Creating a Children-Friendly Reading Environment via Joint Learning of Content and Human Attention

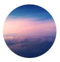
Conference Paper · July 2020

DOI: 10.1145/3397271.3401062

7 authors, including:

Guoxiu He
Wuhan University
10 PUBLICATIONS **25** CITATIONS
SEE PROFILE

Zhuoren Jiang
Zhejiang University
51 PUBLICATIONS **220** CITATIONS
SEE PROFILE



Liu Jiawei
Wuhan University
2 PUBLICATIONS **1** CITATION
SEE PROFILE

Xiaozhong Liu
Indiana University Bloomington
145 PUBLICATIONS **821** CITATIONS
SEE PROFILE

Creating a Children-Friendly Reading Environment via Joint Learning of Content and Human Attention

Guoxiu He[1,2], Yangyang Kang[2], Zhuoren Jiang[3], Jiawei Liu[1,2], Changlong Sun[2], Xiaozhong Liu[4*]
Wei Lu[1*]
[1]School of Information Management, Wuhan University, Wuhan, China
[2]Alibaba Group, Hangzhou, China
[3]School of Public Affairs, Zhejiang University, Hangzhou, China
[4]Indiana University Bloomington, Bloomington, United States
guoxiu.he@whu.edu.cn;yangyang.kangyy@alibaba-inc.com;jiangzhuoren@zju.edu.cn
laujames2017@whu.edu.cn;changlong.scl@taobao.com;liu237@indiana.edu;weilu@whu.edu.cn

ABSTRACT

Technological advancements have led to increasing availability of erotic literature and pornography novels online, which can be alluring to adolescence and children. Unfortunately, because of the inherent complexity of these indecent contents and training data sparseness, it is a challenging task to detect these readings in the Cyberspace while children can easily access them. In this study, we propose a novel framework, Joint Learnin**G** **O**f C**O**ntent an**D** Hu**M**an **A**ttentio**N** (**GoodMan**), to identify indecent readings by augmenting natural language understanding models with large scale human reading behaviors (dwell time per page) on portable devices. From the text modeling viewpoint, the innovative joint attention trained by joint learning is employed to orchestrate the content attention and human behavior attention via the BiGRU. From the data augmentation perspective, various users' reading behaviors on the same text can generate considerable training instances with joint attention, which can be effective to address the cold start problem. We conduct an extensive set of experiments on an online ebook dataset (with human reading behaviors on portable devices). The experimental results show insights into the task and demonstrate the superiority of the proposed model against alternative solutions.

CCS CONCEPTS

• **Human-centered computing → User models**; • **Computer systems organization → Neural networks**; • **Information systems → *Content analysis and feature selection***;

KEYWORDS

user modeling, reading behavior, nature language understanding, neural networks, long text

ACM Reference Format:
Guoxiu He, Yangyang Kang, Zhuoren Jiang, Jiawei Liu, Changlong Sun, Xiaozhong Liu and Wei Lu. 2020. Creating a Children-Friendly Reading Environment via Joint Learning of Content and Human Attention. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '20), July 25–30, 2020, Virtual Event, China.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3397271.3401062

1 INTRODUCTION

Children's curiosity and adolescence's hormone spike can trigger their energetic exploration and discovery of sexual information. Technological advancements, unfortunately, make such information acquisition much easier than ever before. Increasing accessibility of such indecent content (e.g. pornography novels and erotic literature), hidden among massive cyber-readings, exposes young people to the panorama of a distorted view of human sexuality, which can threaten their mental and physical health [1, 3]. Efforts need to be made to detect those readings and create a children-friendly reading environment [46].

Unlike prior classification problems, indecent (pornography) text can be highly creative. And, with limited training data, we can hardly cope with this problem effectively. "*He sits up again and trails a spoonful of ice cream down the center of my body, across my stomach, and into my navel where he deposits a large dollop of ice cream ...*", such seductiveness in *Fifty Shades of Grey* employs few explicit sexual words. Thus, it can be somehow camouflaged for classical detecting solutions, e.g., sensitive words based rules [34, 43] can hardly understand such complex semantics. Similarly, classical data-driven text models [10, 22, 27, 33] could be employed to detect suspect text. However, acquiring decent training data to satisfy machine/deep learning optimization can be expensive.

With an eye-tracking device, we probably find readers' attention, or to say reading focuses. Compared with ordinary readings, these focuses distribute quite differently on indecent text. Readers can pay additional attention to the seductive content in the reading and go through the normal sentences quickly [30, 50]. Scholars won't surprise if eye-tracking devices can essentially empower the detection model, while, practically, we cannot afford the eye-tracking cost of thousands of readers on millions of ebooks.

Reading on personal portable devices, like smartphones and tablets, is increasingly fashion-forward in recent years. In this study, we propose an innovative content plus reading behavior

[*]Corresponding authors.

SIGIR '20, July 25–30, 2020, Virtual Event, China
© 2020 Association for Computing Machinery.
ACM ISBN 978-1-4503-8016-4/20/07…$15.00
https://doi.org/10.1145/3397271.3401062

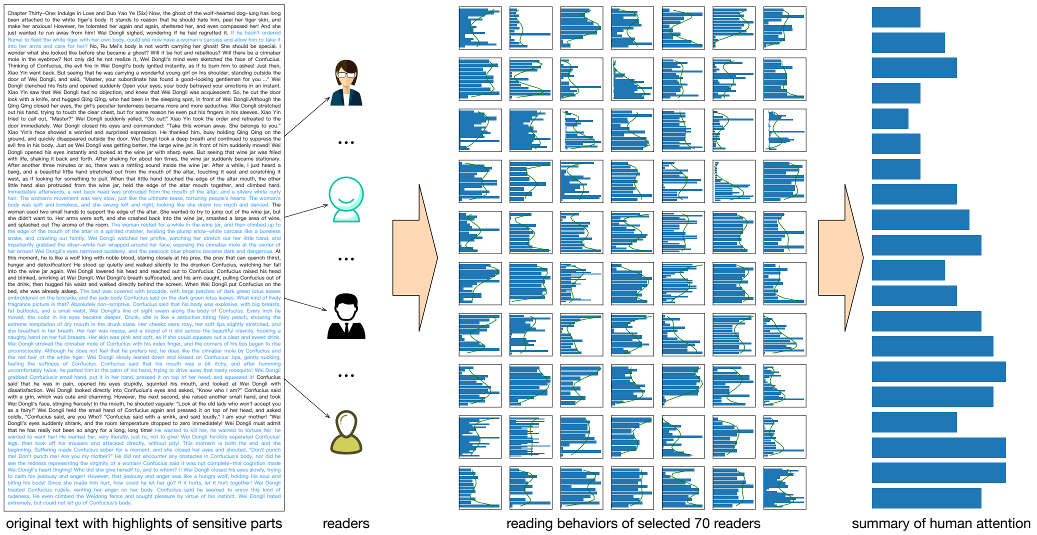

original text with highlights of sensitive parts
readers
reading behaviors of selected 70 readers
summary of human attention

Figure 1: An example of how human reading behaviors (dwell time per page) can provide insight into text understanding. From left to right, the first part is a children-unfriendly (indecent) text. Sensitive parts in this text are highlighted as blue by experts for qualitative analysis. The second part and third part present selected 70 reading behaviors, which are dwell-time distributions of pages per reader. Though reading behavior varies from reader to reader, long-dwell pages are always close to sensitive parts. Finally, the summary of all reading behaviors is illustrated in the fourth part. And this comprehensive distribution is related to sensitive parts of the original text.

tracking based model by using finger screen flip data (collected from portable devices). Unlike eye-tracking with cost and privacy concerns, recording dwell time of each page can be much easier, economical, and applicable. While traditional monitors can display a large amount of textual information, portable device screens often host much a smaller amount of text, which can be ideal to characterize human behavior-based attention [5].

