2-2021

Evidence aware neural pornographic text identification for child protection

Kaisong SONG

Yangyang KANG

Wei GAO
Singapore Management University, weigao@smu.edu.sg

Zhe GAO

Changlong SUN

See next page for additional authors

Evidence Aware Neural Pornographic Text Identification for Child Protection

Kaisong Song[1], Yangyang Kang[1], Wei Gao[2], Zhe Gao[3], Changlong Sun[1], Xiaozhong Liu[4]
[1]Alibaba Group, China
[2]School of Information Systems, Singapore Management University, Singapore
[3]Ant Financial Services Group, China
[4]Indiana University Bloomington, USA
kaisong.sks@alibaba-inc.com, yangyang.kangyy@alibaba-inc.com, weigao@smu.edu.sg
gaozhe.gz@alibaba-inc.com, changlong.scl@taobao.com, liu237@indiana.edu

Abstract

Identifying pornographic text online is practically useful to protect children from access to such adult content. However, some authors may intentionally avoid using sensitive words in their pornographic texts to take advantage of the lack of human audits. Without prior knowledge guidance, real semantics of such pornographic text is difficult to understand by existing methods due to its high context-sensitivity and heavy usage of figurative language, which brings huge challenges to the porn detection systems used in social media platforms. In this paper, we approach to the problem as a document-level porn identification task by locating and integrating sentence-level evidence and propose a novel **E**vidence-Aware **N**eural **P**orn **C**lassification (**eNPC**) model. Specifically, we first propose a basic model which locates porn indicative sentences in the document with a multiple instance learning model, and then aggregate the sentence-level evidence to induce document label with self-attention mechanism. Moreover, we consider label dependencies within local context. Finally, we further enhance the sentence representation with prior knowledge produced by an automatic porn lexicon construction strategy. Extensive experimental results show that our model exhibits consistent superiority over competitors on two real-world Chinese novel datasets and an English story dataset.

Introduction

The proliferation of pornographic content online can create salient personal and social issues. Such adult content are not suitable for children and may cause juvenile delinquency. Many platforms have built their content rating systems which rate the suitability of TV broadcasts, movies, comic books, online literature or video games for their audiences[1]. For example, 9/10 boys and 6/10 girls will be exposed to pornography before they turn 18, and the majority of online exposures are unwanted and unwarranted, which may escape from existing content rating systems[2] and bring negative social impact. Pornographic text identification (and rating) is a rarely studied and important problem, which has recently drawn much attention from both research communities (Hu et al. 2007) and industries (He et al. 2020a,b).

[1]https://en.wikipedia.org/wiki/Content_rating

[2]everaccountable.com/blog/how-pornography-affects-teenagers-and-children/

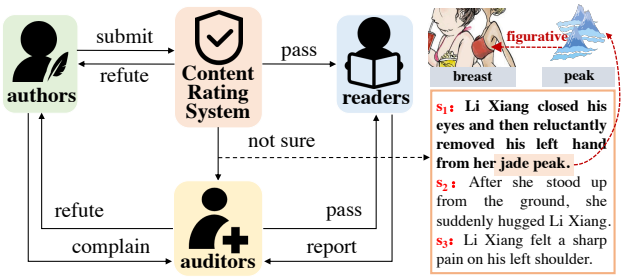

Figure 1: A typical content rating system. A pornographic example with highlighted sentence-level evidences is given.

The workflow of a typical online content rating system is displayed in Figure 1. Texts such as comments, blogs, and novels submitted to websites by authors are first checked by the content rating system which can refute plenty of obvious pornographic texts and send the suspicious texts to the auditor for confirmation. With long time struggling experience against such content rating systems, some authors may purposely avoid using sensitive words in their pornographic texts to get away with the automatic screening process. For example, the pornographic sentence s_1 in Figure 1 may be misidentified as a normal type by using figurative words "*jade peak*" instead of sensitive word "*breast*". Obviously, identifying such pornographic texts is difficult due to highly context-sensitive and figurative arousal content. In particular, the battle against such pornographic texts is regularly considered as one of the key challenges in text classification and semantic understanding of long texts.

Intuitively, pornographic text identification can be formulated as a text classification task (Lee, Hui, and Fong 2002, 2005; Hu et al. 2007). Classification models are trained on documents annotated with labels and used to predict predefined categories such as "*porn*" or "*normal*" given unseen texts. However, this kind of model based on traditional or neural text classification lacks of result interpretability, which cannot provide indicative porn evidence to back up manual reaffirmation and maybe cause bias and misuse. Rooted from the need in real-world applications, we define the pornographic text identification as a task that aims at predicting whether the given document is "*porn*" or "*nor-*

mal" by pinpointing and integrating the sentence-level evidence for classification, *with only document-level labels being available*, which is dubbed as evidence-aware identification of porn text. Document-level annotation is relatively easy to obtain due to the wide use of policy violation reporting function provided by online literature platforms. Recently, Multiple Instance Learning (MIL) is proposed to classify each segment (e.g., sentence) in a bag (e.g., document) with CNN and then aggregate the prediction results as the final prediction of the bag without any sentence-level annotations (Angelidis and Lapata 2018; Hellman et al. 2020). Despite it may be applied to our problem, the model learns representation for each sentence independently and ignores the context dependency among sentences which can provide helpful clues for identifying suspicious porn indicative sentences. Besides, it ignores prior knowledge and cannot handle highly context-sensitive and figurative arousal content which contains explicit or implicit pornographic semantics.

To identify complex pornographic semantics accurately, high-quality porn lexicons should be utilized to provide prior knowledge to guide the classification, like sentiment lexicons used for sentiment classification (Wu et al. 2019; Song et al. 2016; Feng et al. 2015). However, a manually created task-specific lexicon is usually time-consuming and labor-intensive, which will affect the applicability of the model greatly. As far as we know, there are currently no publicly available porn lexicons. Different from previous studies, we aim to automatically generate a high-quality and wide-coverage porn lexicon which can guide sentence representation by focusing more on the keywords while training.

