Semi-supervised Multi-task Learning for Multi-label Fine-grained Sexism Classification

Harika Abburi* IIIT-Hyderabad, India harika.a@research.iiit.ac.in

Pulkit Parikh* IIIT-Hyderabad, India pulkit.parikh@research.iiit.ac.in

Niyati Chhaya Adobe Research, India nchhaya@adobe.com

Vasudeva Varma IIIT-Hyderabad, India vv@iiit.ac.in

Abstract

Sexism, a form of oppression based on one's sex, manifests itself in numerous ways and causes enormous suffering. In view of the growing number of experiences of sexism reported online, categorizing these recollections automatically can assist the fight against sexism, as it can facilitate effective analyses by gender studies researchers and government officials involved in policy making. In this paper, we investigate the fine-grained, multi-label classification of accounts (reports) of sexism. To the best of our knowledge, we work with considerably more categories of sexism than any published work through our 23-class problem formulation. Moreover, we propose a multi-task approach for fine-grained multi-label sexism classification that leverages several supporting tasks without incurring any manual labeling cost. Unlabeled accounts of sexism are utilized through unsupervised learning to help construct our multi-task setup. We also devise objective functions that exploit label correlations in the training data explicitly. Multiple proposed methods outperform the state-of-the-art for multi-label sexism classification on a recently released dataset across five standard metrics.

1 Introduction

Sexism, defined as prejudice, stereotyping, or discrimination based on a person's sex, occurs in various overt and subtle forms, permeating personal as well as professional spaces. While men and boys are also harmed by sexism, women and girls suffer the brunt of sexist mindsets and resultant wrongdoings. With increasingly many people sharing recollections of sexism experienced or witnessed by them, the automatic classification of these accounts into well-conceived categories of sexism can help fight this oppression, as it can better equip authorities formulating policies and researchers of gender studies to analyze sexism.

The detection of sexism differs from and can complement the classification of sexism. In a forum where instances of sexism are mixed with other posts unrelated to sexism, sexism detection can be used to identify the posts on which to perform sexism classification. Moreover, we observe the distinction between sexist statements (e.g., posts whereby one perpetrates sexism) and the accounts of sexism suffered or witnessed (e.g., personal recollections shared as part of the #metoo movement). We also note the prior work on detecting or classifying personal stories of sexual harassment and/or assault (Chowdhury et al., 2019; Karlekar and Bansal, 2018). In this paper, we focus on classifying an account (report) of sexism involving any set of categories of sexism.

Most of the existing research on sexism classification (Anzovino et al., 2018; Jafarpour et al., 2018; Jha and Mamidi, 2017) considers at most five categories of sexism. Further, the majority of prior approaches associate only one category of sexism with an instance of sexism. Having mutually exclusive categories of sexism is unreasonable and limiting, as substantiated by Table 1.

To the best of our knowledge, Parikh et al. (2019) is the only work that explores the multi-label categorization of accounts involving any type(s) of sexism. It provides the largest dataset containing

*Both authors contributed equally.

Proceedings of the 28th International Conference on Computational Linguistics, pages 5810–5820
Barcelona, Spain (Online), December 8-13, 2020

Table 1: An Instance of sexism associated with multiple categories

Account	"A colleague once saw me washing my coffee mug before leaving the office and 'joked' if I was practicing for my 'home duties'. It's sad that he doesn't see the problem with men not bearing half the load of household work."
Associated categories of sexism	Role stereotyping: False generalizations about some roles being more suitable for women; also applies to similar mistaken notions about men
	Moral policing: The promotion of discriminatory guidelines for women under the pretense of morality; also applies to statements that feed into such narratives
	Hostile work environment: Sexism suffered at the workplace; also applies when sexism perpetrated by a colleague elsewhere makes working worrisome for the victim

accounts drawn from 'Everyday Sexism Project'[1]. It contains about 13K textual accounts labeled with at least one of 23 categories of sexism formulated with the help of a social scientist. However, they perform sexism classification among 14 categories derived by merging some sets of categories. This prohibits distinguishing within category pairs such as {moral policing, victim blaming} and {motherhood-related discrimination, menstruation-related discrimination}. We overcome this limitation by carrying out a fine-grained (23-category) classification using the same labeled dataset. Table 1 defines a subset of the 23 categories; the entire set is listed in Figure 3 and defined in Parikh et al. (2019).

Given the limited labeled data and large number of categories for our sexism classification, we explore complementary signals for the learning. As far as we know, this paper presents the first multi-task approach for any type of sexism classification. We propose a semi-supervised multi-task multi-label classification approach involving (up to) three tasks. All three tasks are set up automatically, requiring no manual labeling effort; the labels and/or samples needed are created through unsupervised learning or acquired via weak labeling. We obtain unlabeled accounts of sexism from 'Everyday Sexism Project' for unsupervised topic proportion distribution estimation and clustering. We develop a neural multi-task architecture that allows for shared learning across multiple tasks through common layers/weights and a combined loss function.