As Figure 1 depicts, for the same indecent (pornography) text, various readers will have different reading focal sentences (spend a longer time before flipping to the next page) when reading with portable devices. We hypothesize that these reading behaviors (dwell time of each page) can offer important potential to differentiate indecent readings from others. In this study, we use "dwell time for each page" to represent users' reading behavior. The observations are listed as follows:

- The focal parts of the same text for different readers could be quite diverse.
- Reading attention (dwell time for each page) is selectively allocated in the document rather than uniformly or normally allocated. People generally focus on the parts they are interested in (e.g., seductiveness in the pornography readings) and spend less time on, even skip, the other parts.
- In most cases, people tend to spend more time on the "sensitive" parts, which indicates human attention and sensitive (indecent) content information have a certain consistency.

In this study, we propose a novel framework, Joint Learnin**G** **O**f C**O**ntent an**D** Hu**M**an **A**ttentio**N** (**GoodMan**), to identify indecent content for young readers. In **GoodMan**, each text can be augmented to multiple samples associated with different users' reading attention tracking on portable devices. Then, the enhanced attention is jointly learned by leveraging the human behavior attention along with the content classification attention. This study proofs that, when training data is sparse, content-based attention can be not trustful, and joint attention learning with human reading behaviors can successfully eliminate the model bias.

From the data augmentation perspective, human reading behavior data can be also critical. As aforementioned, following each individual's reading behavior, the GoodMan triggers instance generation by statistically zooming in the most suspicious content in the target ebook. When thousands of readers explore the same ebook, the proposed model can successfully address the label sparseness problem.

Briefly, our main contributions of this work can be summarized as follows:

- We propose an innovative content plus human behavior tracking based model by using finger flipping data (dwell time of each page) on portable devices. The proposed framework, then, is employed to detect indecent ebooks to create a children-friendly reading environment.
- A joint attention mechanism followed by joint learning is proposed by creatively integrating classical content based attention and human behavior based attention, which can be complementary.
- When a large number of users are reading the same content, the proposed model can be used for data augmentation with various flipping dwell time sequences. This mechanism can be generalized to other NLP problems.
- We collect a large text plus reading behavior dataset to validate the proposed model. Experiment results indicate that GoodMan outperforms existing state-of-art solutions.

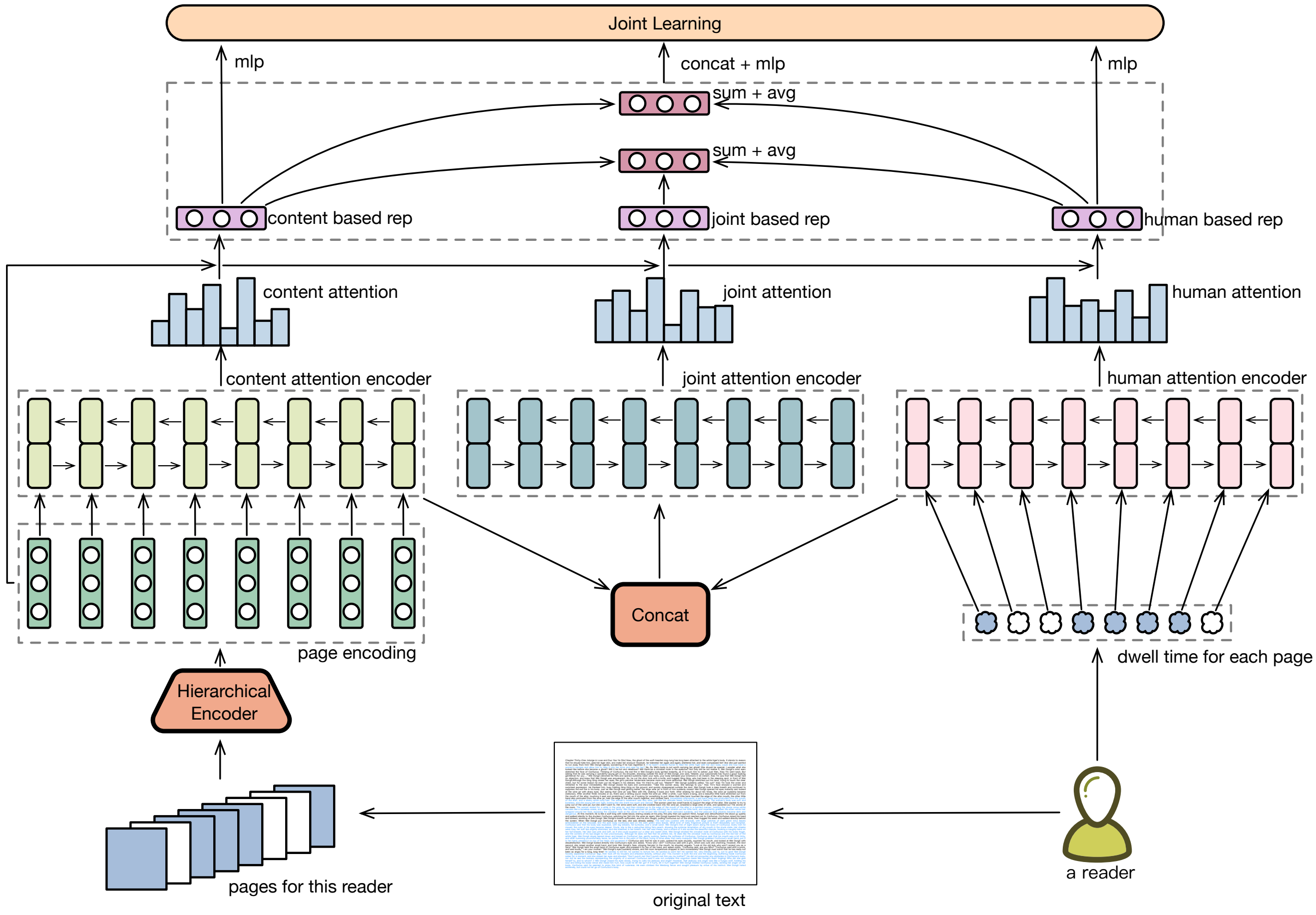


Figure 2: The architecture of the proposed GoodMan for detecting indecent or improper texts. This figure shows how a text can be augmented by a reader. The original text can be divided into several pages with associated reading behavior (dwell time per page) of the reader. After that, a hierarchical encoder, a content attention component, a human attention component, a joint attention component, and a joint learning component are included in the proposed GoodMan to make full use of the human reading behavior and identify the label of the target text.

2 JOINT LEARNING OF CONTENT AND HUMAN ATTENTION

In this section, we propose a novel framework, Joint LearninG Of COntent anD HuMan AttentioN (**GoodMan**). As depicted in Figure 2, GoodMan enables indecent text detection (for Children) for online ebook/reading providers by exploiting text content and users' reading behavior (dwell time per page) via portable devices.

