Appendix: Categorizing Sexism and Misogyny through Neural Approaches

PULKIT PARIKH and HARIKA ABBURI, IIIT-Hyderabad, India
NIYATI CHHAYA, Adobe Research, India and IIIT-Hyderabad, India
MANISH GUPTA* and VASUDEVA VARMA, IIIT-Hyderabad, India

ACM Reference Format:
Pulkit Parikh, Harika Abburi, Niyati Chhaya, Manish Gupta, and Vasudeva Varma. 2020. Appendix: Categorizing Sexism and Misogyny through Neural Approaches. 1, 1 (December 2020), 6 pages.

1 HYPER-PARAMETER VALUES

Using experiments on a validation set, which was merged into the training set during the test runs, for each method, we choose the values of (up to) four hyper-parameters: the LSTM dimension, the attention dimension, the number CNN filters for kernel sizes 2, 3, and 4 each, and the k value in k-max-over-time pooling. The values used for all proposed methods and deep learning baselines for which we report results in our paper are provided in Tables 1, 2, 3, 4, 5, 6, and 7. The same LSTM and attention dimensions are used in all parts of a model. Table 3 also contains the layer sizes for the stacked autoencoder models.

The hyper-parameter values for the traditional machine learning baselines for sexism classification are as follows. For SVM, we set soft margin (C) to 1.0. For RF (Random Forest) and GBT (Gradient Boosting Trees), we use 100 estimators. For extracting character and word n-grams, the maximum number of features used, word n-gram range, and character n-gram range are set to 10000, (1,2), and (1,5) respectively. The hyper-parameter values for misogyny classification are the same as above except that the number of estimators for RF and GBT is 50.

For misogyny detection, the hyper-parameter values for the traditional ML methods are as follows. For SVM, 1 is used as the value of soft margin (C). For RF and GBT, we set the number of estimators to 100. For extracting character and word n-grams, the maximum number of features used, word n-gram range, and character n-gram range are 5000, (1,3), and (1,5) respectively.

For the two methods that achieved the best results for Subtask A and Subtask B of the Evalita 2018 shared task on Automatic Misogyny Identification (AMI) that we use as baselines for misogyny detection and misogyny classification respectively, we use the hyper-parameter values mentioned in the corresponding papers.

*The author is also an applied researcher at Microsoft.

Authors' addresses: Pulkit Parikh, pulkit.parikh@research.iiit.ac.in; Harika Abburi, harika.a@research.iiit.ac.in, IIIT-Hyderabad, India; Niyati Chhaya, nchhaya@adobe.com, Adobe Research, India, IIIT-Hyderabad, India; Manish Gupta, manish.gupta@iiit.ac.in; Vasudeva Varma, vv@iiit.ac.in, IIIT-Hyderabad, India.

Manuscript submitted to ACM

Table 1. Hyper-parameter values corresponding to multi-label sexism classification results with the EBCE loss for deep learning methods (the sub-columns for the proposed methods specify different configurations as to the ways in which the input is embedded and/or processed; different colors/styles represent different sentence-level concatenations in a method; colors/styles have no connections across rows)

		Approach				**LSTM dim.**	**Attention dim.**	**#CNN filters of each kernel size**	**k in k-max-over-time pooling**
Baselines (with ELMo)		biLSTM-Attention				200	300	N.A.	N.A.
		Hierarchical-biLSTM-Attention				300	400	N.A.	N.A.
		CNN-biLSTM-Attention				300	400	100	N.A.
		C-biLSTM				N.A.	N.A.	N.A.	N.A.
		CNN-Kim				N.A.	N.A.	150	N.A.
		biLSTM				300	N.A.	N.A.	N.A.
		BERT-biLSTM-Attention				200	500	N.A.	N.A.
		USE-biLSTM-Attention				300	600	N.A.	N.A.
		InferSent-biLSTM-Attention				100	200	N.A.	N.A.
Proposed methods		***biL-att* applied on**	***c-k-max* applied on**	***c-biL-att* applied on**	**Text Encoder**				
	Flat	ELMo			tBERT	200	400	N.A.	N.A.
		ELMo, GloVe			tBERT	200	400	N.A.	N.A.
		concat(ELMo, GloVe)			tBERT	200	500	N.A.	N.A.
		Ling, ELMo, GloVe			tBERT	200	300	N.A.	N.A.
		ELMo, GloVe			tBERT, USE	200	400	N.A.	N.A.
		Ling, fastText, ELMo, GloVe			tBERT, USE, InferSent	200	500	N.A.	N.A.
			ELMo		tBERT	N.A.	N.A.	100	1
			ELMo, GloVe		tBERT, USE	N.A.	N.A.	200	1
				ELMo, GloVe	tBERT	200	400	150	1
		ELMo	ELMo		tBERT	200	500	100	1
		ELMo, GloVe	ELMo, GloVe		tBERT, USE	100	300	100	1
	Hierarchical				tBERT	300	600	N.A.	N.A.
		ELMo			tBERT	300	600	N.A.	N.A.
		ELMo, GloVe			tBERT	100	200	N.A.	N.A.
		ELMo, *GloVe*			tBERT	100	200	N.A.	N.A.
		ELMo, GloVe			*tBERT*	200	400	N.A.	N.A.
		concat(ELMo, GloVe)			tBERT	100	100	N.A.	N.A.
		concat(ELMo, GloVe)			*tBERT*	300	600	N.A.	N.A.
			ELMo, GloVe		tBERT	300	500	100	1
				ELMo, GloVe	tBERT	100	100	100	1
		Ling, ELMo, GloVe			tBERT	300	600	N.A.	N.A.
		ELMo, GloVe			tBERT, USE	300	600	N.A.	N.A.
		ELMo, GloVe			*tBERT*, USE	100	100	N.A.	N.A.
		ELMo, GloVe			*tBERT, USE*	100	200	N.A.	N.A.
		ELMo, GloVe	ELMo, GloVe		tBERT	300	500	100	1
		Ling, fastText, ELMo, GloVe			tBERT, USE, InferSent	200	400	N.A.	N.A.