In a multi-label setting, the presence of one label may affect the likelihood of the presence of another. While the label co-occurrences in the training data are indirectly accessible to the standard classifier training (since the entire training data is used during the course of an epoch), we propose principled ways of utilizing them explicitly. We formulate loss functions for multi-label classification relating to pair-wise label correlations. We compute pair-wise label co-occurrence statistics from the training data and use them as targets for the proposed objective functions. Through our loss functions, we seek to bring together the actual conditional co-occurrence probabilities and the corresponding model-based probability estimates. In addition to a generic loss function, we present a loss focusing on only non-co-occurring label pairs. We combine a proposed semi-supervised multi-task neural model with our generic loss function to form a method that outperforms numerous baselines with a clear margin. Our key contributions are summed up below.

- To the best of our knowledge, this is the only work to consider as many as 23 categories for sexism classification.
- We introduce a semi-supervised multi-task neural approach for sexism classification. Three appropriate auxiliary tasks are prepared automatically through unsupervised learning and weak labeling.
- We propose loss functions aimed at tapping label correlations in the multi-label data explicitly.
- Our best-performing multi-task method and loss function yield better results than many, diverse baselines across five different metrics individually. Combining them enhances the performance further.

[1] https://everydaysexism.com

2 Related Work

In this section, we describe prior work on the classification of sexism. Though our work involves accounts of sexism, existing work on the classification of sexist or misogynous statements (e.g., tweets wherein one perpetrates sexism or misogyny) and some distantly related work are also included in this review. We end this section with a brief review of multi-task learning.

Melville et al. (2018) apply topic modeling to data obtained from The Everyday Sexism Project and maps the semantic relations between topics. ElSherief et al. (2017) study user engagement with posts related to gender based violence and their language nuances. Since sexism classification can be preceded by sexism detection to remove posts unrelated to sexism, we note that sexism is detected by some hate speech classification methods that include sexism as a category of hate (Badjatiya et al., 2017; Waseem and Hovy, 2016; Zhang and Luo, 2018; Davidson et al., 2017). Frenda et al. (2019) present an approach for detecting sexism and misogyny from tweets. Given our focus on sexism classification, we do not delve into prior work related to hate speech or cyber-bullying (Van Hee et al., 2015; Agrawal and Awekar, 2018).

Karlekar and Bansal (2018) explore CNN, RNN, and a combination of them for categorizing personal experiences of sexual harassment into one or more of three classes. In Yan et al. (2019), a density matrix encoder inspired by quantum mechanics is used for the classification of personal stories of sexual harassment. Khatua et al. (2018) employ deep learning methods to classify sexual violence into one of four categories. In Anzovino et al. (2018), tweets identified as misogynist are classified as stereotype and objectification, discredit, sexual harassment and threats of violence, dominance, or derailing using features involving Part of Speech (POS) tags, n-grams, and text embedding. Jafarpour et al. (2018) perform a 4-class categorization of sexist tweets. In Jha and Mamidi (2017), tweets are classified as benevolent, hostile, or non-sexist using biLSTM with attention, SVM, and fastText. While its categorization of sexism pertains to how it is stated, our work concentrates on aspects such as what an instance of sexism involves, where it occurs, and who perpetrates it.

Parikh et al. (2019) explore multi-label categorization of accounts reporting any kind(s) of sexism. They provide the largest dataset for sexism classification and the state-of-the-art classifier for it. Their neural approach can combine sentence embeddings from pre-trained models like BERT with those generated using biLSTM with attention and CNN. As far as we know, our work presents the first semi-supervised approach for the multi-label classification of accounts describing any type(s) of sexism that goes further than using unlabeled instances only for fine-tuning pre-trained models.

Multi-Task Learning (MTL) is inspired by human learning activities wherein people often apply the knowledge learned from previous tasks to help learn a new task (Zhang and Yang, 2017). It is useful for multiple (related) tasks to be learned jointly so that the knowledge learned in one task can benefit other tasks. There has been considerable interest in applying MTL to a variety of problems including text classification using deep neural networks (DNNs) (Liu et al., 2019; Xu et al., 2019; Guo et al., 2018; Ruder et al., 2019). MTL provides an effective way of leveraging labeled data from auxiliary tasks for the core task, especially when labeled data available for single-task learning is not large. In this work, we adopt MTL for fine-grained multi-label sexism classification using several auxiliary tasks.

3 Proposed Semi-supervised Multi-task Approach for Sexism Classification

Our problem statement is to classify an account of sexism (also referred to as a post henceforth) into one or more of 23 categories of sexism. This section introduces a semi-supervised multi-task approach for it. We begin with the auxiliary task setup and then provide the architecture details. Henceforth, we refer to the labeled training set as L and the unlabeled set as U (w.r.t. categories of sexism).

3.1 Formulating Auxiliary Tasks

We construct three auxiliary tasks that 1) could complement sexism classification in learning terms and 2) involve unlabeled accounts of sexism obtained from 'Everyday Sexism Project' that substantially outnumber the labeled instances available. The labels for all these tasks and additional samples for one are obtained through unsupervised learning or weak labeling, as depicted in Fig. 1 and detailed below.