2.1 Overview

In this work, we distinguish indecent contents from given texts by taking advantage of their associated readers' reading behavior tracking (dwell time of each page on portable devices). Each input is represented as a triple $(p, \boldsymbol{H}, y)$, where p is the original text, $\boldsymbol{H}$ denotes the reading behaviors of readers on the text, and $y \in \{0, 1\}$ indicates whether the text is indecent. More specifically, the original text p is a sequence of words and reading behaviors of readers $\boldsymbol{H}$ is a set of dwell time sequences, which are represented as below:

$$\begin{aligned} p &= [w_1, \cdots, w_{n_p}]\,, \\ \boldsymbol{H} &= \{\boldsymbol{h}_1, \cdots, \boldsymbol{h}_{n_H}\}\,, \\ \boldsymbol{h}_i &= [h_{i1}, \cdots, h_{in_h}]\,, \end{aligned} \tag{1}$$

where n_p is the number of words in the input text, n_H is the number of readers who have read the text, and n_h is the number of pages for the $i-$th reader. Since at least one human has read the given text, n_H is not less than 1. And n_h depends on the screen size and the font size of the users' reading devices. In the proposed GoodMan, each word w in text p is mapped into a d_e dimensional word embedding $\boldsymbol{w} \in \mathbb{R}^{d_e}$ by looking up the embedding matrix $\boldsymbol{E} \in \mathbb{R}^{n_w \times d_e}$, where n_w is the number of vocabulary size.

The goal of the proposed GoodMan is to explore the combination of content attention and human attention from the original text and the users' reading behaviors respectively. After giving the $i-$th human's reading behavior of the text, the probability of whether the target text is indecent is estimated by learning the parameters θ:

$$y' = argmax_{y \in Y} Pr(y|p, \boldsymbol{H}_i, \theta) \ . \tag{2}$$

2.2 Hierarchical Encoder

Given the $i-$th reader, the input text p can be augmented as a page sequence $p_i = [s_{i1}, s_{i2}, \cdots, s_{in_h}]$. Hence, the $j-$th page in the text p for the $i-$th reader can be represented as a matrix $\boldsymbol{S}_{ij} \in \mathbb{R}^{n_s \times d_e}$:

$$\boldsymbol{S}_{ij} = lookup(\boldsymbol{E}, s_{ij}) \ , \tag{3}$$

where n_s is the number of words in this page.

For a long text, though it has been cut up to several pages according to the given reading behavior, there are still many words on each page. Hence, it seems not advisable to use recurrent neural networks (RNN) based models to characterize semantics of each page. Meanwhile, the near neighbors of each word play an important role to understand its meaning. Hence, we pad both k words at the head and tail for the given page, and then apply a convolutional neural network (CNN) with $2 \cdot k + 1$ kernel size followed by average pooling to obtain the encoding of each page/screen $\boldsymbol{s}_{ij} \in \mathbb{R}^{d_c}$:

$$\boldsymbol{s}_{ij} = AvgPooling(Conv(\boldsymbol{S}_{ij})) \ , \tag{4}$$

where d_c is the number of the CNN kernels.

Then, the target text which is augmented by the $i-$th reader can be represented as a matrix $\boldsymbol{P}_i \in \mathbb{R}^{n_h \times d_c}$:

$$\boldsymbol{P}_i = [\boldsymbol{s}_{i1}, \boldsymbol{s}_{i2}, \cdots, \boldsymbol{s}_{in_h}] \ . \tag{5}$$

Similarly, near neighbors of each page can provide helpful information for understanding the given page. Hence, a $2 \cdot k + 1$ kernel size based CNN is applied to get the final hierarchical encoding for all pages $\hat{\boldsymbol{P}}_i \in \mathbb{R}^{n_h \times d_c}$:

$$\hat{\boldsymbol{P}}_i = Conv(\boldsymbol{P}_i) \ . \tag{6}$$

2.3 Content Attention Component

Intuitively, the order of the pages is quite important to understand the relationship among these pages. Therefore, we employ a bidirectional gate recurrent unit (BiGRU) to encode the sequence of pages as $\boldsymbol{R}_i^c \in \mathbb{R}^{n_h \times d_r}$:

$$\boldsymbol{R}_i^c = BiGRU(\hat{\boldsymbol{P}}_i) \ , \tag{7}$$

where d_r is the dimension of the BiGRU.

Then, a fully connected neural network (FC) with a *relu* function is applied for every page to estimate whether the target page is indecent. Finally, a *softmax* function is utilized to get the content based attention:

$$\boldsymbol{a}_i^c = softmax(relu(FC(\boldsymbol{R}_i^c))) \ , \tag{8}$$

where $\boldsymbol{a}_i^c \in \mathbb{R}^{n_h}$.

After getting the content attention, the content based representation can be calculated as a weighted sum of hierarchical encoding $\boldsymbol{r}_i^c \in \mathbb{R}^{d_c}$:

$$\boldsymbol{r}_i^c = \boldsymbol{a}_i^c \cdot \hat{\boldsymbol{P}}_i \ . \tag{9}$$

2.4 Human Attention Component

Though we could utilize the human reading behavior (dwell time of each page) to augment the model, the dwell time sequence of pages can be quite noisy (also demonstrated in Figure 1). That means, there are some bias and some abnormal dwell time for some pages. In addition, as a sequential list, each dwell time is influenced by the forward and backward records. Hence, we employ a BiGRU followed by a FC with *relu* and *softmax* functions as a self-smooth encoder to extract a more accurate attention $\boldsymbol{a}_i^h \in \mathbb{R}^{n_h}$:

$$\begin{aligned} \boldsymbol{R}_i^h &= BiGRU(\boldsymbol{h}_i) \ , \\ \boldsymbol{a}_i^h &= softmax(relu(FC(\boldsymbol{R}_i^h))) \ . \end{aligned} \tag{10}$$

Finally, the human based representation $\boldsymbol{r}_i^h \in \mathbb{R}^{d_c}$ is the human attention weighted sum of content representation shown as below:

$$\boldsymbol{r}_i^h = \boldsymbol{a}_i^h \cdot \hat{\boldsymbol{P}}_i \ . \tag{11}$$

2.5 Joint Attention Component

Though the content attention and human attention can provide valuable perspectives to understand the given text and produce a reasonable representation individually, the human attention can be more useful when combined with content. Hence, we use a BiGRU to encode the combination of outputs of content attention encoding and human attention encoding, and then apply a fully connected neural network to estimate the risk for each page as the joint attention:

$$\begin{aligned} \boldsymbol{R}_i^j &= BiGRU(Concat(\boldsymbol{R}_i^c, \boldsymbol{R}_i^h)) \ , \\ \boldsymbol{a}_i^j &= softmax(relu(FC(\boldsymbol{R}_i^j))) \ , \end{aligned} \tag{12}$$

where $\boldsymbol{R}_i^j \in \mathbb{R}^{n_h \times d_r}$ and $\boldsymbol{a}_i^j \in \mathbb{R}^{n_h}$.