In this paper, we address the rarely studied task of pornographic text identification by proposing a novel and extensible Evidence-Aware Neural Porn Classification (eNPC) model, which conducts document-level porn identification and locates porn indicative sentences simultaneously only under the supervision of document-level labels. The contributions of this paper are three-folds:

- We propose a MIL-based model eNPC, which classifies all the sentences and then aggregates the sentence-level predictions into document-level prediction via attention weights calculated by capturing pornographic context clues, i.e., context semantic and context label predictions.
- We propose an automatic porn lexicon construction strategy which can provide prior knowledge to guide sentence representation by enriching word inputs with type embeddings from lexicons and adjusting word attention appropriately via an auxiliary word-level classification task.
- Extensive experiments conducted on two real-world Chinese novel paragraph datasets and a English personal story dataset demonstrate the effectiveness of our model.

Related Work

Social media data generated by netizens have already been studied to address various social issues, including sexual harassment (Chowdhury et al. 2019; Karlekar and Bansal 2018; Khatua, Cambria, and Khatua 2018), spam content detection (He et al. 2019; Jiang et al. 2019, 2020) and pornographic text identification (Hu et al. 2007). Pornographic text identification is important but rarely studied before. Existing methods are mainly based on keyword matching and statistics and on text classification. Lee et al. (2002; 2005) counted the frequencies with which keywords appear in a text. The frequencies, together with the relevant web page features, are used as the input to the Kohonen self-organizing neural network (KSOM) which will determine whether a text can be classified as pornographic. Du et al. (2003) extracted feature vectors from pornographic and normal texts, and proposed a feature matching method. Hu et al. (2007) trained three classifiers respectively based on continuous text pages, discrete text pages and image pages. A fusion alogirthm, based on Bayes theory, was proposed to fuse the results from texts and images. Recently, He et al. (2020a) proposed a Skim and Intensive Reading Model (SIRM) for detecting harmful contents polluting the web space. Later, He et al. (2020b) further studied to identify indecent readings by augmenting neural network models with human reading behaviors. Compared with previous studies, we aim to identify pornographic texts while locating porn indicative sentences, which is proved more conducive to the platform's control over pornographic content.

Multiple Instance Learning (MIL) is first proposed in image processing field, and recently applied to natural language processing. Angelidis et al. (2018) proposed a MIL-based model to perform document-level and sentence-level sentiment classification with only document labels. Wang et al. (2018) and Song et al. (2019) applied it to peer-reviewed research papers and customer satisfaction analysis. Davani et al. (2019) and Wang et al. (2016) learned using sentence-level evidence for document-level classification. However, all these methods cannot adapt to our porn identification scenario well because of failing to understand implicit pornographic semantics in long texts. Besides, all these methods ignore incorporating prior task-specific knowledge.

Incorporating task-specific prior knowledge will provide helpful information for model training. Wu et al. (2019) took advantage of multi-task learning to learn task-specific word embeddings and word attentions simultaneously. Chen et al. (2019b) proposed an auxiliary tagging task to integrate sentiment commonsense into sequential neural networks. Chen et al. (2019a) retrieved similar concepts from external knowledge base to enhance the semantic representation of short texts. However, these studies relied on existing task-specific lexicons and their modeling strategies are kind of rough. Different from previous studies, we propose an automatic porn lexicon construction strategy called semantic composition (CS) which provides high-quality and wide-coverage pornographic knowledge to enhance sentence representation learning while training.

Our approach to this task is based on multi-instance learning framework which integrates effective techniques designed for our task specifically, such as porn lexicon construction, prior knowledge incorporation, labels of local context integration, and multiple instance learning for locating relevant evidence. To our best knowledge, none of previous work in this domain considers such level of thoroughness on modeling and has achieved comparable performance to ours.

Evidence-Aware Neural Porn Classification

In this section, we first propose a context-sensitive MIL-based model eNPC that classifies each document and all its consisted sentences into binary categories under the supervision of document annotation. The overall model architecture is displayed in Figure 2, which consists of three neural network layers: *Sentence Representation Layer*, *Sentence Classification Layer* and *Document Classification Layer*.

Sentence Representation Layer

Let any document $d \in \mathcal{D}$ containing $|d|$ sentences as $[s_1, ...s_i..., s_{|d|}]$, where $s_i = [w_{i1}, ...w_{it}..., w_{i|s_i|}]$ is the i-th sentence and $|s_i|$ is the sentence length. We transform each sentence s_i into a sequence of low-dimensional dense vectors $\mathbf{e}_i = [\mathbf{e}_{i1}, ...\mathbf{e}_{it}..., \mathbf{e}_{i|s_i|}]$ via a look-up table $\mathbf{E} \in \mathcal{R}^{V \times K}$, where V is the vocabulary size and K is the dimension of word embeddings. Then, $\mathbf{e}_i$ is fed into a Bi-LSTM (Schuster and Paliwal 1997) to produce hidden states $\{\mathbf{h}_{it}\}$ of words, where $\mathbf{h}_{it} \in \mathcal{R}^H$ summarizes the information of the whole sentence centered around the word w_{it} and H is the hidden size. Afterwards, we use attention mechanism to select important words to obtain informative sentence representation. The attention weight $\alpha_{it} \in (0, 1)$ of the t-th word w_{it} in any sentence s_i can be formulated as below:

$$\begin{aligned} \mathbf{u}_{it} &= \tanh(\mathbf{W}_w \mathbf{h}_{it} + \mathbf{b}_w) \\ \alpha_{it} &= \frac{\exp(\mathbf{u}_{it}^{\mathrm{T}} \mathbf{U}_w)}{\sum_{k=1}^{|s_i|} \exp(\mathbf{u}_{ik}^{\mathrm{T}} \mathbf{U}_w)} \end{aligned} \tag{1}$$

where $\mathbf{W}_w$, $\mathbf{b}_w$ and $\mathbf{U}_w$ are learnable model parameters and $\exp(\cdot)$ is an exponential function. The final sentence representation $\mathbf{v}_i \in \mathcal{R}^H$ is the weighted summation of all the hidden states $\{\mathbf{h}_{it}\}$ by the following formula:

$$\mathbf{v}_i = \sum_{t \in [1, |s_i|]} \alpha_{it} \mathbf{h}_{it} \tag{2}$$

Sentence Classification Layer

The sequential sentence vectors $[\mathbf{v}_1, ...\mathbf{v}_i..., \mathbf{v}_{|d|}]$ are then fed into another Bi-LSTM and produce hidden states $[\mathbf{h}_1, ...\mathbf{h}_i, ..., \mathbf{h}_{|d|}]$, where $\mathbf{h}_i \in \mathcal{R}^H$ summarizes the information of the whole document centered around the sentence s_i. Afterwards, each $\mathbf{h}_i$ is fed into a linear layer and then a $softmax$ function to obtain the sentence label $l_i \leftarrow softmax(\mathbf{W}_l \mathbf{h}_i + \mathbf{b}_l)$, where $l_i \in \mathcal{G} = \{0, 1\}$, $\mathbf{W}_l$ and $\mathbf{b}_l$ are learnable parameters shared across all the sentences, labels 1 and 0 denote porn and normal classes, respectively. It is hypothesized that contextual sentences $[s_{i-I}, ...s_i..., s_{i+I}]$ within window size $2I + 1$ near any target sentence s_i can provide useful clues as porn sentences may be close to each other. Therefore, we obtain local context label $\hat{l}_i = max\big\{[l_{i-I}, ...l_i..., l_{i+I}]\big\}$. Finally, we can obtain the predicted probability distribution $\mathbf{p}_{s_i} \in \mathcal{R}^{|\mathcal{G}|}$ for each sentence over classes $\mathcal{G}$ by:

$$\mathbf{p}_{s_i} = softmax\big(\mathbf{W}_s[\mathbf{h}_i; \hat{l}_i] + \mathbf{b}_s\big) \tag{3}$$

where $\mathbf{W}_s$ and $\mathbf{b}_s$ are learnable model parameters, and notation $[;]$ denotes concatenation operation.

Document Classification Layer

To obtain probability distribution for a document d, we opt for an attention mechanism to reward sentences that are more likely to be good predictors. Therefore, we measure the importance of each sentence s_i through a scoring function as below:

$$\begin{aligned} \mathbf{u}_i &= \tanh\big(\mathbf{W}_d[\mathbf{h}_i; \hat{l}_i] + \mathbf{b}_d\big) \\ \beta_i &= \frac{\exp(\mathbf{u}_i^{\mathrm{T}} \mathbf{U}_d)}{\sum_{k=1}^{|d|} \exp(\mathbf{u}_k^{\mathrm{T}} \mathbf{U}_d)} \end{aligned} \tag{4}$$

where $\mathbf{W}_d$, $\mathbf{b}_d$ and $\mathbf{U}_d$ are learnable parameters. Finally, we obtain the document-level probability distribution $\mathbf{p}_d \in \mathcal{R}^{|\mathcal{G}|}$ as the weighted sum of predicted probability distributions of all the sentences:

$$\mathbf{p}_d = \sum_{i \in [1, |d|]} \beta_i \mathbf{p}_{s_i} \tag{5}$$

Note that our approach only needs document labels while sentence labels are unobserved.

We use the cross-entropy loss to minimize the error between the distribution of the predicted document labels and the gold labels of the documents:

$$\mathcal{L}_d = -\frac{1}{|\mathcal{D}|} \sum_{j=1}^{|\mathcal{D}|} \sum_{c=1}^{|\mathcal{G}|} g_{d_j}^c \log\Big(\mathbf{p}_{d_j}^c\Big) \tag{6}$$

where $\mathcal{D}$ is the training set, $g_{d_j}^c$ is 1 or 0 indicating whether class c is a correct answer for the j-th training instance, and $\mathbf{p}_{d_j}^c$ is the predicted probability for class c. We use back propagation to calculate the gradients of all the parameters, and update them with Momentum optimizer (Qian 1999).

Pornographic Semantics Modeling

Identifying pornographic texts is difficult due to both obvious and implied semantics. On one hand, existing models lack prior knowledge to attend on obvious pornographic sensitive words while training, which may lead to poor sentence representation. On the other hand, implicit pornographic text is commonly composed of normal words, which is highly context-sensitive and figurative. In the Figure 2 (Left), we first introduce an effective strategy to build a high-quality porn lexicon automatically from corpus, and then use it to enhance sentence presentation by enriching inputs and guiding attention calculation, simultaneously.

Porn Lexicon Automatic Construction

Pornographic semantics is difficult to understand due to high context-sensitivity and heavy usage of figurative language without any prior knowledge. Thus, we propose an automatic porn lexicon construction strategy called semantic composition (CS), which mainly includes three key steps:

- **Step 1**: Let the annotated corpus as a document-word *tf-idf* matrix $\mathbf{M}_{dw} \in \mathcal{R}^{|\mathcal{D}| \times V}$ and a document-label matrix $\mathbf{M}_{dc} \in \mathcal{R}^{|\mathcal{D}| \times |\mathcal{G}|}$, and then a word-label matrix $\mathbf{M}_{wc} \in \mathcal{R}^{V \times |\mathcal{G}|}$ can be obtained by matrix multiplication as blow:

$$\mathbf{M}_{wc} = \mathbf{M}_{dw}^{\mathrm{T}} \mathbf{M}_{dc} \tag{7}$$

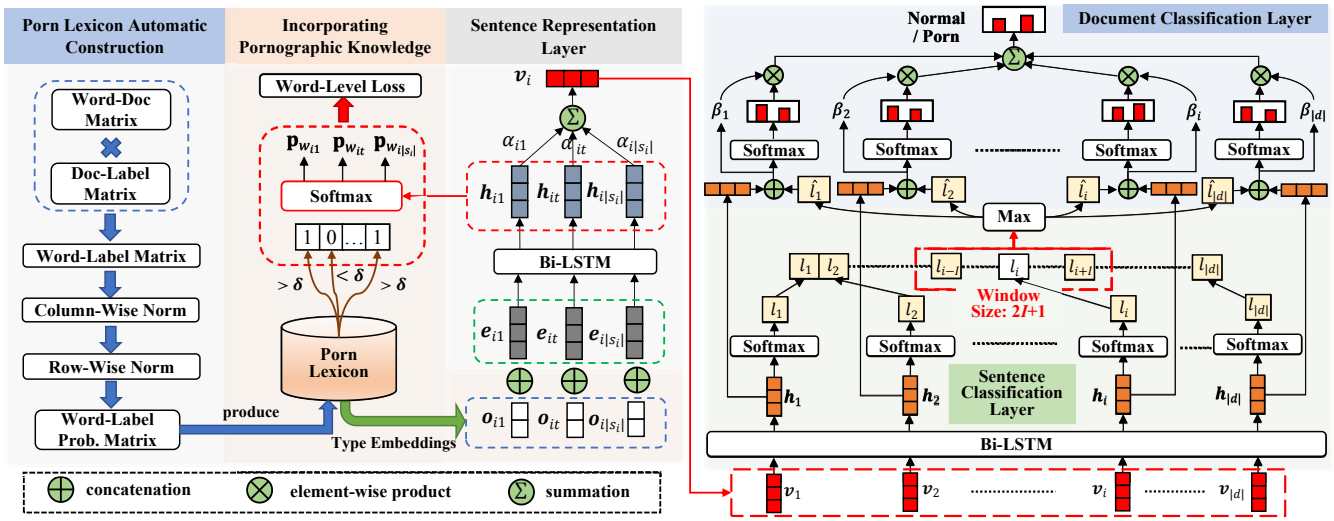

Figure 2: The architecture of our eNPC model with prior knowledge enhancement. The left figure includes the construction of a porn lexicon (*left module*) and prior knowledge enhanced sentence representation layer (*middle&right module*), and the right figure includes the sentence classification layer (*lower module*) and the document classification layer (*upper module*).

- **Step 2**: After that, a normalized matrix $\widetilde{\mathbf{M}}_{wc} \in \mathcal{R}^{V\times|\mathcal{G}|}$ can be obtained by first applying column-wise *min-max* normalization to $\mathbf{M}_{wc} \in \mathcal{R}^{V\times|\mathcal{G}|}$, which makes all the elements under the same label are comparable.

$$\widetilde{\mathbf{M}}_{wc(i,j)} = \frac{\mathbf{M}_{wc(i,j)} - \min\left(\mathbf{M}_{wc(\cdot,j)}\right)}{\max\left(\mathbf{M}_{wc(\cdot,j)}\right) - \min\left(\mathbf{M}_{wc(\cdot,j)}\right)} \tag{8}$$

- **Step 3**: Finally, we scale each row-wise data of $\widetilde{\mathbf{M}}_{wc(i,\cdot)}$ that sums up to one, where each row denotes a porn probability distribution for a word.

$$\widehat{\mathbf{M}}_{wc(i,j)} = \frac{\widetilde{\mathbf{M}}_{wc(i,j)}}{\sum_{t=1}^{|\mathcal{G}|}\widetilde{\mathbf{M}}_{wc(i,t)}} \tag{9}$$

In order to produce a high-quality lexicon, we only reserve the words with high-frequency γ and restrict their part of speech (POS) to nouns, verbs, adjectives and adverbs by using existing POS tagging tools. For any *normal* word, its probability distribution is close to $[1, 0]$; for any *obvious porn* word, its probability distribution is close to $[0, 1]$.

Incorporating Pornographic Knowledge

To enrich input representations, we first concatenate word embedding $\mathbf{e}_{it}$ of each input word w_{it} with its corresponding porn-indication embedding $\mathbf{o}_{it} \in \mathcal{R}^{|\mathcal{G}|}$ by a learnable look-up matrix $\mathbf{E}^p \in \mathcal{R}^{|\mathcal{G}|\times V}$. For the words in lexicon, matrix elements are initialized with their porn probability distributions (see $\widehat{\mathbf{M}}_{wc(i,j)}$ in Equation 9), otherwise randomly initialized. Thus, the original inputs of word sequence $\{\mathbf{e}_{it}\}$ can be further enriched by the formula as below:

$$\mathbf{e}_{it} \leftarrow [\mathbf{e}_{it}; \mathbf{o}_{it}] \tag{10}$$

where the new word vector $\mathbf{e}_{it} \in \mathcal{R}^{K+|\mathcal{G}|}$.

Besides, we aim to make the pornographic words $\{\mathbf{h}_{it}\}$ being paid more attention when modeling sentences. Specifically, we resort to multi-task learning by jointly learning an auxiliary word-level porn classifier which is complementary to our document-level porn classifier. We consider the words with high pornography probabilities larger than a threshold δ (i.e., $\delta = 0.8$) as *porn*, otherwise as *normal*. To this end, we first feed the hidden state $\mathbf{h}_{it}$ of each word into a $softmax$ function to compute word class probability:

$$\mathbf{p}_{w_{it}} = softmax(\mathbf{W}_w\mathbf{h}_{it} + \mathbf{b}_w) \tag{11}$$

where $\mathbf{W}_w$ and $\mathbf{b}_w$ are learnable model parameters. Then, we use the cross-entropy function to minimize the loss between the gold word labels and predicted class probability distribution for each word:

$$\mathcal{L}_w = -\frac{1}{\mathcal{N}}\sum_{j=1}^{|\mathcal{D}|}\sum_{i=1}^{|d_j|}\sum_{t=1}^{|s_i|}\sum_{c=1}^{|\mathcal{G}|} g^c_{w_{it}} \log\left(\mathbf{p}^c_{w_{it}}\right) \tag{12}$$

where $\mathcal{N}$ is the number of words in training set. The final loss function is the weighted sum of $\mathcal{L}_d$ and $\mathcal{L}_w$ as below:

$$\mathcal{L} = \mathcal{L}_d + \lambda\mathcal{L}_w \tag{13}$$

where $\lambda \in [0, 1]$ is an adjustable trade-off factor to measure the relative importance of different parts.