1. **Estimating the Topic Proportion Distribution**: The *Topic-p* task is the prediction of topic proportion distributions, signifying the degrees to which a given post relates to a fixed number of topics. For estimating these distributions (which act as target labels in our subsequent multi-task learning) and topics, we employ lda2vec (Moody, 2016). It produces post-to-topic proportions by mixing word2vec's skip-gram architecture with Dirichlet-optimized sparse topic mixtures. We fix the number of topics to 10. While we train lda2vec on $L \cup U$, we use only the topic proportion distributions associated with L for multi-task learning. Experiments that substitute them with their U-based counterparts underperform.
2. **Predicting the Cluster Label**: Another auxiliary task that we explore is the prediction of the cluster to which a given post belongs (*Cl-pred*). We perform k-means clustering on vector representations of the posts in $L \cup U$ and thereby augment L with cluster labels for setting up this k-class classification. The post representations are created using a BERT (Devlin et al., 2018) model tuned using unlabeled accounts of sexism (henceforth called *BERT-t*). The number of clusters k is a hyper-parameter that we tune.
3. **Detecting an Account of Sexism**: Identifying whether a given post is an account of sexism is adopted as a third task (*S-det*). We obtain the (weakly labeled) negative data for this task from the Blog Authorship Corpus (Schler et al., 2006) through two types of filtering. To help select accounts as opposed to commentary, the presence of a few keywords and one past tense POS tag is mandated. Moreover, we exclude posts that exceed the word and sentence count maximums of L. This filtering is also used while randomly selecting posts from U as the positive data. L itself can also be (additionally) used for this purpose.

3.2 Proposed Architecture

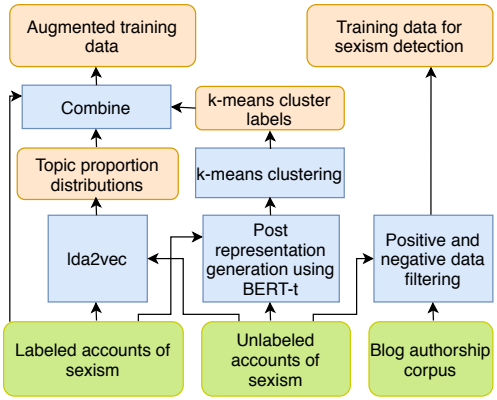

Figure 1: Our unsupervised data creation for multi-task sexism classification

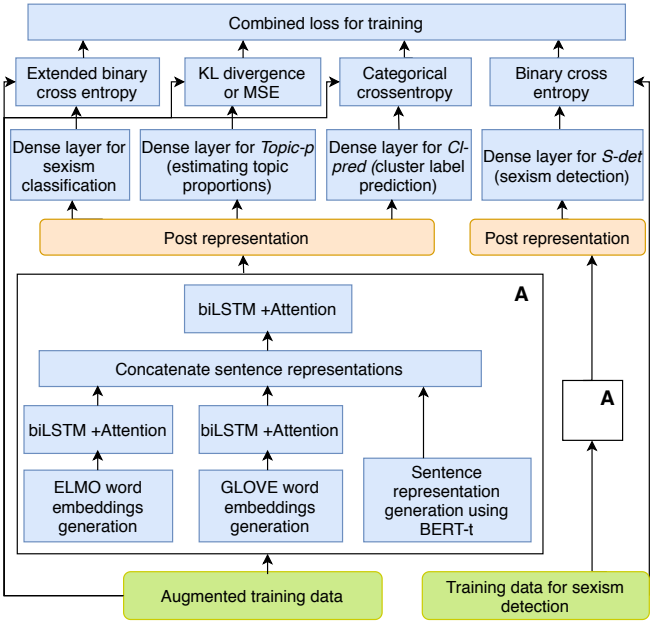

Figure 2: Proposed multi-task neural classification

Fig. 2 presents our multi-task neural architecture. If the list of chosen tasks consists of sexism detection, the training input is a batch of tuples (samples), each of which consists of a post from L and $|L|$ posts randomly picked from the sexism detection training data each. Each of the two posts in a sample is processed hierarchically for the creation of its post representation similar to Parikh et al. (2019) using the same set of layers/weights (block A in Fig. 2). First, the words of each sentence (in each post) are embedded separately using ELMo (Peters et al., 2018) and GloVe (Pennington et al., 2014). The two word vector matrices are passed through biLSTM linked with an attention scheme (Yang et al., 2016), yielding two sentence representations. This is complemented by another sentence embedding generated using *BERT-t*. Next, the concatenation of the three sentence vectors is fed to biLSTM + attention to

create the post representation.

The vector representation for the post from L is passed to a fully connected layer custom-made for each of sexism classification, *Topic-p*, and *Cl-pred*. The other post representation is fed to a dense layer designed for *S-det*. For sexism classification and *S-det*, the sigmoid activation is used; for the other two tasks, we use softmax. For sexism classification, we employ the extended binary cross entropy loss addressing the multi-label aspect (Parikh et al., 2019). For the loss for *Topic-p*, we explore KL divergence (as lda2vec outputs a probability distribution) but find mean squared error (MSE) more effective. For *Cl-pred* and *S-det*, we use categorical and binary cross entropy losses respectively. We incorporate class imbalance neutralizing weights into the loss functions where needed. The final loss that we train with is a weighted combination of these losses, where the weights are hyper-parameters.

We experiment with variants of this architecture corresponding to some subsets of the three auxiliary tasks described. We also attempt transfer learning wherein a variant of the proposed model comprising only one auxiliary task's dense layer is used to pre-train the weights of block A used in another variant meant for the core sexism classification task. Since this does not perform well, it is omitted.