Table 2. Hyper-parameter values corresponding to multi-label sexism classification results for the proposed ensemble approach (the hierarchical architecture with one sentence-level group is used; the number of learners equals *max_label_subsets* where our automatic label subset selection method is used)

***biL-att* applied on**	**Text Encoder**	**Ensemble settings**	**LSTM dim.**	**Attention dim.**	**#CNN filters of each kernel size**	**k in k-max-over-time pooling**
ELMo, GloVe	tBERT	*max_label_subsets* = 2, *label_membership* = 1	200	400	N.A.	N.A.
ELMo, GloVe, fastText, Ling	tBERT, USE, InferSent	*max_label_subsets* = 2, *label_membership* = 1	100	100	N.A.	N.A.
ELMo, GloVe	tBERT	*max_label_subsets* = 3, *label_membership* = 1	200	400	N.A.	N.A.
ELMo, GloVe	tBERT	*max_label_subsets* = 3, *label_membership* = 2	200	400	N.A.	N.A.
ELMo, GloVe	tBERT	8 *omnipresent labels*, *max_label_subsets* = 2, *label_membership* = 1	200	400	N.A.	N.A.
ELMo, GloVe, fastText, Ling	tBERT, USE, InferSent	8 *omnipresent labels*, *max_label_subsets* = 2, *label_membership* = 1	200	400	N.A.	N.A.
ELMo, GloVe	tBERT	8 *omnipresent labels*, *max_label_subsets* = 3, *label_membership* = 1	200	500	N.A.	N.A.
ELMo, GloVe	tBERT	8 *omnipresent labels*, *max_label_subsets* = 3, *label_membership* = 2	100	200	N.A.	N.A.
ELMo, GloVe	tBERT	all *omnipresent labels*, 2 learners	300	500	N.A.	N.A.
ELMo, GloVe, fastText, Ling	tBERT, USE, InferSent	all *omnipresent labels*, 2 learners	200	300	N.A.	N.A.
ELMo, GloVe	tBERT	all *omnipresent labels*, 3 learners	300	600	N.A.	N.A.

Table 3. Hyper-parameter values corresponding to multi-label sexism classification results with the autoencoder-based method for using unlabeled data (using the hierarchical approach with *biL-att* on ELMo and GloVe and BERT for embedding sentences)

Encoder Settings	**Layer Sizes**	**LSTM dim.**	**Attention dim.**	**#CNN filters of each kernel size**	**k in k-max-over-time pooling**
5 stacked layers, quickly decreasing layer size	922, 819, 717, 614, 512	300	500	N.A.	N.A.
3 stacked layers, quickly decreasing layer size	922, 819, 717	200	400	N.A.	N.A.
5 stacked layers, slowly decreasing layer size	973, 922, 870, 819, 768	300	400	N.A.	N.A.
3 stacked layers, slowly decreasing layer size	973, 922, 870	100	200	N.A.	N.A.

Table 4. Hyper-parameter values corresponding to multi-label sexism classification results under various settings

Multi-Label Setting	**Approach**	**LSTM dim.**	**Attention dim.**	**#CNN filters of each kernel size**	**k in k-max-over-time pooling**
Label Powerset with class imbalance correction	biLSTM-Attention	300	600	N.A.	N.A.
	Hierarchical-biLSTM-Attention	200	400	N.A.	N.A.
	Best proposed method	100	100	N.A.	N.A.
Label Powerset without class imbalance correction	biLSTM-Attention	200	400	N.A.	N.A.
	Hierarchical-biLSTM-Attention	300	600	N.A.	N.A.
	Best proposed method	200	300	N.A.	N.A.
Binary Relevance	biLSTM-Attention	100	300	N.A.	N.A.
	Hierarchical-biLSTM-Attention	100	300	N.A.	N.A.
	Best proposed method	100	300	N.A.	N.A.
NCE loss	biLSTM-Attention	200	500	N.A.	N.A.
	Hierarchical-biLSTM-Attention	200	400	N.A.	N.A.
	Best proposed method	100	200	N.A.	N.A.