Experiments and Results

In this section, we conduct extensive experiments on three real-world datasets to evaluate the effectiveness of our eNPC. The datasets are described in detail as follows:

- **MiDu** is a Chinese novel paragraph dataset collected by ourselves from an online literature reading platform called MiDu App[3], where each paragraph has up to 30 sentences.

[3] http://www.midureader.com/

MiDu	Train	Dev	Test	Total
#Paragraphs	50,300	6,288	6,288	62,876
#Porn Paragraphs	5,947	709	704	7,360
#Normal Paragraphs	44,353	5,579	5,584	55,516
#Porn sentences	6,852	811	809	8,472
#Normal sentences	309,494	38,841	38,773	387,108
AliWX	**Train**	**Dev**	**Test**	**Total**
#Paragraphs	21,739	2,717	2,718	27,174
#Porn Paragraphs	5,704	725	745	7,174
#Normal Paragraphs	16,035	1,992	1,973	20,000
Safecity	**Train**	**Dev**	**Test**	**Total**
#Stories	7,156	894	895	8,945
#Porn stories	4,868	601	615	6,084
#Normal stories	2,134	293	280	2,707

Table 1: The statistics of the three datasets we used.

We adopt the strictest standard, i.e., any paragraph related to the topics of *sexual description, sexual behavior, sexology knowledge, and sexual vulgarity* are annotated as *Porn*, otherwise *Normal*. Finally, we build a corpus with 62,876 paragraphs including 7,360 pornographic paragraphs and 55,516 normal paragraphs. In order to verify the performance of evidence locating, the dataset is also annotated with sentence-level labels. Note that sentence labels are only used for testing. For the dataset, two well-trained annotators labeled each paragraph and the Cohen's kappa coefficient of inter-rater agreement is 0.85, and then labeled the sentences and the Cohen's kappa coefficient is 0.82. A third annotator made the final decision in case of disagreement.

- **AliWX** is the only publicly available industrial novel dataset collected from Alibaba Literature[4], which contains lots of obscure pornographic content since the authors of these novels may purposely avoid using explicit and sensitive words instead of figurative words because of the censorship (He et al. 2020a). Compared with **MiDu**, **AliWX** is more balanced and its paragraphs are longer.
- **Safecity** is a public English story dataset (Karlekar and Bansal 2018) derived from public sexual harassment personal stories in Safecity online forum[5]. As we focus on binary classification, we ignore the differences among tags by annotating the post with any sexual harassment tag as *Porn*, otherwise *Normal*. Since *Safecity* is small and each consisted story is short, we use *Safecity* as an auxiliary dataset due to the lack of public English porn dataset.

The Chinese datasets are tokenized and POS tagged by a Chinese word segmentation utility called jieba[6]. The English dataset is POS tagged by NLTK[7]. After preprocessing, all the datasets are partitioned into training, development and testing sets with 80/10/10 split. The statistics of the datasets are given in Table 1. All the resources have been released[8].

[4]https://www.aliwx.com.cn/

[5]The dataset contains 3 forms of tags (i.e., groping, ogling and commenting).

[6]https://pypi.org/project/jieba/

[7]http://www.nltk.org/

[8]https://sites.google.com/view/aaai-2021

Experimental Settings

For all the methods, we apply fine-tuning for the word vectors. The Chinese word embeddings are obtained by training CBOW (Mikolov et al. 2013) on our Chinese corpora (i.e., Midu and AliWX). Similarly, the English word embeddings are obtained by training CBOW on our English corpus (i.e., Safecity). The word vectors are initialized by the word embeddings, where the dimension is 300 and the English/Chinese vocabulary size is 10K/155.5K. All the learnable model parameters are initialized by sampling values from a uniform distribution $\mathcal{U}(-0.01, 0.01)$. The hyperparameters are tuned to the best on the development set. The size of hidden states H is 50, the dropout rate is 0.1, the learning rate is 0.1, the learning rate decay is 0.9, the trade-off factor λ is 0.3, the batch size is 32, the window size $I = 2$, and the number of epochs is 10. The lexicon size for MiDu/AliWX/Safecity dataset is 18,129/28,139/2,075 by reserving words with frequency $\gamma \geq 10/10/5$. We use *Macro F1* and *Accuracy* as the evaluation metrics.

Comparative Study

We compare our approach with several state-of-the-art classification methods which are partitioned into two groups that ignore lexicons (models 1-4) or use lexicons (models 5-7).

- **LSTM** is the standard Long Short Term Memory network (Hochreiter and Schmidhuber 1997).
- **HAN** is the hierarchical attention network for document classification (Yang et al. 2016).
- **MILNET** conducts both document-level and sentence-level classification under the supervision of document labels (Angelidis and Lapata 2018).
- **SIRM** is the Skim and Intensive Reading Model which identifies spam texts by using three components: skim reading component, intensive reading component, and adversarial training component (He et al. 2020a).
- **Lex** is an easy heuristic method which identifies pornographic text by judging whether it contains porn words appearing in our constructed porn lexicon.
- **SCSNN** is the Sentiment Aware Attention and Word Embeddings which addresses sentiment analysis by using multi-task learning to conduct sentence-level and word-level classification simultaneously (Chen et al. 2019b).
- **SAATWE** is the Sentiment Commonsense Induced Sequential Neural Networks which uses multi-task learning to learn task-specific word embeddings and word attentions for document classification (Wu et al. 2019).