4 Our Label Co-occurrence based Loss Functions for Multi-label Classification

In this section, we propose loss functions which explicitly use label correlations existing in the multi-label data. In order to take advantage of label correlations this way, we compute statistics pertaining to pair-wise label co-occurrences from the training set L and supply them as targets to the loss functions. First, we compute the symmetric label co-occurrence (count) matrix M. We then estimate the probability of label l_j occurring given the occurrence of label l_i, $P(j|i)$, as $M(i,j)/f(i)$, where $f(i)$ denotes the frequency of label l_i. Next, we compute the set of label pairs correlated to a specified degree; $S = \{i, j \mid P(j|i) \leq t,\ i \neq j,\ 1 \leq i, j \leq q\}$, where the threshold t is a hyper-parameter. The proposed loss function *L-cor* is then given as follows.

$$\textit{L-cor} = \frac{1}{|S|} \sum_{(i,j) \in S} |\frac{\sum_{k=1}^{n} \hat{p}_{ki}\hat{p}_{kj}}{\sum_{k=1}^{n} \hat{p}_{ki}} - P(j|i)| \tag{1}$$

Here, n and q are the numbers of posts and labels respectively. $\hat{p}_{ki}$ is the estimated probability of label l_i applying to post x_k. *L-cor* is aimed at minimizing the L1 norms of the differences between the actual conditional co-occurrence probabilities $P(j|i)$ and their estimated model-based counterparts $\hat{P}(j|i)$ for the label pairs in S. The $\hat{P}(j|i)$ estimates are derived from the probabilities $\hat{p}_{ki}$ output by the model. To include all non-diagonal labels pairs in the loss computation, t is set to 1. Conversely, to impose penalties for only non-co-occurring (non-diagonal) label pairs, we set t to 0 and can remove $||$ from Eq. 1 (as $\forall (i,j) \in S,\ P(j|i) = 0$ when $t = 0$).

We also devise another loss function targeting non-co-occurring label pairs. We first compute the set of uncorrelated pairs (as per the training data); $S_u = \{i, j \mid M(i,j) = 0,\ i < j,\ 1 \leq i, j \leq q\}$. Using that, our loss function *L-unc* is computed using the following expression.

$$\textit{L-unc} = \frac{1}{|S_u|} \sum_{(i,j) \in S_u} \frac{(\sum_{k=1}^{n} \hat{p}_{ki}\hat{p}_{kj})(\sum_{k=1}^{n} \hat{p}_{ki} + \sum_{k=1}^{n} \hat{p}_{kj})}{(\sum_{k=1}^{n} \hat{p}_{ki})(\sum_{k=1}^{n} \hat{p}_{kj})} \tag{2}$$

In effect, for each $(i,j) \in S_u$, we sum the model-based conditional co-occurrence probabilities $\hat{P}(j|i)$ and $\hat{P}(i|j)$ here. Since S_u comprises only non-co-occurring label pairs, these co-occurrence probability estimates should be low for the optimal model weights. Hence, we minimize their sum.

We also explore two variants of the *L-cor* and *L-unc* losses. Replacing the L1 norm with the L2 norm in Eq. 1 constitutes one of them. The other involves replacing the sum of the model-based conditional co-occurrence probabilities in Eq. 2 with a model-based co-occurrence score. Both underperform *L-cor* and *L-unc*.

Each proposed label co-occurrence based loss is weighed by a hyper-parameter and added to extended binary cross entropy. We use some of the combined loss functions in conjunction with the existing state-of-the-art model (Parikh et al., 2019) as well as our multi-task model.

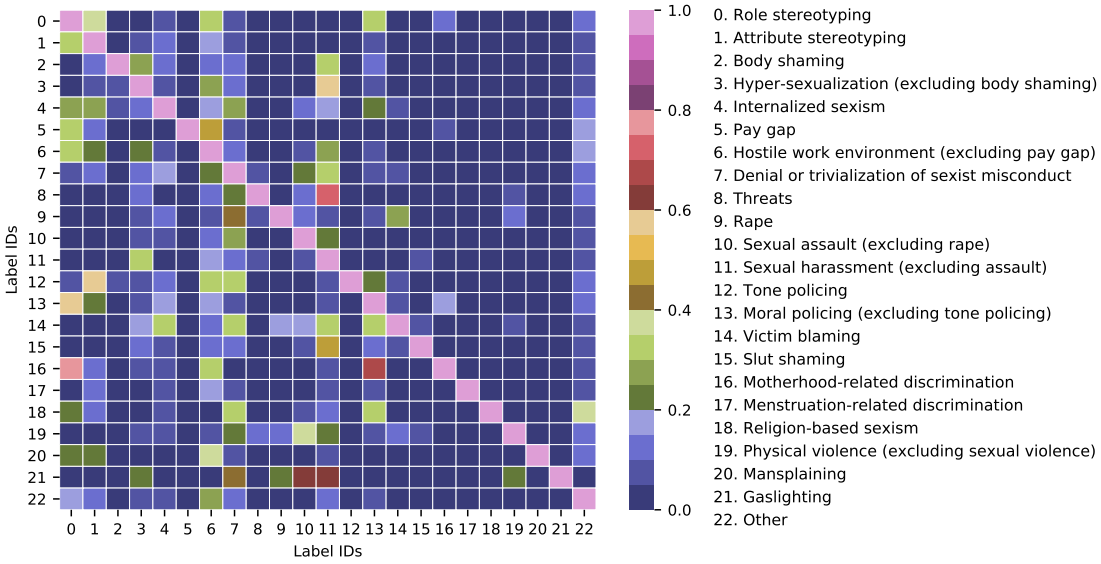

Figure 3: Pairwise conditional label co-occurrence matrix for the training data

5 Experiments

This section provides the experimental evaluation of the proposed sexism classification methods against a number of baseline methods and presents analysis. Our code, all hyper-parameter values used are available on GitHub[2]

