

For each method, all the hyper-parameter tuning is done using the validation data. The hyper-parameters values for the baselines and proposed methods are as follows.

Traditional machine learning hyper-parameters:

- For SVM, soft margin (C) is set to 1.0 and class imbalance correction is applied.
- For Random Forest, the number of estimators is 100.
- For logistic regression, class imbalance correction is applied.
- The maximum number of features used are 10000, and character word-ngram range is (1,2).

Fixed parameters across all deep learning (DL), semi supervised baselines and for proposed methods:

- The learning rate was set to 0.001.
- Number of epochs set to 10.
- Batchsize set to 64.
- We employ 0.25 dropouts after each input as well as before the final fully connected layer.

Tuned hyper-parameter values for DL baselines

Approach	LSTM dim.	Attention dim.	CNN Filters
biLSTM	300	N.A.	N.A.
biLSTM-Attention	300	600	N.A.
Hierarchical-biLSTM-Attention	200	400	N.A.
BERT-biLSTM-Attention	300	600	N.A.
USE-biLSTM-Attention	300	600	N.A.
CNN-biLSTM-Attention	300	500	100
CNN-Kim	0	0	150
C-biLSTM	300	0	N.A.

Tuned hyper-parameter values for semi supervised baselines

Approach	Threshold (T)	LSTM dim.	Attention dim.
BERT-t-biLSTM-Attention	N.A.	300	600
Self-training	0.8	200	300
(Parikh et al., 2019)	N.A.	200	400

Tuned hyper-parameter values for proposed multi-task methods

Approach	Topic-p loss weight	Cl-pred loss weight	S-det loss weight	LSTM dim.	Attention dim.
Topic-p	0.3	N.A.	N.A.	200	400
Cl-pred	N.A.	0.1	N.A.	100	100
Topic-p, Cl-pred	0.3	0.1	N.A.	200	300
S-det	N.A.	N.A.	0.1	200	400
S-det, Cl-pred	N.A.	0.1	0.3	300	600
S-det, Topic-p, Cl-pred	0.1	0.1	0.1	200	300
S-det, Topic-p	0.3	N.A.	0.1	300	600

The loss weight for sexism classification is set to 1.0. For k-means, the tune hyper parameter value for the number of clusters k is 7.

Tuned hyper-parameter values for proposed objective functions

Approach	Threshold (t)	Beta	LSTM dim.	Attention dim.
L-unc	N.A.	0.01	200	300
L-cor with t=0	0.0	0.01	200	400
L-cor	0.05	0.05	100	200
L-cor with t=1	1.0	0.05	200	400

Here, Beta is the loss weight with which each proposed label co-occurrence based loss is multiplied (each Beta-scaled proposed loss is added to extended binary cross entropy to form the corresponding overall loss).

Tuned hyper-parameter values for our best method

Approach	Threshold	Beta	Topic-p loss weight	S-det loss weight	LSTM dim	Attention dim
S-det, Topic-p, L-cor	0.05	0.05	0.3	0.1	300	600