24th European Conference on Artificial Intelligence - ECAI 2020
Santiago de Compostela, Spain

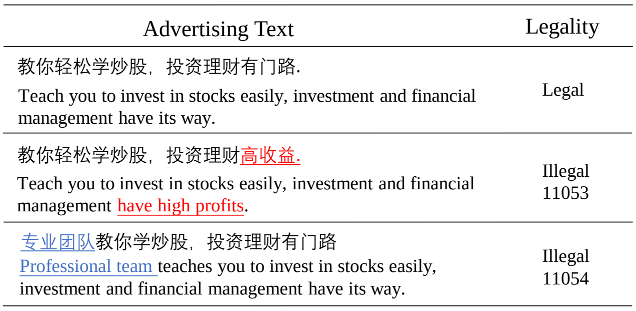
IAD: A Benchmark Dataset and a New Method for Illegal Advertising Classification

Zebo Liu [1] and **Kehan Li**[1] and **Xu Tan**[2] and **Jiming Chen**[1]

Abstract. While online advertising becomes ubiquitous and the pillar of the economy in Internet industry, there are increasing illegal ads which contain misleading or deceptive content and hinder the healthy development of online advertising. How to detect illegal advertising and classify it according to the provisions it violates, is critical for legal supervision. However, due to the difficulty of dataset acquisition and the lack of expert knowledge in advertising, benchmark datasets and methods for illegal advertising classification are scarce. In this paper, we collect and release a large-scale dataset for illegal advertising classification (called *IAD*, short for illegal ads), which contains the content of illegal ads and the corresponding violated provisions. IAD dataset has been released. Based on the IAD dataset, we further propose a novel method called *IAD-Net* to classify the violated provisions of the illegal ads. IAD-Net mainly adopts an interactive attention-based parallel LSTM network, where the parallel structure integrates the provision into classification process, equivalent to using prior information to supervise the classification. Besides, IAD-Net introduces an auxiliary embedding layer to enhance the semantics of lexicons in short ads, and an interactive attention mechanism to capture the relationship between lexicons in ads and its legality. We conduct comprehensive study on the IAD dataset and benchmark several previous methods as well as the proposed IAD-Net for illegal advertising classification. Experimental results demonstrate that IAD-Net achieves good accuracy and outperforms all the previous methods on IAD dataset. We believe the proposed IAD dataset and IAD-Net will help accelerate the research in the area of illegal advertising classification.

1 INTRODUCTION

Advertising serves as a commercial campaign in which commercial operators or service providers directly or indirectly promote the merchandise or service by certain media forms. Considering the exponential growth of the Internet users [11, 5], online advertising has gradually become the dominant way of the advertising market. [4] shows that online advertising, as a business with billions of dollars, is growing with a ratio of double-digit, where the illegal advertising is also growing rapidly due to its high profit margin ranging from 86% to 93%. Meanwhile, the unsound regulatory system leads to more illegal advertising such as misleading or deceptive ads to seek high profits. Hence, it is crucial to strengthen the supervision and control of Internet ads. Discriminating the legitimacy of ad is essential but far from enough. Identifying the violated provisions of illegal ads precisely is much more difficult and urgent, due to that it can guide how to supervise and punish the owner of illegal ads.



Advertising Text	Legality
教你轻松学炒股，投资理财有门路. Teach you to invest in stocks easily, investment and financial management have its way.	Legal
教你轻松学炒股，投资理财高收益. Teach you to invest in stocks easily, investment and financial management have high profits.	Illegal 11053
专业团队教你学炒股，投资理财有门路 Professional team teaches you to invest in stocks easily, investment and financial management have its way.	Illegal 11054

Figure 1. Illegality of ads while a few words change in ads. 11053 and 11054 are different provisions [3].