5.1 Datasets

We use the dataset contributed by Parikh et al. (2019) comprising $13,023$ accounts of sexism, each labeled with at least one of 23 categories of sexism. The diverse categories of sexism, derived in consultation with a social scientist, range from body shaming and menstruation-related discrimination to role stereotyping and victim blaming. Figure 3 lists the 23 categories and shows the pairwise conditional label co-occurrence matrix for the training data L. We caution that coverage/co-occurrence statistics related to sexism shown in this work do not represent their real-world counterparts.

We obtain unlabeled instances of sexism from 'Everyday Sexism Project', which has already received several hundred thousand accounts of sexism from survivors and observers. We shortlist $70,000$ shortest instances containing a minimum of 7 words each to form U. Short posts are preferred to maximize the resemblance to the labeled data. Additionally, we use Blog Authorship Corpus (Schler et al., 2006) to obtain weakly labeled negative data for the sexism detection auxiliary task; the keywords mandated therein to help select accounts as opposed to commentary are 'i','we', and 'he'. The negative and positive sets each comprise $20,000$ posts.

5.2 Evaluation Metrics

Multi-label classification, wherein classes can co-occur, is evaluated differently from the single-label case. We adopt a number of established metrics, namely instance-based F1 referred to as F_{ins}, instance-based accuracy Acc, F1 macro F_{mac}, F1 micro F_{mic}, and Subset Accuracy SA (Zhang and Zhou, 2014; Parikh et al., 2019).

5.3 Baselines

All deep learning architectures below end with a dense layer with the sigmoid activation and are trained using the extended binary cross entropy loss.

- **Random**: For each test sample in training data, labels are selected randomly based on their normalised frequencies.

[2] https://github.com/Harikavuppala1a/Semisupervised_Multitask_Learning

- **Traditional Machine Learning (TML)**: We report the performance using Support Vector Machine (SVM), Logistic Regression (LR), and Random Forests (RF), each applied on two feature sets, namely the average of the ELMo vectors for a post's words (referred to as ELMO) and TF-IDF on word unigrams and bigrams (called Word-ngrams). This gives rise to six combinations: ELMO with SVM (ELMO-SVM), ELMO with LR (ELMO-LR), ELMO with RF (ELMO-RF), word-ngrams with SVM (word-ngrams-SVM), word-ngrams with LR (word-ngrams-LR), and word-ngrams with RF (word-ngrams-RF).
- **Deep Learning (DL)**:
  biLSTM and biLSTM-Attention: The word embeddings for a post are passed through a bidirectional LSTM with and without the attention scheme from Yang et al. (2016).
  Hierarchical-biLSTM-Attention: In an architecture similar to Yang et al. (2016) with GRU replaced with LSTM, the word embeddings are first fed to biLSTM with attention to create a representation for each sentence. These sentence embeddings are then passed through another instance of biLSTM with attention.
  BERT-biLSTM-Attention and USE-biLSTM-Attention: Sentence representations are generated using BERT via bert-as-service (Xiao, 2018) and USE (Cer et al., 2018) each and fed to a biLSTM with attention.
  CNN-Kim: Convolutional and max-over-time pooling layers are applied to the word vectors for a post in this method similar to Kim (2014).

Table 2: Results for the proposed methods as well as the traditional machine learning (TML), deep learning (DL), and semi-supervised baselines

	Approach	$\mathbf{F_{ins}}$	$\mathbf{F_{mac}}$	$\mathbf{F_{mic}}$	**Acc**	**SA**
	Random	0.035	0.090	0.192	0.022	0.003
TML baselines	Word-ngrams-SVM	0.453	0.227	0.413	0.331	0.107
	Word-ngrams-LR	0.544	0.188	0.492	0.454	0.287
	Word-ngrams-RF	0.538	0.246	0.482	0.444	0.272
	ELMO-SVM	0.546	0.261	0.501	0.431	0.206
	ELMO-LR	0.576	0.261	0.535	0.475	0.279
	ELMO-RF	0.374	0.100	0.330	0.307	0.185
DL baselines	biLSTM	0.627	0.451	0.577	0.472	0.147
	biLSTM-Attention	0.648	0.445	0.597	0.499	0.176
	Hierarchical-biLSTM-Attention	0.664	0.485	0.616	0.516	0.191
	BERT-biLSTM-Attention	0.591	0.397	0.546	0.431	0.089
	USE-biLSTM-Attention	0.566	0.398	0.525	0.402	0.061
	CNN-Kim	0.658	0.481	0.617	0.513	0.195
	CNN-biLSTM-Attention	0.421	0.284	0.387	0.278	0.035
	C-biLSTM	0.485	0.316	0.446	0.328	0.038
Semi-supervised baselines	*BERT-t*-biLSTM-Attention	0.632	0.435	0.580	0.472	0.127
	Self-training	0.704	0.513	0.654	0.557	0.220
	(Parikh et al., 2019)	0.714	0.546	0.665	0.572	0.242
Proposed multi-task methods	**Auxiliary tasks**					
	Topic-p	0.716	0.560	0.669	0.574	0.241
	Cl-pred	0.716	0.558	0.668	0.577	0.251
	Topic-p, Cl-pred	0.720	0.552	0.673	0.581	0.256
	S-det	0.726	0.559	0.679	0.589	0.273
	S-det, Cl-pred	0.724	0.558	0.679	0.589	0.275
	S-det, Topic-p, Cl-pred	0.720	0.550	0.673	0.583	0.266
	S-det, Topic-p	0.728	0.565	0.677	0.590	0.276
Proposed objective functions with (Parikh et al., 2019)	*L-unc*	0.716	0.546	0.668	0.575	0.242
	L-cor with $t = 0$	0.711	0.550	0.663	0.570	0.247
	L-cor	0.723	0.550	0.672	0.584	0.264
	L-cor with $t = 1$	0.718	0.559	0.670	0.577	0.245
Our best method	*S-det, Topic-p with L-cor*	**0.731**	**0.573**	**0.681**	**0.595**	**0.281**