Results and Analysis: From Table 2, we can find that Lex performs worst because of ignoring contextual semantics. LSTM and Bi-LSTM can not compete with other models because of ignoring different importance of words. HAN and MILNET perform much better by using a two-layer document architecture, but they can not capture complex pornographic semantics well due to the simplicity of their model architectures. SIRM simulates human reading behavior and can better understand semantics via a skim-reading module and intensive-reading module. However, SIRM still can

Methods	Midu		AliWX		Safecity	
	MacroF1	Accuracy	MacroF1	Accuracy	MacroF1	Accuracy
LSTM	84.38	94.15	85.01	88.12	77.65	81.01
Bi-LSTM	84.53	94.08	85.62	88.48	79.09	81.56
HAN	87.11	94.86	88.07	90.62	79.01	81.68
MILNET	85.10	94.23	85.33	88.70	78.57	82.23
SIRM	86.37	94.80	87.94	90.69	80.85	82.91
eNPC (Basic)	88.25	95.67	88.96	91.13	79.85	83.02
Lex	64.83	89.48	54.97	61.03	35.95	38.99
SCSNN	88.11	95.52	88.52	90.77	80.97	83.24
SAATWE	87.93	95.23	88.96	91.46	78.87	82.12
eNPC (Final)	**89.70**	**95.99**	**89.51**	**91.69**	**81.66**	**84.13**

Table 2: Comparison among different classification models.

Methods	Midu		AliWX		Safecity	
	MacroF1	Accuracy	MacroF1	Accuracy	MacroF1	Accuracy
eNPC (Basic)	88.25	95.67	88.96	91.13	79.85	83.02
eNPC-$\hat{l}_i$	87.96	95.55	88.05	90.73	79.45	82.79
eNPC+wp	89.36	95.79	89.21	91.46	80.73	83.46
eNPC+pe	88.81	95.74	89.02	91.46	80.18	82.91
eNPC+wp+pe	**89.70**	**95.99**	**89.51**	**91.69**	**81.66**	**84.13**

Table 3: Comparison among different model configurations.

not outperform ours because its CNN-based modules fail to capture sequential information well and ignore label dependencies of contexts. All these methods ignore prior pornographic knowledge which helps improve the performance.

SCSNN and SAATWE incorporate lexicon into modeling, but they ignore label dependencies of context and can not locate porn indicative sentences. Besides, constructing a high-quality porn lexicon is usually time-consuming and labor-intensive, these models may face large usage restrictions. Compared with all the competitors, eNPC achieves the best result because it makes the most of context information and porn knowledge derived from an automatic lexicon construction method. This again verifies the effectiveness and applicability of our method.

Ablation Study

Different model configurations can largely affect the model performance. In Table 3, we implement several model variants for ablation tests by removing ("-") or adding ("+") different model components. **eNPC-$\hat{l}_i$** ignores labels of local context from Basic eNPC. **eNPC+pe** concatenates porn-indication embeddings with the word embeddings of the Basic eNPC. **eNPC+wp** adds word-level porn classification. **eNPC+wp+pe** is our fully configured model which considers word-level porn classification and porn-indication embeddings, simultaneously.

From Table 3, we can find that eNPC+wp and eNPC+pe outperforms the eNPC because porn lexicon provides helpful prior knowledge on guiding sentence representation. eNPC-$\hat{l}_i$ performs worst because of ignoring the fact that porn sentences are close to each other and target label can be derived from its context labels. Our eNPC+wp+pe (Final eNPC) performs best because of accurately locating key pornographic words and understanding context semantics, which implies the effectiveness of different components.

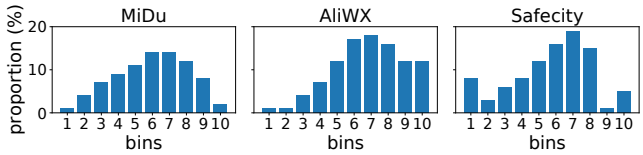

Figure 3: The proportion of lexicon words over probabilities.

words	(normal, porn)	words	(normal, porn)
jade peak	(0.0682, 0.9318)	breast	(0.3843, 0.6157)
plump	(0.3624, 0.6376)	relax	(0.4997, 0.5003)
gasp	(0.5076, 0.4924)	fight	(0.3718, 0.6282)
study	(0.8420, 0.1580)	bracelet	(0.7111, 0.2889)
make love	(0.1589, 0.8411)	caress	(0.2829, 0.7171)

Table 4: The probability distribution of 10 example words.

Porn Lexicon Study

The quality of constructed porn lexicons may influence the classification performance greatly. We first study the porn distribution of all the words in our lexicons. Since we only reserve the words with specified Part-Of-Speech (i.e., *noun*, *verb*, *adjective* and *adverb*), our lexicon is only a subset of the vocabulary. Specifically, we partition all the words into 10 bins according to their porn probabilities, i.e., $\text{bin}_1 = [0, 0.1)$, $\text{bin}_2 = [0.1, 0.2)$, ..., $\text{bin}_9 = [0.8, 0.9)$, $\text{bin}_{10} = [0.9, 1]$. The results are displayed in Figure 3. We can find that the words concentrated in the area of [0.4, 0.8] occupy a larger proportion. This is reasonable because most words can appear in either porn sentences or normal sentences, and these words usually have relatively balanced distributions.

In Table 4, we display 10 example words of our lexicon constructed from the AliWX dataset. We can find that some words are obvious pornographic, such as "*make love*" and "*caress*", some words are normal or pornographic irrelevant, such as "*bracelet*" and "*gasp*", and the remaining words can be used everywhere, such as "*relax*" and "*fight*". Note that the words "*jade peak*" are used as a metaphor for the shape of target " *breast*", thus it is obvious pornographic.

The size of porn lexicons may also influence the classification performance greatly. The sizes of lexicons in MiDu, AliWX and Safecity are 18,129/28,139/2,075, respectively. Therefore, we randomly sample $x\%$ $(x = 0, 25, 50, 75, 100)$ words from the original lexicons, and further study the impact of porn lexicon size on the classification performance. The results are displayed in Figure 4. We can observe that our eNPC performs worst when ignoring incorporating prior knowledge (i.e., 0%), which again verifies the necessity of enhancing sentence representation with porn knowledge. Besides, the performance on MiDu and AliWX gradually improves until stable when sampling proportion is $\geq 50\%$. The results on Safecity achieves consistently improvements because lexicon contributes more to the small dataset. Note that the classification performance may decrease slightly because automatically constructed lexicons may contain noises without manual intervention.