However, identifying the violated provisions of illegal ads is still an unexplored area due to the following reasons: (1) High-quality dataset is difficult to obtain due to that data collection and labeling requires professional knowledge in the fields of ads and law. To the best of our knowledge, few datasets can be leveraged for illegal ads classification. (2) Due to the rigorousness of lexicons and the incorporation of provisions, previous methods cannot handle the classification of illegal ads well. In the legal field, subtle changes in one or two words can subvert the legitimacy or completely migrate the provisions violated [28], which is especially noticeable in Chinese ads as shown in Figure 1.

To address the above problems, in this paper, we construct and release a large-scale dataset called *IAD* (short for illegal ads) from scratch, to facilitate the study on classification and more in-depth studies of illegal advertising on Internet. Based on the proposed IAD dataset, we further propose a novel interactive attention-based parallel LSTM network called *IAD-Net* to identify the violated provisions of illegal ads precisely. The IAD-Net consists of three major components: (1) An auxiliary embedding layer to enhance semantics and solve the sparseness of short ads effectively. (2) An interactive attention mechanism to capture the interaction between illegal ads and provisions, extract local semantic changes in ads, capture illegal expressions of ads and elements of provisions simultaneously. (3) A parallel structure to enable IAD-Net to process ads and provisions simultaneously, which effectively utilizes the content of provisions.

We conduct comprehensive study on IAD dataset and compare the accuracy of the proposed IAD-Net with previous methods. The classification accuracy of IAD-Net outperforms ten strong baselines.

[1] Zhejiang University, China, email: {liu_zzz,khli,cjm}@zju.edu.cn
[2] Microsoft Research Asia, China, email: xuta@microsoft.com

[3] 11053 stipulates that investment ads can not guarantee the income. 11054 stipulates that investment ads can not be in the name of a professional institution.

IAD-Net can locate the key word pairs between ads and provisions for interpretability and explore the specific illegal expressions of ads. The main contributions of this paper are summarized as follows:

- We construct and release a dataset called IAD for Internet illegal advertising.
- We propose a novel method called IAD-Net to identify the violated provisions of the illegal ads.
- We conduct comprehensive experimental study on the proposed IAD dataset and IAD-Net, and demonstrate that our IAD-Net outperforms several state-of-the-art regular text classification models.

2 RELATED WORK

Most popular classification models are not perfectly suitable for multi-label illegal advertising text classification based on their performances in our experimental results, because these models lack targeted design to process linguistic particularity of ads and the relevance between ads and provisions.

In general, text classification models can be briefly divided into two classes, including inner semantic analysis and external corpus-assisted models. Models from the first class, such as LDA [1], mainly focus on the inner semantic structure of text and extract potential semantic features. Many models modify the topic model to a semi-supervised or supervised one so as to make use of prior knowledge on the distribution of topics [36, 35, 32, 25], but they cannot capture the correspondence of subtle lexicons changes in ads, nor make full use of provision context. The second class utilizes unlabeled data or external corpus, which produce considerate improvement in learning accuracy [34, 23, 2]. To some extent, unlabeled ads may enhance the semantic richness of labeled ads. However, the external corpus may change the illegality of ads even with a synonym replacement. Recently, models based on deep learning methods have been proposed. For instance, the HFT-CNN model learns hierarchical category structure for multi-label [22] and the Att-BLSTM model captures crucial semantic information in a sentence [39]. Besides, TextCNN[14] and FastText[13] are also useful in text classification tasks. However, deep models are struggling in the sparseness of ads [35].

3 DATASET

In this section, we will illustrate the data source, detailed construction process and statistics of our IAD dataset.

3.1 Data Source

We cooperate with China Internet Advertising Monitoring Center and crawl a large amount of suspected adverting data on various websites. We adopt three ways of collecting data, which are SDK acquisition, search engine acquisition and specific websites acquisition. For each website visited, we go to three levels of web pages at least and collect 1000+ locations. Distributed crawling is conducted across the Internet, mainly focusing on search engine websites, trading platforms and other websites with large traffic. The crawled data mainly consists of three sources including video ads, picture ads and text ads. The advertising contents of video and picture will be identified and transferred to text. Data cleaning and labeling are based on this raw text data. Automatically categorizing the ads without the help of legal professionals is an open challenge that leads to the deficiency of Internet illegal ads dataset, so we have legal professionals to participate in the construction of datatset and ensure the correctness and rigorousness of annotation.