Table 5. Hyper-parameter values corresponding to results for misogyny detection (baselines use ELMo embeddings; proposed methods use the flat architecture)

	Approach				**LSTM dim.**	**Attention dim.**	**#CNN filters of each kernel size**	**k in k-max-over-time pooling**
Baselines	CNN-Kim				N.A.	N.A.	100	N.A.
	biLSTM				200	N.A.	N.A.	N.A.
	biLSTM-Attention				200	400	N.A.	N.A.
	C-biLSTM				100	N.A.	100	N.A.
	BERT				N.A.	N.A.	N.A.	N.A.
	USE				N.A.	N.A.	N.A.	N.A.
	InferSent				N.A.	N.A.	N.A.	N.A.
Proposed methods	***biL-att*** **applied on**	***c-k-max*** **applied on**	***c-biL-att*** **applied on**	**Text Encoder**				
				tBERT	N.A.	N.A.	N.A.	N.A.
	ELMo			BERT	100	200	N.A.	N.A.
	ELMo			tBERT	100	300	N.A.	N.A.
	ELMo			USE	200	400	N.A.	N.A.
		ELMo		BERT	N.A.	N.A.	150	4
	ELMo, GloVe			BERT	100	300	N.A.	N.A.
	ELMo, GloVe			BERT, USE	300	600	N.A.	N.A.
	concat(ELMo, GloVe)			BERT, USE	100	300	N.A.	N.A.
	ElMo, GloVe, Ling			BERT	100	100	N.A.	N.A.
	ElMo, GloVe, Ling			BERT, USE	100	300	N.A.	N.A.
			ElMo, GloVe, fastText, Ling		100	200	100	N.A.
	ElMo, GloVe, fastText, Ling			BERT, USE, InferSent	100	200	N.A.	N.A.
		ElMo, GloVe, fastText, Ling		BERT, USE, InferSent	N.A.	N.A.	100	1
	concat(ELMo, GloVe, fastText, Ling)			BERT, USE, InferSent	200	400	N.A.	N.A.
	ElMo, GloVe, fastText, Ling	ElMo, GloVe, fastText, Ling		BERT, USE, InferSent	300	600	100	1

Table 6. Hyper-parameter values corresponding to results for misogyny classification (baselines use ELMo embeddings; proposed methods use the flat architecture)

	Approach				**LSTM dim.**	**Attention dim.**	**#CNN filters of each kernel size**	**k in k-max-over-time pooling**
Baselines	CNN-Kim				N.A.	N.A.	100	N.A.
	biLSTM				300	N.A.	N.A.	N.A.
	biLSTM-Attention				100	300	N.A.	N.A.
	C-biLSTM				300	N.A.	100	N.A.
	BERT				N.A.	N.A.	N.A.	N.A.
	USE				N.A.	N.A.	N.A.	N.A.
	InferSent				N.A.	N.A.	N.A.	N.A.
Proposed methods	***biL-att* applied on**	***c-k-max* applied on**	***c-biL-att* applied on**	**Text Encoder**				
				tBERT	N.A.	N.A.	N.A.	N.A.
	ELMo			tBERT	200	500	N.A.	N.A.
		ELMo		tBERT	N.A.	N.A.	150	1
			ELMo	tBERT	100	300	100	1
	ELMo			USE	100	200	N.A.	N.A.
	concat(ELMo, GloVe)			tBERT	100	100	N.A.	N.A.
	concat(ELMo, GloVe)			USE	100	200	N.A.	N.A.
		ELMo, GloVe		tBERT	N.A.	N.A.	100	1
	ELMo, GloVe			USE	100	100	N.A.	N.A.
	ELMo, GloVe			tBERT	100	100	N.A.	N.A.
			ELMo, GloVe	tBERT	100	300	100	2
	ELMo, GloVe, Ling			tBERT	100	200	N.A.	N.A.
	concat(ELMo, GloVe, Ling)			tBERT	100	200	N.A.	N.A.
	ELMo, GloVe			tBERT, USE	200	400	N.A.	N.A.
	concat(ELMo, GloVe)			tBERT, USE	100	200	N.A.	N.A.
	ELMo	ELMo		tBERT	200	400	100	1
	ELMo, GloVe	ELMo, GloVe		tBERT	200	500	200	1
	ELMo, GloVe, fastText, Ling			tBERT, USE, InferSent	100	200	N.A.	N.A.

Table 7. Hyper-parameter values corresponding to multi-label sexism classification results with the EBCE loss without weight-based class imbalance correction for deep learning methods (the sub-columns for the proposed methods specify different configurations as to the ways in which the input is embedded and/or processed; different colors/styles represent different sentence-level concatenations in a method; colors/styles have no connections across rows)

Proposed Hierarchical Architecture				**LSTM dim.**	**Attention dim.**	**#CNN filters of each kernel size**	**k in k-max-over-time pooling**
***biL-att* applied on**	***c-k-max* applied on**	***c-biL-att* applied on**	**Text Encoder**				
Ling, fastText, ELMo, GloVe			tBERT, USE, InferSent	100	100	N.A.	N.A.