Finally, the joint attention based representation is computed by weighted sum as $\boldsymbol{r}_i^j \in \mathbb{R}^{d_c}$:

$$\boldsymbol{r}_i^j = \boldsymbol{a}_i^j \cdot \hat{\boldsymbol{P}}_i \ . \tag{13}$$

2.6 Joint Learning Component

After getting the content based representation $\boldsymbol{r}_i^c$, human based representation $\boldsymbol{r}_i^h$, and joint based representation $\boldsymbol{r}_i^j$, we calculate the following combinations:

$$\begin{aligned} \boldsymbol{r}_i^1 &= (\boldsymbol{r}_i^c + \boldsymbol{r}_i^h)/2 \ , \\ \boldsymbol{r}_i^2 &= (\boldsymbol{r}_i^c + \boldsymbol{r}_i^h + \boldsymbol{r}_i^j)/3 \ . \end{aligned} \tag{14}$$

Then, the final representation can be extracted as follows:

$$\boldsymbol{r}_i^f = relu(FC(Concat(\boldsymbol{r}_i^c, \boldsymbol{r}_i^h, \boldsymbol{r}_i^j, \boldsymbol{r}_i^1, \boldsymbol{r}_i^2))) \ , \tag{15}$$

where $\boldsymbol{r}_i^f \in \mathbb{R}^{d_c}$

Hence, the label can be estimated as y':

$$y' = sigmoid(FC(\boldsymbol{r}_i^f)) \ . \tag{16}$$

Besides, since the content attention and human attention are the essential components for the final representation, we want the two components can benefit from the end-to-end learning directly. Hence, another two outputs can be estimated as y^c and y^h:

$$\begin{aligned} y^c &= sigmoid(FC(\boldsymbol{r}_i^c)) \ , \\ y^h &= sigmoid(FC(\boldsymbol{r}_i^h)) \ . \end{aligned} \tag{17}$$

In the end, the objective function can be defined as follows:

$$\zeta(y, y', y^c, y^h) = \zeta(y, y') + \zeta(y, y^c) + \zeta(y, y^h) \ . \tag{18}$$

And the proposed GoodMan can be trained by using stochastic gradient descent (SGD) methods, such as Adam [26]. More implementation details will be given in Section 4 (Experiments).

2.7 Inference

For each text, there can be multiple users' reading behaviors. In the inference stage, we count how many times that the texts augmented by reading behaviors are detected as indecent via GoodMan. And then, we calculate the proportion of the suspects to the total augmented number. Furthermore, we set a threshold to determine whether the text is indecent based on the proportion. For example, if the threshold is set to 0, the text will be seen as indecent text once there is one suspect. If the threshold is set to 0.9, the text is considered indecent only when the proportion is greater than 0.9. Finally, we rank all texts according to the proportion as the final result for the online textual content provider.

3 DATA COLLECTION

To the best of our knowledge, no public indecent text dataset associated with human reading behavior information (dwell time per page of text) is available. In order to address this problem, we collect Indecent Content Detection Dataset (ICDD) from Alibaba Literature[1], which is one of the largest Chinese novel platforms. ICDD contains selected ebooks and their related human reading behaviors. Based on ICDD, we can provide insights into this problem, and train and evaluate the proposed GoodMan.

In ICDD, there are 2,000 indecent ebooks and 10,000 normal ebooks. All these ebooks are primarily selected by sensitive words based rules[2] and labeled by 3 experts afterwards. For each ebook, users' reading behaviors are extracted from the log of the ebooks provider's website. In this study, one human reading behavior refers to a user's reading of an ebook. We collect the user's flip dwell time on each page during his/her reading. In total, there are 512,263 reading behaviors, which consists of 90,428 reading behaviors in indecent ebooks and 421,835 behaviors in normal ebooks. ICDD is divided into 3 parts: training set, validation set, and test set. There are 500 indecent ebooks and 500 normal ebooks in training and validation set respectively. The rest ebooks are used as test set. 5-fold cross-validation is applied to avoid evaluation bias. Table 1 exhibits the detailed statistics of ICDD.

Table 1: Details of the ICDD including ebooks and associated reading behavior informations (dwell time per page).

		train	val	test
		#behaviors	#behaviors	#behaviors
1	indecent	22,692	22,063	45,673
	normal	21,315	20,629	379,891
	total	44,007	42,692	425,564
2	indecent	23,455	21,721	45,252
	normal	20,853	20,501	380,481
	total	44,308	42,222	425,733
3	indecent	22,067	23,097	45,264
	normal	20,279	21,249	380,307
	total	42,346	44,346	425,571
4	indecent	23,232	23,574	43,622
	normal	21,290	20,917	379,628
	total	44,522	44,491	423,250
5	indecent	21,387	23,571	45,470
	normal	21,031	21,078	379,726
	total	42,418	44,649	425,196

To gain an insightful understanding of ICDD, we analyze the dataset from multiple dimensions. The statistical results are shown in Figure 3. From the preliminary analysis of ICDD, we can draw the following observations:

- Basically, the ebooks in ICDD are long texts. Most of them contain around 2000 words and 3000 characters (refer to a and b in Figure 3). This phenomenon may challenge general semantics representation models.
- The numbers of readers of indecent and normal ebooks are close: both of them retain around 20 users' reading behaviors (refer to g in Figure 3).
- Statistically, the indecent and normal ebooks are indistinguishable. For instance, there is no significant difference between indecent and normal ebooks about average pages per text (refer to c and d in Figure 3); the dwell time of each page or whole text for indecent and normal ebooks is similar (refer to e, f, and h in Figure 3); the top words distributions for indecent and normal ebooks are consistent (refer to i - l in Figure 3).