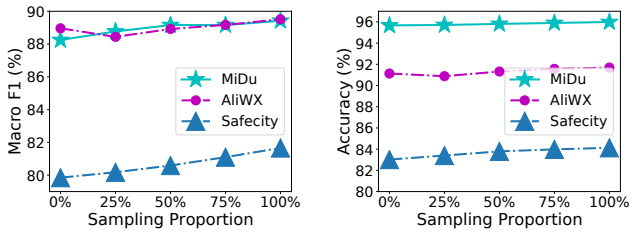

Figure 4: The influences of porn lexicon size

Methods	Top 50%		Last 50%		Midu	
	MacroF1	Accuracy	MacroF1	Accuracy	MacroF1	Accuracy
Lex	52.97	88.73	50.72	94.23	52.23	92.27
eNPC (Basic)	89.41	98.26	76.63	99.20	86.12	98.87
MILNET	86.35	97.88	69.80	98.90	81.63	98.53
eNPC (Final)	**89.79**	**98.51**	**75.78**	**99.35**	**86.48**	**99.05**

Table 5: Comparison among different sentence-level evidence locating methods. The best results are highlighted.

Id	Example Sentences (Chinese → English)
s_1	Help, **bad guys, rogue, assault me!** **Lex:**✓ **MILNet:**✓ **eNPC:**✓ **Truth:**Porn
s_2	Li held breath and listened to the sound in the room. **The warm breath brought a trace of coquettish voice, which clearly meant what was going on.** **Lex:**✓ **MILNet:**✗ **eNPC:**✓ **Truth:**Porn
s_3	Xiao Yan's mouth showed a slight smile, her sexy thin lips came to the base of Ye Xiaomo's ears, and said softly: be quite, don't be angry. **Lex:**✗ **MILNet:**✓ **eNPC:**✓ **Truth:**Normal
s_4	The beauty only wore a big t-shirt and a pair of shorts , her **breast** was high and her skin was fair as fat. **Lex:**✗ **MILNet:**✗ **eNPC:**✓ **Truth:**Normal
s_5	Haha, **his little brother is so small, 25mm**. Hahaha, I have never seen such a little brother. **Lex:**✗ **MILNet:**✗ **eNPC:**✗ **Truth:**Porn

Table 6: The identified results of five example sentences in MiDu. The pornographic fragments are highlighted and the words hitting the porn lexicon are underlined.

Results on Sentence-level Classification

To study the validity of evidence locating ability, we compare the **eNPC** (Basic/Final) with **MILNet** and **Lex** on sentence-level porn classification. Besides, we also study the influences of paragraph lengths on the evidence locating considering our weakly-supervision learning strategy. Thus, we partition the testing set of MiDu into two groups with the same size according to the number of sentences within each paragraph. The results of comparisons are displayed in Table 5. Note that the testing set in MiDu is imbalanced because there are much more normal sentences than porn sentences (i.e., $\frac{\#normal}{\#porn} \approx 47.92$). Thus, the metric of Macro F1 is more suitable and accurate than Accuracy.

From Table 5, we can observe that basic eNPC outperforms MILNet and Lex, which proves the effectiveness of modeling context dependencies among sentences. Our final eNPC using prior knowledge performs best, which again proves the effectiveness of our approach on locating porn indicative sentences. Besides, we can observe that as the bag (i.e., paragraph) contains more segments (i.e., sentences), the classification performance decreases to a certain extent. This is because sentence-level prediction results are inferred indirectly from document labels. A long paragraph with more sentences will bring more noise to pornographic evidence locating. Compared with MILNET, our eNPC model performs much better on longer paragraphs, which proves the robustness of the method.

Case Study

Table 6 displays five example sentences (i.e., s_1-s_5) with the identified results of **Lex**, **MILNET** and our **eNPC**. We also give the gold standards of these examples for reference. The correct identification is marked as ✓, otherwise ✗. The original Chinese sentences in MiDu dataset have been translated into English sentences for understandability.

In Table 6, we can observe that the simple sentence s_1 is identified correctly by all the methods because it has obvious pronographic semantics and hits sensitive words with porn probabities larger than the threshold $\delta = 0.8$ (e.g., "*bad guys*", "*rogue*" and "*assult*") in the porn lexicon. For the sentence s_2, we can find that our eNPC outperforms the MILNet by accurately identifying descriptions of pornographic scenes by considering prior knowledge. The Lex method can also classify the sentence correctly by hitting sensitive words "*coquettish voice*". For the sentence s_3, both eNPC and MILNet perform better than Lex because of capturing semantics of the sentence accurately that does not contain any obvious porn words. The sentence s_4 is normal, but it is misclassified by MILNet which ignores context dependencies and by Lex which mishits the porn word "*breast*". Obviously, the description about clothing and figure does not belong to pornography. For the bad case s_5, all the methods cannot handle this metaphorical description well because of involving common sense reasoning. Through the combination of the normal words "*little brother*" and "*25mm*", it can be inferred that the sentence s_5 is related to pornography. The sentence s_5 is the most difficult and we will leave this for future work.

Conclusion and Future Work

In this paper, we study a porn identification task for child protection by proposing an extensible eNPC approach. It can classify all the sentences by considering context semantics and labels, and then aggregate the sentence-level prediction results into the document-level prediction. A porn lexicon is built automatically from the corpus via a semantic composition strategy and used as prior knowledge to guide model training in both input layer and attention layer. Our model can be customized to protect children from reaching porn text by adopting porn labels of different levels of strictness. Experiments conducted on three real-world datasets indicate the effectiveness of our approach. In the future, we will study fine-grained classification on pornographic texts.

Acknowledgements

We thank all the reviewers for their comments. This work is supported by National Key R&D Program of China (2020YFC0832505).

References

Angelidis, S.; and Lapata, M. 2018. Multiple Instance Learning Networks for Fine-Grained Sentiment Analysis. *TACL* 6: 17–31.

Chen, J.; Hu, Y.; Liu, J.; Xiao, Y.; and Jiang, H. 2019a. Deep Short Text Classification with Knowledge Powered Attention. In *Proceedings of the 33rd AAAI Conference on Artificial Intelligence*, 6252–6259.