CNN-biLSTM-Attention: In this baseline similar to Wang et al. (2016), each sentence's word embeddings are passed through convolutional and max-over-time pooling layers. The resultant representations are then passed through a biLSTM with attention.
C-biLSTM: This is a variant of the C-LSTM architecture (Zhou et al., 2015) somewhat related to a method used in Karlekar and Bansal (2018). After applying convolution on a post's word vectors, the feature maps are stacked along the filter dimension to generate a series of window vectors, which are subsequently fed to biLSTM.

- **Semi-supervised**
  BERT-t-biLSTM-Attention: This is the same as BERT-biLSTM-Attention except that the pre-trained BERT model used is fine-tuned using unlabeled instances of sexism (Parikh et al., 2019).
  (Parikh et al., 2019): The state-of-the-art model concatenates sentence representations obtained using a *BERT-t* with those created from ELMo and GloVe embeddings separately using biLSTM with attention. The combined sentence embeddings are passed through biLSTM with attention.
  Self-training: We adapt self training (Yarowsky, 1995) for multi-label classification. Iteratively, we train (Parikh et al., 2019), use it to pseudo-label unlabeled samples, and augment the training set with those for which the mean of the class-wise model-given probabilities exceeds a threshold.

5.4 Results

Table 2 provides the results produced by various proposed multi-task methods, objective functions and the baselines. We set aside 15% from original labeled data for validation and testing each. The validation set was merged into the training set during the testing phase. For each deep learning method, for each metric, the mean of the results obtained over three runs is given.

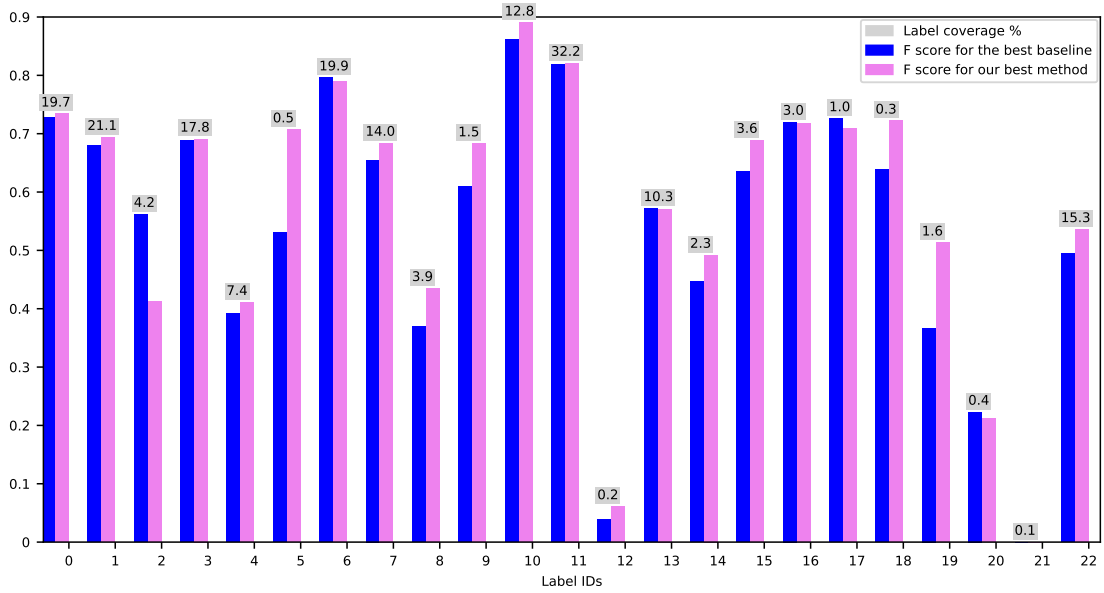

Figure 4: Class-wise sexism classification F-scores for the best-performing baseline (Parikh et al., 2019) and our best method (*S-det, Topic-p* with *L-cor*). Each grey box contains the % of the training data samples that the corresponding label applies to. The class names follow the same order as Fig. 3.