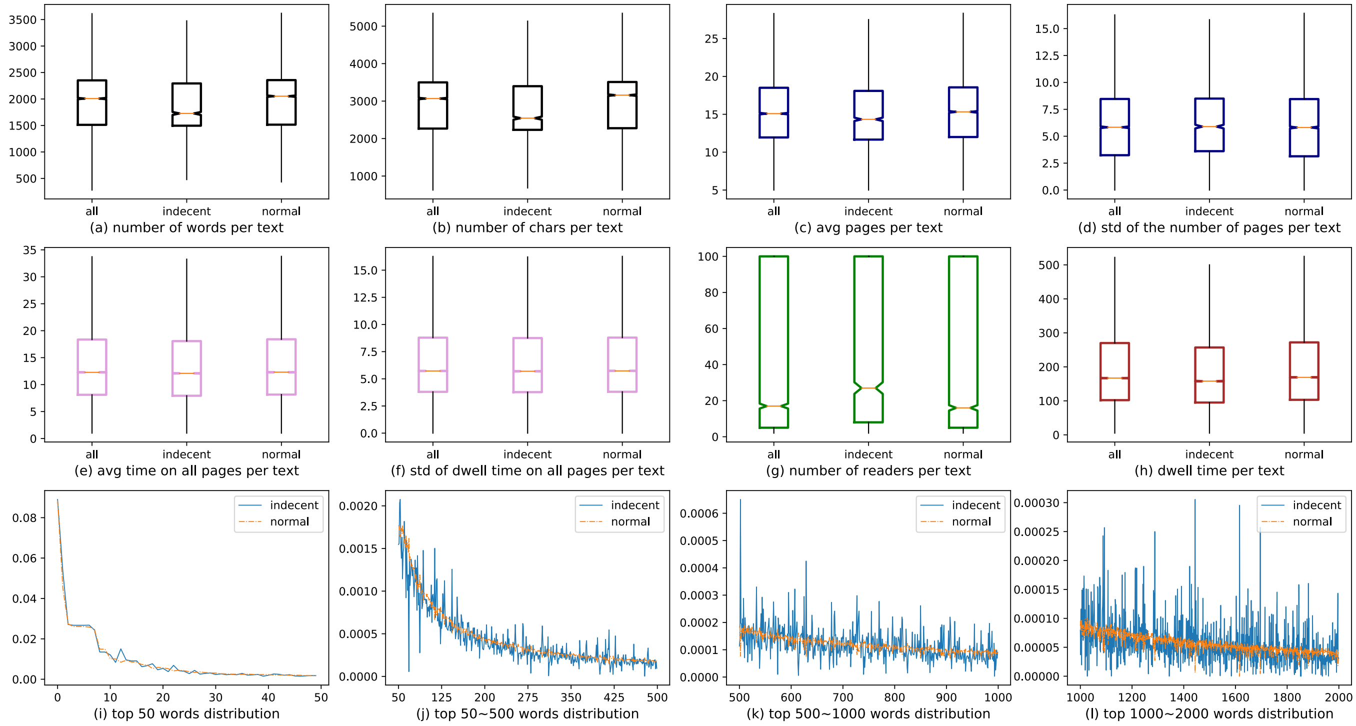
4 EXPERIMENTS

In this section, we introduce extensive experiments to evaluate the proposed GoodMan against a number of alternative solutions, including content only based baselines, straightforward combination of content and human based baselines, and ablation models of GoodMan. In particular, we aim to address the following research questions:

- **RQ1**: Why is it necessary to introduce the human reading behaviors (dwell time per page) into the text understanding model?
- **RQ2**: Is each component in the proposed GoodMan model indispensable?
- **RQ3**: Could different attention components provide different weighting knowledge? And is there any difference in attention between indecent and normal texts?

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

[2] The sensitive contents are reported by Alibaba Literature users. We summarized the sensitive words in these contents and used them to retrieve ebooks that may contain indecent content.



(a) number of words per text (b) number of chars per text (c) avg pages per text (d) std of the number of pages per text

(e) avg time on all pages per text (f) std of dwell time on all pages per text (g) number of readers per text (h) dwell time per text

(i) top 50 words distribution (j) top 50~500 words distribution (k) top 500~1000 words distribution (l) top 1000~2000 words distribution

Figure 3: Data statistics of ICDD. The statistics include number of words/chars per text (a, b), avg/std pages per text (c, d), avg/std dwell time for each page per text (e, f), number of readers per text (g), dwell time per text (h), and word distribution of the dataset (i - l).

As a byproduct of this study, we release the data and codes with the hyper-parameter settings to benefit other researchers[3].

4.1 Alternative Solutions

4.1.1 Content-Only Based Baselines. For models listed below, the input is just content.

Support Vector Machine (SVM) [18]: is a strong and robust machine learning model for a limited training dataset.

Data Augmentation Based SVM (Augmented SVM): More data we have, better performance we can achieve. Therefore, appropriate data augmentation is useful to boost up model performance [48]. If the dwell time information for each page is not available, we can randomly sample sentences (70% in this experiment) from each text to augment the whole dataset. For inference, we use the same voting function as GoodMan (Subsection 2.7). And we set the threshold to 0.

WordAvg [42]: is a simple model based on word embeddings with average pooling. We set the dimension of the word embedding as 64.

SimpleCNN [25]: is a simple CNN model with average pooling using different kernels. There are 7 kinds of filters whose widths are from 1 to 7 and each has 64 different ones.

DeepCNN [25]: Similar to SimpleCNN, we use 3 layers of CNN in DeepCNN.

HieraAttenRNN [49]: is Hierarchical Attention RNN (HieraAttenRNN) which contains hierarchical structure from sentence level to paragraph level. The HieraAttenRNN implements an attention mechanism to the output matrix from each layer in order to aggregate a representation. Here, we use both GRU (HieraAttenGRU) and LSTM (HieraAttenLSTM) as the RNN unit. We use the recommended hyper-parameters from authors.

DPCNN [21]: is Deep Pyramid CNN (DPCNN), which is a low-complexity word-level deep CNN architecture for text categorization, which can efficiently represent long-range dependency in text. We use the default hyper-parameters from authors.

Transformer [45]: This is the state-of-the-art model to encode the deep semantic information via self-attention mechanism[4]. We minimize the batch size to 4 due to the limited GPU memory.

There are two reasons why we can't compare the powerful model BERT [9] directly: 1) BERT can only handle the input whose length is less than 512, but the average length of ebooks is about 2000. 2) Even if we cut the novel into several pieces (the length of each piece is less than 512), we cannot make sure that every piece maintains the same label of the whole document.

4.1.2 Content Plus Human Behavior Based Baselines. For models listed below, the input is a combination of content and associated human reading behavior (dwell time of each page).

[3] https://github.com/GuoxiuHe/GoodMan

[4] https://github.com/tensorflow/models/tree/master/official/transformer

SVM: To show the ability of the human reading information in the simple machine learning models, we select the page, whose dwell time is greater than the average dwell time for each user per text, as input. For inference, we also employ the same voting strategy as GoodMan (SubSection 2.7). We set the threshold to 0.0.

WordAvg: Since the Word Embedding Average is not sensitive to the position of words, we could utilize the same human augmented data as SVM. The difference is that the threshold we use is 0.9 for inference.

HieraAttenLSTM: HieraAttenLSTM can be employed in the content and human based scenario when the input text has been divided into pages depending on the dwell time sequence. Besides the pages (sentences) level attention, a normalized dwell time sequence is used as human attention. For HieraAttenLSTM, we set the threshold to 0.9 for inference.

It is not straightforward to employ the human reading behavior into other content based baselines mentioned in Section 4.1.1.

4.2 Ablation Models of GoodMan

To evaluate the importance and necessity of each component in GoodMan, we conduct extensive ablation tests on GoodMan:

Subtract Joint Learning (SubJoLea): We remove the joint learning component from GoodMan. That means we remove the $\zeta(y, y^c)$ and $\zeta(y, y^h)$ from the final loss function.

Subtract Joint Attention (SubJoAtten): We remove the joint attention component from the GoodMan. That means the combination of the content attention and the human attention is used to estimate the label directly. Note that, this model still maintains the joint learning component.

Subtract Joint Attention and Joint Learning (SubJoAtten& Lea): We remove both the joint attention component and the joint learning component from GoodMan.

Subtract Content Attention (SubContentAtten): We remove the content attention from GoodMan. In that case, only human attention works in the proposed model.

Subtract Human Attention (SubHumanAtten): We remove the human attention from GoodMan. That means, only content attention is useful in GoodMan.

4.3 Evaluation Metrics

In this study, all evaluations and empirical analyses are reported by accuracy, precision, recall, F1 score, F2 score, and Average Precision (AP) with respect to indecent ebooks. Since there are 5 fold experimental datasets, we choose to report the average and standard deviation of all folds' experimental results. Specifically, in this task, due to the harmfulness of indecent content, the recall of all indecent ebooks is relatively more important than the precision. From a comprehensive evaluation viewpoint, we use F1 score, F2 score, and Average Precision as the major indicators to evaluate the models' performance and robustness. In addition, the statistical significance is conducted via the student t-test with *p-value*$\leq$ 0.001.

