collocate Usage in Context
Approach: Reference to the groomer's intention to meet with the child.		622	”...lots more peaceful lol i know rightand i *could come over* right?”
Compliance Testing: Checking likelihood of victim agreeing to proposed behaviour.		23	”do u *like* talking to *older guys*?”
Deceptive Trust Development: Building trust with the victim with the ulterior motive of eliciting sexual activities.	Activities	61	”ok so any *plans* for *this weekend*?”
	Personal Information	33	”so can you *tell me how* me about how far it is from you to allendale”
	Relationship	357	”i couldn't *stop thinking about* u”
Isolation: Groomer distance the victim physically/emotionally from their support circle.		112	”*we meet* some *where* alone near your neighborhood...”
Sexual Gratification: Groomers attempt to involve their victim in sexual talk/activities.		892	”just you and me touching each other ... feeling each other”

  - Base model #2: 768

- Pre-trained Glove embedding: 840B-300D crawl.
- Pre-trained XLNet: xlnet-base-cased (`https://github.com/zihangdai/xlnet/` & `https://github.com/huggingface/transformers`).
- Out-of-vocabulary (OOV) default embedding vector: random coordinates following a normal distribution of mean 0 and std 1 (i.e. close to the centre of the embedding manifold).
- LSTM hidden size (both base models): 256
- LSTM # layers (both base models): 2
- Classification layer (both base models): 1 fully connected layer
- Dropout rate between LSTM layers and classification layer: 20%
- Training/validation split: 70/30 (stratified)
- Maximum sequence length: 2,000 - sorted & bucketed batches
- λ (weights of the additional losses): 1, 1, 1/3 for Stimulation of LSTM, Stimulation of attention, and Aux. OG process estimations, respectively
- Random weight initialisation seed: 42

3.3 Tokenisation details

As a standard step in NLP, we tokenise named entities prior to OG classification. Our criteria for tokenisation and word replacement are as follows:

- All Spacey entities (see `https://spacy.io/api/annotation#named-entities`) are encoded to their respective categories, in addition to *LONGWORD* for words with more than 35 characters, and *URL* for URLs.
- Stemming using SnowballStemmer (NLTK).
- Tokenised using Spacey 'en' (English) model (`https://spacy.io/models`).
- Tokens with less than 5 occurrences in the corpus are replaced by OOV.

4 Further experimental results

We provide here additional metrics and experimental results. The evaluations of the individual CL-augmentations are further detailed in Table 7 with additional accuracy, precision, and recall metrics. Accuracy is also provided for all compared OG classifiers in Table 8 and progressive additions of prior knowledge in Table 11. Accuracy is to be considered carefully considering the strong class imbalance in our dataset.

We verify in Table 10 that the 3 strategies for selective normalisation of WSR based on word variants preserve the average pairwise distances between non-pairs of variants. This is a pre-requisite for the WSR space to preserve its semantic descriptive power, although not a sufficient condition, as shown in Section 5.1.

In Table 11, integration strategies are added to base model #1 iteratively to measure the performance improvements with each augmentation.

Table 7: Impact of each CL augmentation on OG classification. Bold are improved results.