As expected, the random baseline performs extremely poorly, reflecting the challenging nature of this fine-grained multi-label classification. Among the combinations of classifiers and features experimented with for traditional machine learning, ELMo with logistic regression yields the best results. The best deep learning baseline is Hierarchical-biLSTM-Attention, and it outperforms its traditional ML counterpart. Overall, the semi-supervised method by Parikh et al. (2019) is confirmed as the best baseline.

Most proposed methods outperform all baselines across all metrics. The maximum performance improvement is observed for subset accuracy (SA), which is the most stringent metric. Our best performing multi-task method involves *S-det* and *Topic-p* as the auxiliary tasks (along with the primary sexism clas-

sification task). Among the proposed objective functions capitalizing on label correlations, used with the state-of-the-art model (Parikh et al., 2019), *L-cor* leads to the best results for most metrics. This loss seeks to minimize the L1 norms of the differences between the actual and model-based conditional co-occurrence probabilities. The best performance is seen with a combined proposed method involving auxiliary tasks *S-det* and *Topic-p* along with our *L-cor* loss.

Table 3: Test samples illustrating the improved performance of our combined method

Account of sexism	**True labels**	**Avg. label coverage**
One day my housemate was followed all the way home (15 minute walk) by a young guy who verbally abused her the entire way threatening to rape her and at one point trying to push her into a alley.	Sexual harassment (excluding assault), Threats, Physical violence (excluding sexual violence)	12.60%
Woman just told me that a job "might be more for a man" as it involves driving	Role stereotyping, Attribute stereotyping, Internalized sexism	16.11%
I was struggling at work and my manager told me that I needed to be noticed more. Wearing more make up, dying my hair and wearing heels would enable other men in the office to take note of me more. The aim was that when I walked in the office, the men would look my way.	Sexual harassment (excluding assault), Hyper-sexualization (excluding body shaming), Hostile work environment (excluding pay gap)	23.34%

Figure 4 compares the class-wise performance of our best performing model with that of the best baseline (Parikh et al., 2019). We report each label's coverage, i.e. the proportion (%) of samples belonging to that label in the training data, to bring out the imbalanced distribution. For a majority of the classes, the proposed method outperforms the baseline. We note a significant improvement over the baseline for several low-coverage classes. Table 3 provides accounts of sexism from the test set for which our combined method made the right predictions but the best baseline did not. For each account, the associated labels and the average of their coverage values are also reported. Our method can be seen to work well across different labels and coverage values.

We analyze the impact of the multi-label nature of the problem on the performance in Table 4. The performance for one run of our best method across different numbers of labels per post (1 to 6) is compared against the counterpart in Parikh et al. (2019). We observe improved performance with the proposed method across all metrics for a majority of the cases. Figure 5 shows the improvement in F_{mac} observed with increasing training data. Our best method produces greater improvement than the best baseline throughout and also yields a better result for the lowest training data % (i.e. 50).

#labels per post	**Approach**	$\mathbf{F_{ins}}$	$\mathbf{F_{mac}}$	$\mathbf{F_{mic}}$	**Acc**	**SA**
1	Best proposed method	**0.688**	**0.487**	**0.585**	**0.565**	**0.343**
	Best baseline method	0.667	0.414	0.554	0.537	0.314
2	Best proposed method	**0.727**	**0.566**	**0.678**	**0.589**	**0.247**
	Best baseline method	0.719	0.532	0.667	0.575	0.215
3	Best proposed method	**0.740**	**0.628**	**0.710**	**0.593**	0.157
	Best baseline method	0.739	0.535	0.705	0.592	**0.174**
4	Best proposed method	**0.767**	**0.649**	**0.748**	**0.626**	**0.104**
	Best baseline method	0.721	0.592	0.721	0.594	0.094
5	Best proposed method	**0.776**	0.713	**0.762**	**0.635**	**0.121**
	Best baseline method	0.740	**0.716**	0.728	0.592	0.091
6	Best proposed method	**0.836**	**0.865**	**0.827**	**0.726**	**0.168**
	Best baseline method	0.800	0.789	0.784	0.671	0.167

Table 4: Performance variation across #labels per post

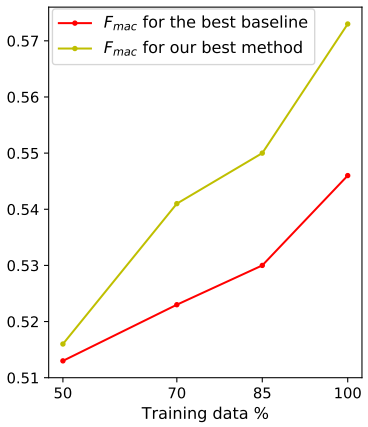

Figure 5: F_{mac} for varying training data percentages

6 Conclusion

We investigated the 23-class fine-grained classification of accounts of sexism in this work. We explored neural multi-task learning for addressing this using three auxiliary tasks automatically set up through unsupervised learning from unlabeled accounts of sexism and weak labeling. Moreover, we formulate loss functions through which we seek to utilize correlations existing between labels in the training data. Our best loss function and multi-task method outperform a number of varied baselines across five standard metrics on their own. We achieved even better results when we combined them. Next, we plan to incorporate semi-supervised proxy labeling into our approach and leverage language models.