4.4 Experiment Settings of GoodMan

For the proposed GoodMan, the near-neighbor size k in Hierarchical Encoder is set to 1. The dimension d^r of the BiGRU in all attention encoders are set to 10, and the dimension d of all other layers are all set to 64. In addition, the learning rate is 1×10^{-3} and the batch size is 8. For the ebooks dataset, we use *JIEBA*[5] for tokenization. For inference, we set the threshold to 0.9.

5 RESULTS AND ANALYSIS

This section provides detailed insights into the experimental results, and we also discuss and summarize the experimental outcomes.

5.1 Performance Comparison of GoodMan and Baselines (RQ1)

Table 2 addresses **RQ1** by comparing the performance of GoodMan with all baselines including content based models and content + human based models. The result proofs that GoodMan consistently achieves the best performance in terms of Accuracy (92.47%), F1 Score (67.81%), F2 Score (74.18%), and Average Precision (78.70%).

As the classical machine learning model, SVM achieves decent and comparable results especially in terms of recall. Though the model can produce many misjudgments with low precision and AP, SVM is still the first choice to solve the data sparseness problem. After applying a general data augmentation function, the augmented SVM can successfully enhance the F1 score and AP comparing with SVM. However, although the voting threshold is very low (the threshold is set to 0), this method is limited by gaining higher precision at the expense of recall. And this limitation also results in a lower F2 score. This finding suggests that we should explore more sophisticated approaches for data augmentation.

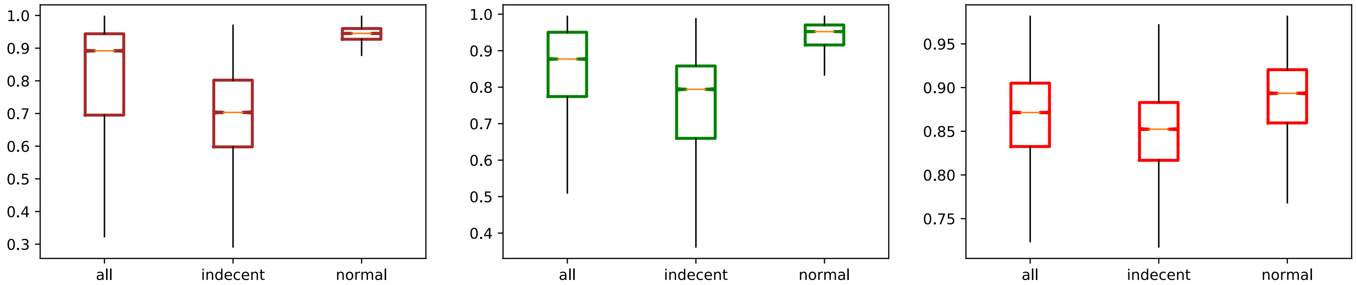
Deep learning comparisons, including some state-of-the-art models, can also be found in the first part in Table 2. Comparing with SVM, an unsatisfactory performance was obtained by the average of word embedding (WordAvg). Furthermore, the document based hierarchical models, such as HieraAttenGRU and HieraAttenLSTM, fail to characterize the semantics of the input text. Though the higher precision of these models leads to a better AP comparing with SVM, the long input limits these models' performance. CNN based models, e.g., SimpleCNN, DeepCNN, and DPCNN, accomplish the best performance so far, i.e., a higher precision or recall to balance the F1 score, F2 score, and Average Precision. These phrase based models can effectively address the cold start problem. With the multiple layers of CNN, the DeepCNN and DPCNN could both encode the short and long dependency, which implement very comparable results in content based baselines. However, the more complex model, Transformer, cannot perform decently because of training data sparseness.

In addition, several straightforward combinations of content and human reading behaviors, which are introduced in Section 4, are listed in the second part in Table 2. With the same voting threshold, human reading behavior enhanced SVM significantly outperforms the simple data augmentation based SVM and the classical SVM, which demonstrates the usefulness of human reading behavior information. Similarly, human based WordAvg and HieraAttenLSTM achieve a better result than content based WordAvg and HieraAttenLSTM according to all indicators. Besides, due to the simple combination mechanism and impropriate text representation model, all these models fail to outperform the best content-only

[5] https://github.com/fxsjy/jieba

Table 2: Experimental results of performance comparison among GoodMan and all alternative models including content based baselines, content + human based baselines, and ablation models of GoodMan. *underline* shows the best performance for baselines and * indicates that GoodMan significantly outperforms the best-perform baselines according to the main indicators (F1 Score, F2 Score, and AP) (p-value ≤ 0.001).

	Model	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)	F2 Score (%)	AP (%)
Content-Only Based Baselines	SVM	83.46±1.45	36.13±2.17	84.08±2.00	50.49±1.94	66.37±1.23	30.40±1.73
	Augmented SVM	37.98±1.44	95.15±0.68	37.00±1.66	53.26±1.69	42.14±1.72	35.90±1.52
	WordAvg	83.12±3.92	35.84±4.15	81.10±7.93	49.38±3.50	64.24±2.74	58.93±6.53
	HieraAttenGRU	88.39±2.30	45.49±7.57	56.24±7.75	49.39±2.13	53.00±3.91	51.69±2.12
	HieraAttenLSTM	86.43±0.96	39.17±2.05	63.62±2.84	48.42±1.37	56.49±1.60	51.90±2.53
	SimpleCNN	87.70±0.97	43.95±2.02	81.18±4.08	56.94±1.09	69.31±1.67	65.63±3.30
	DeepCNN	89.10±0.81	47.48±2.36	79.72±5.23	59.41±1.78	70.08±2.99	68.79±2.06
	DPCNN	88.36±1.20	45.83±2.95	83.98±5.97	59.11±1.52	71.79±2.73	66.88±5.77
	Transformer	88.51±2.32	46.64±7.34	71.70±6.98	55.83±2.97	64.11±2.57	59.49±1.70
Content + Human Based Baselines	SVM	41.64±2.23	93.41±0.62	41.89±2.32	57.81±2.27	47.07±2.36	40.36±2.38
	WordAvg	87.77±1.75	44.61±3.97	84.28±4.85	58.11±2.18	71.29±1.09	74.04±4.76
	HieraAttenLSTM	87.06±0.95	42.46±1.90	81.85±2.33	55.88±1.10	69.00±0.33	70.20±2.32
GoodMan (Ablation)	SubHumanAtten	87.66±1.46	44.34±3.38	**86.56**±3.68	58.49±1.96	72.54±0.49	75.02±2.23
	SubContentAtten	90.12±0.96	50.59±3.04	77.18±3.47	61.02±1.90	69.74±1.96	77.22±3.06
	SubJoAtten& Lea	88.71±0.52	47.58±1.24	85.82±2.30	60.61±0.63	73.00±0.88	75.35±2.63
	SubJoAtten	90.66±0.33	52.30±1.21	76.90±1.91	62.23±0.33	70.26±0.84	76.93±1.93
	SubJoLea	89.57±0.57	48.81±1.53	84.80±2.72	61.93±1.24	73.87±1.60	73.35±2.05
	GoodMan	**92.47**±0.62*	**59.50**±2.50	79.22±2.25	**67.81**±1.09*	**74.18**±1.71*	**78.70**±3.87 *



(a) content attention vs. human attention (b) content attention vs. joint attention (c) human attention vs. joint attention

Figure 4: Detailed difference comparison among the content attention, human attention, and joint attention learnt in GoodMan via cosine similarity.

based models. And the extreme threshold settings (mentioned in Section 4) try their best to balance the precision and recall.