Model	Strategy		Accuracy (%)	Precision	Recall	F_1
#1	No augmentation		99.12	0.867	0.794	0.829
	Supervised WSR modification		98.96	0.834	0.765	0.798
	Manifold learning		98.91	0.849	0.723	0.820
	Elastic pulling		99.19	0.878	0.808	0.841
	Aux. OG process detection		99.13	0.890	0.768	0.825
	Stim. attention	supervised	99.06	0.839	0.804	0.821
		excitation (Eq. 3)	99.04	0.822	0.817	0.820
		excitation (Eq. 4)	99.16	0.870	0.808	0.838
		superv.+excit. (Eq.3)	99.16	0.859	0.819	0.838
		superv.+excit. (Eq.4)	99.15	0.929	0.741	0.824
	Stim. LSTM	supervised	99.17	0.924	0.752	0.829
		excitation	99.10	0.856	0.797	0.825
		superv.+excit.	99.20	0.906	0.781	0.839
#1 w. GloVe	No augmentation		99.15	0.879	0.789	0.832
	Supervised WSR modification		99.00	0.868	0.739	0.798
	Manifold learning		99.00	0.896	0.708	0.791
	Elastic pulling		99.15	0.880	0.772	0.823
#2	No augmentation		99.41	0.900	0.871	0.886
	Aux. OG process detection		99.42	0.918	0.861	0.889
	Stim. attention	supervised	99.42	0.919	0.862	0.890
		excitation (Eq. 3)	99.41	0.894	0.885	0.889
		excitation (Eq. 4)	99.43	0.916	0.866	0.891
		superv.+excit. (Eq.3)	99.39	0.891	0.881	0.886
		superv.+excit. (Eq.4)	99.40	0.918	0.862	0.889
	Stim. LSTM	supervised	99.47	0.938	0.857	0.896
		excitation	99.40	0.896	0.896	0.887
		superv.+excit.	99.49	0.960	0.846	0.899

Table 8: Comparative evaluation of OG classification methods

Method	Acc (%)	Precision	Recall	AUPR	F_1	$F_{0.5}$
Naive Bayes	91.69	0.240	**0.974**	0.727	0.385	0.283
SVM	98.22	**0.997**	0.337	0.748	0.504	0.716
Decision Tree	98.28	0.693	0.642	0.637	0.667	0.682
Random Forest	98.38	0.987	0.400	0.718	0.569	0.763
Liu et al. 2017	99.11	0.919	0.735	0.885	0.817	0.875
BERT	98.86	0.837	0.711	0.815	0.711	0.808
Base model #1	98.96	0.867	0.794	0.867	0.829	0.851
Base model #1 + L1 Regularisation	99.08	0.880	0.759	0.857	0.815	0.853
Base model #1 + L2 Regularisation	99.18	0.896	0.783	0.890	**0.992**	0.871
Base model #2	99.41	0.900	0.871	0.940	0.886	0.894
Base model #2 + L1 Regularisation	99.38	0.885	0.881	0.940	0.883	0.883
Base model #2 + L2 Regularisation	99.42	0.913	0.865	0.941	0.888	0.903
Augmented model #1	99.25	0.930	0.777	0.924	0.847	0.895
Augmented model #2	**99.49**	0.953	0.853	**0.948**	0.900	**0.931**

Table 9: Progressive additions of CL-augmentations to a simple LSTM model similar to base model #1 with no pre-training of WSR

Method	Acc (%)	Precision	Recall	AUPR	F_1	$F_{0.5}$
Standard LSTM	98.96	0.850	0.741	0.808	0.792	0.826
+ Superv. & excit. LSTM	99.20	0.933	0.757	0.872	0.836	0.891
+ Elastic pulling	99.22	0.913	0.783	0.883	0.843	0.884
+ Superv. & excit. attn	99.21	0.915	0.779	0.913	0.841	0.884

Table 10: Average distance between pairs of selected variants $\overline{\mathcal{D}}_{var}$ and all other pairs of words $\overline{\mathcal{D}}_{non\,var}$ in the WSR spaces

Method	$\overline{\mathcal{D}}_{var}$	$\overline{\mathcal{D}}_{non\,var}$
Base Model #1's original WSR	3.72	2.86
Supervised WSR modification	0.91	2.78
Manifold Learning	1.29	2.86
Elastic Pull	0.61	2.82

Table 11: Progressive additions of CL-augmentations to base model #1 with no WSR pre-training

Method	Precision	Recall	AUPR	F_1	$F_{0.5}$
Standard LSTM	0.850	0.741	0.808	0.792	0.826
+ Superv. & excit. LSTM	0.933	0.757	0.872	0.836	0.891
+ Elastic pulling	0.913	0.783	0.883	0.843	0.884
+ Superv. & excit. attn	0.915	0.779	0.913	0.841	0.884