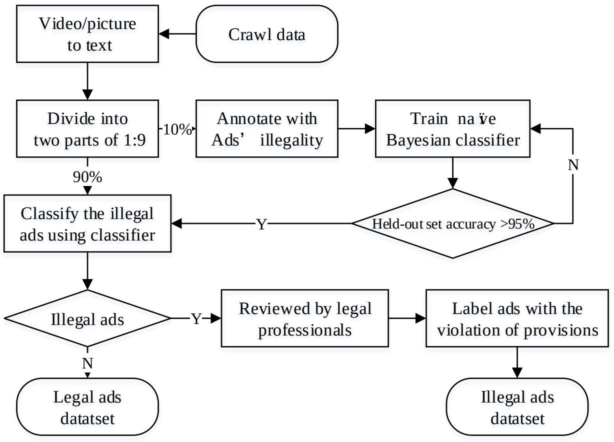

Video/picture to text
Crawl data
Divide into two parts of 1:9
10%
Annotate with Ads' illegality
Train naïve Bayesian classifier
N
90%
Classify the illegal ads using classifier
Y
Held-out set accuracy >95%
Illegal ads
Y
Reviewed by legal professionals
Label ads with the violation of provisions
N
Legal ads datatset
Illegal ads datatset

Figure 2. The flowchart of construction process of IAD dataset.

3.2 Construction Process

Figure 2 demonstrates the overall construction process. Dataset is built via the following steps. First, we crawl ads from the Internet, 10% of which are labeled as illegal or legal by legal professionals according to their illegality and the remaining 90% are unlabeled. Second, these annotated data is used to train a polynomial naive Bayesian classifier, whose accuracy has achieved more than 95% in held-out set. TF-IDF is used for feature selection during training Bayesian classifier. The illegal ads are extracted automatically from the remaining 90% unannotated ads by the classifier. It should be noted that the classifier can only discriminate whether an ad is illegal but cannot determine the specific violation of the provisions. Besides, we train other classifiers such as SVM for comparison. The accuracy of SVM in held-out set achieves 94%, which means it is a relatively simple task to judge whether an ad is illegal or not. Third, legal professionals review the illegal ads extracted from the polynomial naive Bayesian classifier in the second step and label them by the provisions they violated according to *The Advertisement Law of People's Republic of China*. All illegal ads have at least one label. Moreover, we can get the dataset of legal ads at the same time. All manual labeling processes are cross-labeling, which means each ad is labeled by three legal professionals at least to ensure its correctness and rigorousness. At this point, we construct the Internet illegal ads dataset labeled with the violation of the provisions. We open source the IAD dataset for reference[4] .

3.3 Statistics

Table 1. Statistical information of the dataset.

# of labels	# of ads	# of ad categories	Avg length per ad	Vocab size
13	48294	78	55	191832

Table 1 shows the statistical information of the dataset. Some provisions are violated by few ads, therefore, we retain the illegal ads for labels with a sufficient sample size to avoid the imbalance of the

[4] IAD dataset: https://github.com/socknice/IAD-dataset/tree/master

Table 2. Sample of dataset and its translation. Ads will be removed during experiments if its length is less than 10 after segmentation.