In general, when the high quality training dataset is not available, both classical data augmentation methods, machine learning models, and start-of-art deep learning approaches cannot address this problem effectively. Furthermore, we need to deeply explore the tailored combination mechanism of content and human reading behavior information.

5.2 Analysis of Ablation Models of GoodMan (RQ2)

To address **RQ2**, the proposed GoodMan and its ablation models are presented in the third part of Table 2. Most of the GoodMan family models establish superiority over other baseline methods in the light of main indicators.

Compared with the proposed GoodMan, all the ablation models acknowledge performance drop on all the main indicators. In particular, SubHumanAtten hurts precision excessively. The content attention, without user behavior facilitated data augmentation, cannot characterize the content well. In contrast, maintaining the human attention only in GoodMan, SubContentAtten significantly achieves a better performance than SubHumanAtten. SubJoAtten&Lea receives a lower precision than SubJoAttn, which means the joint learning component could provide essential information to enhance both content attention and human attention. More importantly, based on the result of SubJoLea, we find that the joint learning can be especially useful after obtaining a sophisticated combination of representations.

In summary, each component plays an important role in the proposed GoodMan.

5.3 Analysis of Attentions of GoodMan (RQ3)

In this section, we try to answer **RQ3**. From Figure 4, we provide extensive detailed comparisons among the content attention, the human attention, and the joint attention learnt in GoodMan. It is obvious that various attention mechanisms address the target problem differently.

Human attention inherits essential information from original dwell time. Joint attention, with data heterogeneity, integrates user behavior information and text information.

Based on the result, attentions can be more different in indecent text, which means that, compared with normal text, the human reading behavior information in indecent text could provide more useful information (various aspects) to identify the indecent content.

In a word, all attention components play important roles in GoodMan, and joint attention can significantly enhance the model performance when user reading behavior data (dwell time per page) is available. While lab behavior tracking devices, e.g., eye-tracking devices, are not available, finger flipping tracking (dwell time per page) on portable devices, as an economical and applicable alternative, can be useful and trustful.

6 RELATED WORKS

6.1 Human Reading Behavior

Reading is one of the most essential approaches to get information and learn knowledge for human [50]. Such a process consists of vision processing, language understanding, information gaining, nerves controlling, et al. [8, 30]. The research on how people read has attracted lots of attentions for past decades in various fields, e.g. psychology, linguistics, computer science, and neurology. [40] originally studied the reading process by collecting users' eye movement data in the cognitive psychology field. From the psychologist's perspective, the bilateral cooperative model [8] assumed that when people read words or phrases, they are relating the meaning of these words/phrases with real-world semantics at the same time. The EZ Reader model [41] considered how word identification, visual processing, attention, and oculomotor control as joint determinants on eye movement control. [23] introduced that greater human's processing loads, such as difficult words or implied sentences, make longer pauses via allocation mechanism of eye fixations during reading. These general reading models provide insights into the understanding of the human reading behavior patterns.

Besides general models mentioned above, several specific reading behavior based models are proposed to solve the special task. For example, Two-Stage Examination Model [32] and Reading Model in Relevance Judgment [29] are proposed to model the examination behavior on search engine result pages and the reading behavior patterns during relevance judgment process, respectively. Berrypicking Tree Model [15] is proposed to identify improper content in the E-commerce systems [17] by exploiting users' seeking information. [11] utilizes finger tracking data into a mobile interaction technique that facilitates the recording of audio e-books and their synchronization.

It seems helpful to employ human reading behavior for achieving a good understanding performance.

6.2 Text Classification

Existing text representation and classification studies mainly rely on word embeddings [37] and deep neural networks [28]. A large number of CNNs and RNNs with potential benefits have attracted many researchers' attention. Extensive efforts mainly focus on the application of LSTM [19, 38, 39], GRU [6, 7], SRU [44], and CNNs [12, 14, 21, 24] based on word embeddings [35, 37] drawing on the idea of either language model [4, 36] or spatial parameter sharing. And all these models have demonstrated impressive results in NLP applications. Many previous works have shown that the performance of deep neural networks can be improved by attention mechanism [2, 13]. In addition, self-attention mechanism with position embedding characterizes the mutual relationship between one and others as a dependency to capture the semantic encoding information [45]. There are some other works that combine RNN and CNN for text classification [16, 47, 51] or use a hierarchical neural structure for long document modeling [31, 49]. There are also some studies [20] combining content information and graph information to identify camouflaged text. Though BERT [9] is a powerful pre-trained deep neural model, it is not straightforward and suitable for this task due to the limitation of input length and the sparseness of the training data.

In summary, all these neural networks can not capture real human reading attention and need sufficient data to train satisfactory models. Hence, In this work, instead of eye movement tracks, we take advantage of flip dwell time distribution for each text in order to augment text understanding models.

7 CONCLUSION

In an active cyber-ecosystem, indecent (e.g. pornography) readings can be irresistible and destructive for children and teens. Efforts need to be made to create a children-friendly reading environment for online textual content providers. Unfortunately, sparseness and inventiveness of indecent literature challenge the classical text classification algorithms. In this paper, we propose an innovative model, GoodMan, to detect the indecent content from a large ebook collection. "Attention", in this study, returns to its primeval meaning when originating from human reading behavior data (flip dwell time of each page). Furthermore, we proof that joint attention trained by joint learning, encapsulating both content and human behavior information, can be more trustful. Data augmentation via human reading behavior tracking is nontrivial, and it can be potentially applied to a number of NLP problems. Unlike data collected from eye-tracking devices, finger flip dwell time data on portable devices can be much more affordable. We conduct extensive experiments on an online ebook data. The results validate the effectiveness of the proposed model.

8 ACKNOWLEDGMENTS

The authors are grateful to the anonymous reviewers for their insightful feedback and comments. This work is supported by the National Natural Science Foundation of China (71673211, 71704137, 61876003), Guangdong Basic and Applied Basic Research Foundation (2019A1515010837), and a scholarship from the China Scholarship Council (201906270034).

REFERENCES

[1] Yaman Akdeniz. 2003. Controlling illegal and harmful content on the Internet. In *Crime and the Internet*. Routledge, 125–152.

[2] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. *ICLR* (2015), 1–15.

[3] Jo Bell. 2014. Harmful or helpful? The role of the internet in self-harming and suicidal behaviour in young people. *Mental Health Review Journal* 19, 1 (2014), 61–71.

[4] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A neural probabilistic language model. *Journal of machine learning research* 3, Feb (2003), 1137–1155.

[5] Luca Chittaro. 2006. Visualizing information on mobile devices. *Computer* 39, 3 (2006), 40–45.

[6] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *EMNLP* (2014), 1724–1734.

[7] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555* (2014).

[8] Robert G Crowder and Richard K Wagner. 1992. *The psychology of reading: An introduction.* Oxford university press.

[9] 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* (2018).

[10] Cicero Dos Santos and Maira Gatti. 2014. Deep convolutional neural networks for sentiment analysis of short texts. In *COLING*. 69–78.

[11] Carrie Demmans Epp, Cosmin Munteanu, Benett Axtell, Keerthika Ravinthiran, Yomna Aly, and Elman Mansimov. 2017. Finger tracking: facilitating non-commercial content production for mobile e-reading applications. In *Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and Services*. 1–15.