References

Sweta Agrawal and Amit Awekar. 2018. Deep learning for detecting cyberbullying across multiple social media platforms. In *European Conference on Information Retrieval*, pages 141–153. Springer.

Maria Anzovino, Elisabetta Fersini, and Paolo Rosso. 2018. Automatic identification and classification of misogynistic language on twitter. In *International Conference on Applications of Natural Language to Information Systems*, pages 57–64. Springer.

Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. 2017. Deep learning for hate speech detection in tweets. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pages 759–760. International World Wide Web Conferences Steering Committee.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al. 2018. Universal sentence encoder. *arXiv preprint arXiv:1803.11175*.

Arijit Ghosh Chowdhury, Ramit Sawhney, Rajiv Shah, and Debanjan Mahata. 2019. # youtoo? detection of personal recollections of sexual harassment on social media. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2527–2537.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Eleventh international aaai conference on web and social media*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Mai ElSherief, Elizabeth Belding, and Dana Nguyen. 2017. # notokay: Understanding gender-based violence in social media. In *Eleventh International AAAI Conference on Web and Social Media*.

Simona Frenda, Bilal Ghanem, Manuel Montes-y Gómez, and Paolo Rosso. 2019. Online hate speech against women: Automatic identification of misogyny and sexism on twitter. *Journal of Intelligent & Fuzzy Systems*, 36(5):4743–4752.

Han Guo, Ramakanth Pasunuru, and Mohit Bansal. 2018. Soft layer-specific multi-task summarization with entailment and question generation. *CoRR*, abs/1805.11004.

Borna Jafarpour, Stan Matwin, et al. 2018. Boosting text classification performance on sexist tweets by text augmentation and text generation using a combination of knowledge graphs. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 107–114.

Akshita Jha and Radhika Mamidi. 2017. When does a compliment become sexist? analysis and classification of ambivalent sexism using twitter data. In *Proceedings of the second workshop on NLP and computational social science*, pages 7–16.

Sweta Karlekar and Mohit Bansal. 2018. Safecity: Understanding diverse forms of sexual harassment personal stories. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2805–2811.

Aparup Khatua, Erik Cambria, and Apalak Khatua. 2018. Sounds of silence breakers: Exploring sexual violence on twitter. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 397–400.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751.

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019. Multi-task deep neural networks for natural language understanding. *CoRR*, abs/1901.11504.

Sophie Melville, Kathryn Eccles, and Taha Yasseri. 2018. Topic modelling of everyday sexism project entries. *Frontiers in Digital Humanities*, 5:28.

Christopher E Moody. 2016. Mixing dirichlet topic models and word embeddings to make lda2vec. *arXiv preprint arXiv:1605.02019*.

Pulkit Parikh, Harika Abburi, Pinkesh Badjatiya, Radhika Krishnan, Niyati Chhaya, Manish Gupta, and Vasudeva Varma. 2019. Multi-label categorization of accounts of sexism using a neural framework. 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)*, pages 1642–1652.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proc. of NAACL*.

Sebastian Ruder, Joachim Bingel, Isabelle Augenstein, and Anders Søgaard. 2019. Latent multi-task architecture learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 4822–4829.

Jonathan Schler, Moshe Koppel, Shlomo Argamon, and James W Pennebaker. 2006. Effects of age and gender on blogging. In *AAAI spring symposium: Computational approaches to analyzing weblogs*, volume 6, pages 199–205.

Cynthia Van Hee, Els Lefever, Ben Verhoeven, Julie Mennes, Bart Desmet, Guy De Pauw, Walter Daelemans, and Véronique Hoste. 2015. Detection and fine-grained classification of cyberbullying events. In *Proceedings of the international conference recent advances in natural language processing*, pages 672–680.

Jin Wang, Liang-Chih Yu, K Robert Lai, and Xuejie Zhang. 2016. Dimensional sentiment analysis using a regional cnn-lstm model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 225–230.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*, pages 88–93.

Han Xiao. 2018. bert-as-service. `https://github.com/hanxiao/bert-as-service`.

Yichong Xu, Xiaodong Liu, Yelong Shen, Jingjing Liu, and Jianfeng Gao. 2019. Multi-task learning with sample re-weighting for machine reading comprehension. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2644–2655.

Peng Yan, Linjing Li, Weiyun Chen, and Daniel Zeng. 2019. Quantum-inspired density matrix encoder for sexual harassment personal stories classification. In *2019 IEEE International Conference on Intelligence and Security Informatics (ISI)*, pages 218–220. IEEE.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489.

David Yarowsky. 1995. Unsupervised word sense disambiguation rivaling supervised methods. In *33rd annual meeting of the association for computational linguistics*, pages 189–196.

Ziqi Zhang and Lei Luo. 2018. Hate speech detection: A solved problem? the challenging case of long tail on twitter. *Semantic Web*, pages 1–21.

Yu Zhang and Qiang Yang. 2017. A survey on multi-task learning. *CoRR*, abs/1707.08114.

Min-Ling Zhang and Zhi-Hua Zhou. 2014. A review on multi-label learning algorithms. *IEEE transactions on knowledge and data engineering*, 26(8):1819–1837.

Chunting Zhou, Chonglin Sun, Zhiyuan Liu, and Francis Lau. 2015. A c-lstm neural network for text classification. *arXiv preprint arXiv:1511.08630*.