Chen, S.; Lin, X.; Xiao, Y.; and He, L. 2019b. Sentiment Commonsense Induced Sequential Neural Networks for Sentiment Classification. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, 1021–1030.

Chowdhury, A. G.; Sawhney, R.; Mathur, P.; Mahata, D.; and Shah, R. R. 2019. Speak up, Fight Back! Detection of Social Media Disclosures of Sexual Harassment. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 136–146.

Davani, A. M.; Yeh, L.; Atari, M.; Kennedy, B.; Portillo-Wightman, G.; Gonzalez, E.; Delong, N.; Bhatia, R.; Mirinjian, A.; Ren, X.; and Dehghani, M. 2019. Reporting the Unreported: Event Extraction for Analyzing the Local Representation of Hate Crimes. In *Proceedings of the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP*, 5752–5756.

Du, R.; Safavi-Naini, R.; and Susilo, W. 2003. Web filtering using text classification. In *11th IEEE International Conference on Networks, ICON 2003, September 28 - October 1, 2003 Sydney, NSW, Australia*, 325–330.

Feng, S.; Song, K.; Wang, D.; and Yu, G. 2015. A word-emoticon mutual reinforcement ranking model for building sentiment lexicon from massive collection of microblogs. *World Wide Web* 18(4): 949–967.

He, G.; Gao, Z.; Jiang, Z.; Kang, Y.; Sun, C.; Liu, X.; and Lu, W. 2020a. Think Beyond the Word: Understanding the Implied Textual Meaning by Digesting Context, Local, and Noise. In *Proceedings of the 43rd SIGIR*, 2297–2306.

He, G.; Kang, Y.; Gao, Z.; Jiang, Z.; Sun, C.; Liu, X.; Lu, W.; Zhang, Q.; and Si, L. 2019. Finding Camouflaged Needle in a Haystack?: Pornographic Products Detection via Berrypicking Tree Model. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 365–374.

He, G.; Kang, Y.; Jiang, Z.; Liu, J.; Sun, C.; Liu, X.; and Lu, W. 2020b. Creating a Children-Friendly Reading Environment via Joint Learning of Content and Human Attention. In *Proceedings of the 43rd SIGIR*, 279–288.

Hellman, S.; Murray, W.; Wiemerslage, A.; Rosenstein, M.; Foltz, P. W.; Becker, L.; and Derr, M. 2020. Multiple Instance Learning for Content Feedback Localization without Annotation. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications, BEA@ACL 2020*, 30–40.

Hochreiter, S.; and Schmidhuber, J. 1997. Long Short-Term Memory. *Neural Computation* 9(8): 1735–1780.

Hu, W.; Wu, O.; Chen, Z.; Fu, Z.; and Maybank, S. J. 2007. Recognition of Pornographic Web Pages by Classifying Texts and Images. *IEEE Trans. Pattern Anal. Mach. Intell.* 29(6): 1019–1034.

Jiang, Z.; Gao, Z.; Duan, Y.; Kang, Y.; Sun, C.; Zhang, Q.; and Liu, X. 2020. Camouflaged Chinese Spam Content Detection with Semi-supervised Generative Active Learning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL*, 3080–3085.

Jiang, Z.; Gao, Z.; He, G.; Kang, Y.; Sun, C.; Zhang, Q.; Si, L.; and Liu, X. 2019. Detect Camouflaged Spam Content via StoneSkipping: Graph and Text Joint Embedding for Chinese Character Variation Representation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, 6186–6195.

Karlekar, S.; and Bansal, M. 2018. SafeCity: Understanding Diverse Forms of Sexual Harassment Personal Stories. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2805–2811.

Khatua, A.; Cambria, E.; and Khatua, A. 2018. Sounds of Silence Breakers: Exploring Sexual Violence on Twitter. In *IEEE/ACM 2018 International Conference on Advances in Social Networks Analysis and Mining*, 397–400.

Lee, P. Y.; Hui, S. C.; and Fong, A. C. M. 2002. Neural Networks for Web Content Filtering. *IEEE Intell. Syst.* 17(5): 48–57.

Lee, P. Y.; Hui, S. C.; and Fong, A. C. M. 2005. An intelligent categorization engine for bilingual web content filtering. *IEEE Trans. Multim.* 7(6): 1183–1190.

Mikolov, T.; Chen, K.; Corrado, G.; and Dean, J. 2013. Efficient Estimation of Word Representations in Vector Space. In *1st International Conference on Learning Representations*.

Qian, N. 1999. On the momentum term in gradient descent learning algorithms. *Neural Networks* 12(1): 145–151.

Schuster, M.; and Paliwal, K. K. 1997. Bidirectional recurrent neural networks. *IEEE Trans. Signal Process.* 45(11): 2673–2681.

Song, K.; Bing, L.; Gao, W.; Lin, J.; Zhao, L.; Wang, J.; Sun, C.; Liu, X.; and Zhang, Q. 2019. Using Customer Service Dialogues for Satisfaction Analysis with Context-Assisted Multiple Instance Learning. In *Proceedings of the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP*, 198–207.

Song, K.; Gao, W.; Chen, L.; Feng, S.; Wang, D.; and Zhang, C. 2016. Build Emotion Lexicon from the Mood of Crowd

via Topic-Assisted Joint Non-negative Matrix Factorization. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, 773–776.

Wang, K.; and Wan, X. 2018. Sentiment Analysis of Peer Review Texts for Scholarly Papers. In *Proceedings of 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, 175–184.

Wang, W.; Ning, Y.; Rangwala, H.; and Ramakrishnan, N. 2016. A Multiple Instance Learning Framework for Identifying Key Sentences and Detecting Events. In *Proceedings of the 25th ACM International Conference on Information and Knowledge Management, CIKM*, 509–518.

Wu, C.; Wu, F.; Liu, J.; Huang, Y.; and Xie, X. 2019. Sentiment Lexicon Enhanced Neural Sentiment Classification. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, 1091–1100.

Yang, Z.; Yang, D.; Dyer, C.; He, X.; Smola, A. J.; and Hovy, E. H. 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*, 1480–1489.