[12] Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. 2017. Convolutional Sequence to Sequence Learning. In *ICML*. 1243–1252.

[13] Guoxiu He, Junwei Fang, Haoran Cui, Chuan Wu, and Wei Lu. 2018. Keyphrase Extraction Based on Prior Knowledge. In *JCDL*. 341–342. https://doi.org/10.1145/3197026.3203869

[14] Guoxiu He, Zhe Gao, Zhuoren Jiang, Yangyang Kang, Changlong Sun, Xiaozhong Liu, and Wei Lu. 2020. Think Beyond the Word: Understanding the Implied Textual Meaning by Digesting Context, Local, and Noise. In *SIGIR*. ACM.

[15] Guoxiu He, Yangyang Kang, Zhe Gao, Zhuoren Jiang, Changlong Sun, Xiaozhong Liu, Wei Lu, Qiong Zhang, and Luo Si. 2019. Finding Camouflaged Needle in a Haystack?: Pornographic Products Detection via Berrypicking Tree Model. In *SIGIR*. ACM, 365–374.

[16] Guoxiu He and Wei Lu. 2018. Entire Information Attentive GRU for Text Representation. In *ICTIR*. ACM, 163–166.

[17] Guoxiu He, Yunhan Yang, Zhuoren Jiang, Yangyang Kang, Xiaozhong Liu, and Wei Lu. 2020. Implicit Products in the Decentralized eCommerce Ecosystems. In *JCDL*. ACM.

[18] Marti A. Hearst, Susan T Dumais, Edgar Osuna, John Platt, and Bernhard Scholkopf. 1998. Support vector machines. *IEEE Intelligent Systems and their applications* 13, 4 (1998), 18–28.

[19] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[20] Zhuoren Jiang, Zhe Gao, Guoxiu He, Yangyang Kang, Changlong Sun, Qiong Zhang, Luo Si, and Xiaozhong Liu. 2019. Detect Camouflaged Spam Content via StoneSkipping: Graph and Text Joint Embedding for Chinese Character Variation Representation. In *EMNLP*. 6188–6197.

[21] Rie Johnson and Tong Zhang. 2017. Deep pyramid convolutional neural networks for text categorization. In *ACL*, Vol. 1. 562–570.

[22] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of Tricks for Efficient Text Classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. 427–431.

[23] Marcel A Just and Patricia A Carpenter. 1980. A theory of reading: From eye fixations to comprehension. *Psychological review* 87, 4 (1980), 329.

[24] Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom. 2014. A convolutional neural network for modelling sentences. *ACL* (2014), 655–665.

[25] Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *EMNLP*. 1746–1751.

[26] Diederik P Kingma and Jimmy Ba. [n. d.]. Adam: A method for stochastic optimization. In *ICLR*. 1–15.

[27] Siwei Lai, Liheng Xu, Kang Liu, and Jun Zhao. 2015. Recurrent convolutional neural networks for text classification. In *AAAI*.

[28] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *nature* 521, 7553 (2015), 436.

[29] Xiangsheng Li, Yiqun Liu, Jiaxin Mao, Zexue He, Min Zhang, and Shaoping Ma. 2018. Understanding Reading Attention Distribution during Relevance Judgement. In *CIKM*. ACM, 733–742.

[30] Xiangsheng Li, Jiaxin Mao, Chao Wang, Yiqun Liu, Min Zhang, and Shaoping Ma. 2019. Teach Machine How to Read: Reading Behavior Inspired Relevance Estimation. In *SIGIR*. ACM, 795–804.

[31] Rui Lin, Shujie Liu, Muyun Yang, Mu Li, Ming Zhou, and Sheng Li. 2015. Hierarchical recurrent neural network for document modeling. In *EMNLP*. 899–907.

[32] Yiqun Liu, Chao Wang, Ke Zhou, Jianyun Nie, Min Zhang, and Shaoping Ma. 2014. From skimming to reading: A two-stage examination model for web search. In *CIKM*. ACM, 849–858.

[33] Huma Lodhi, Craig Saunders, John Shawe-Taylor, Nello Cristianini, and Chris Watkins. 2002. Text classification using string kernels. *Journal of Machine Learning Research* 2, Feb (2002), 419–444.

[34] Andrew McCallum and Kamal Nigam. 1999. Text classification by bootstrapping with keywords, EM and shrinkage. In *Unsupervised Learning in Natural Language Processing*.

[35] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).

[36] Tomas Mikolov, Martin Karafiát, Lukas Burget, Jan Cernockỳ, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model.. In *Eleventh Annual Conference of the International Speech Communication Association*, Vol. 2. 3.

[37] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *NeurIPS*. 3111–3119.

[38] Hamid Palangi, Li Deng, Yelong Shen, Jianfeng Gao, Xiaodong He, Jianshu Chen, Xinying Song, and Rabab Ward. 2016. Deep sentence embedding using long short-term memory networks: Analysis and application to information retrieval. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)* 24, 4 (2016), 694–707.

[39] Karl Pichotta and Raymond J Mooney. 2016. Using sentence-level LSTM language models for script inference. In *ACL*. 279–289.

[40] Erik D Reichle, Alexander Pollatsek, Donald L Fisher, and Keith Rayner. 1998. Toward a model of eye movement control in reading. *Psychological review* 105, 1 (1998), 125.

[41] Erik D Reichle, Keith Rayner, and Alexander Pollatsek. 2003. The EZ Reader model of eye-movement control in reading: Comparisons to other models. *Behavioral and brain sciences* 26, 4 (2003), 445–476.

[42] Dinghan Shen, Guoyin Wang, Wenlin Wang, Martin Renqiang Min, Qinliang Su, Yizhe Zhang, Chunyuan Li, Ricardo Henao, and Lawrence Carin. 2018. Baseline needs more love: On simple word-embedding-based models and associated pooling mechanisms. In *ACL*. 440–450.

[43] György Szarvas. 2008. Hedge classification in biomedical texts with a weakly supervised selection of keywords. In *Proceedings of ACL-08: HLT*. 281–289.

[44] Sida I. Wang Hui Dai Tao Lei, Yu Zhang and Yoav Artzi. 2018. Simple Recurrent Units for Highly Parallelizable Recurrence. In *EMNLP*. 4470–4481.

[45] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NeurIPS*. 5998–6008.

[46] Gudrun Wallmyr and Catharina Welin. 2006. Young people, pornography, and sexuality: sources and attitudes. *The Journal of School Nursing* 22, 5 (2006), 290–295.

[47] Chenglong Wang, Feijun Jiang, and Hongxia Yang. 2017. A hybrid framework for text modeling with convolutional RNN. In *SIGKDD*. ACM, 2061–2069.

[48] Jason Weston, Sayan Mukherjee, Olivier Chapelle, Massimiliano Pontil, Tomaso Poggio, and Vladimir Vapnik. 2001. Feature selection for SVMs. In *NeurIPS*. 668–674.

[49] 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*. 1480–1489.

[50] Yukun Zheng, Jiaxin Mao, Yiqun Liu, Zixin Ye, Min Zhang, and Shaoping Ma. 2019. Human Behavior Inspired Machine Reading Comprehension. In *SIGIR*. ACM, 425–434.

[51] Chunting Zhou, Chonglin Sun, Zhiyuan Liu, and Francis Lau. 2015. A C-LSTM neural network for text classification. *arXiv preprint arXiv:1511.08630* (2015).