Components	Ad Title	Ad Content	Ad Media	Ad Onwer	Ad Category	Illegal Level	Ad Provision
Chinese	和信贷	低门槛高收益15%年收益	360搜索	和信电子商务有限公司	金融服务	严重违法	11053
Translation	Hexin Credit	Low barriers, high profit, 15% annual profit	360 Search	Hexin E-commerce Co.	Financial Services	Serious	11053

Ad Provision	Chinese	Translation
11053	招商等有投资回报预期的商品或者服务广告，含有对未来效果、收益或者与其相关的情况作出保证性承诺，明示或者暗示保本、无风险或者保收益等内容	Ads for merchants or services with expected return on investment promise future benefits, express or imply that the capital preservation, risk-free or guaranteed income.

dataset. The final dataset has 13 labels. The average length of illegal ads is only 55, most of which belong to short text, which hinders the classification task. Table 2 shows a sample illegal ad and each ad consists of seven components including Ad Title, Ad Content, Illegal Level, etc. In the experiments, Ad name and Ad content are concatenated as input sequence of Ad Subnet and the provision is the input sequence of Provision Subnet. Figure 3 illustrate the proportion of illegal advertising categories. Although there are 78 categories of illegal ads, but illegal ads in investment related categories (5 of 78) account for more than 50%. Most of the illegal ads are concentrated in the financial related fields. These ads are very likely to cause large property losses, even involving fraud, and have serious negative effects. Figure 4 shows that more than 90% of illegal ads violate only one provision, while a small part of ads violate two or more provisions at the same time. This also reflects that provisions are rigorous and their definitions are not overlapping. Figure 5 shows that fields like ad_title and ad_media are almost intact, but ad_owner has missing contents in both illegal ads and legal ads. However, the proportion of illegal ads is up to 17.68% and nearly twice that of legal ads (9.32%), which means ad owners of illegal ads tend to hide their identity to avoid supervision and punishment.

Based on our dataset, other in-depth studies can be conducted in the field of Internet advertising, such as amending the existing laws, mining the common characteristics of illegal advertising, and improving the comprehensiveness of Internet illegal advertising detection.

4 IAD-NET FOR ILLEGAL ADS CLASSIFICATION

In this section, we come to detailed architecture of IAD-Net for illegal ads classification. The structure of IAD-Net is shown in Figure 6.

4.1 Parallel Subnet

Provisions contain crucial prior information and form an objective reference for the illegal ads. Therefore, parallel sub-networks are proposed to encode information of illegal ads and provisions simultaneously. As shown in Figure 6, the Ad Subnet mainly consists of Bi-LSTM layer and an auxiliary embedding layer, and takes the sequential data of short illegal ads as input vectors. Considering the characteristics of short ads, Ad Subnet adds an additional auxiliary

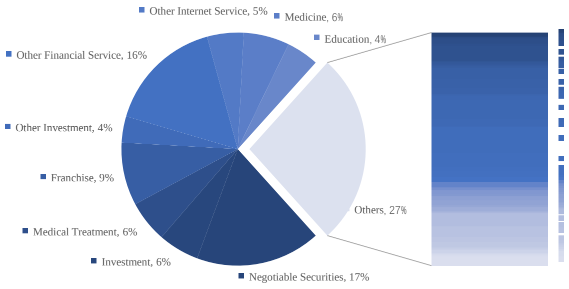


Figure 3. Proportion of illegal advertising categories.

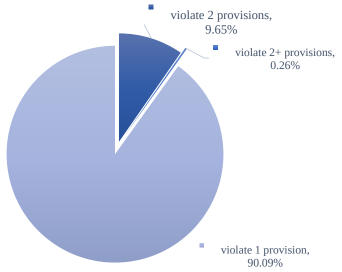


Figure 4. Proportion of illegal ads with different number of violated provisions.

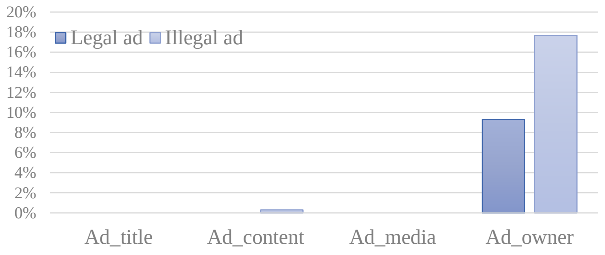


Figure 5. Proportion of missing key fields content in illegal ads data and illegal ad data.

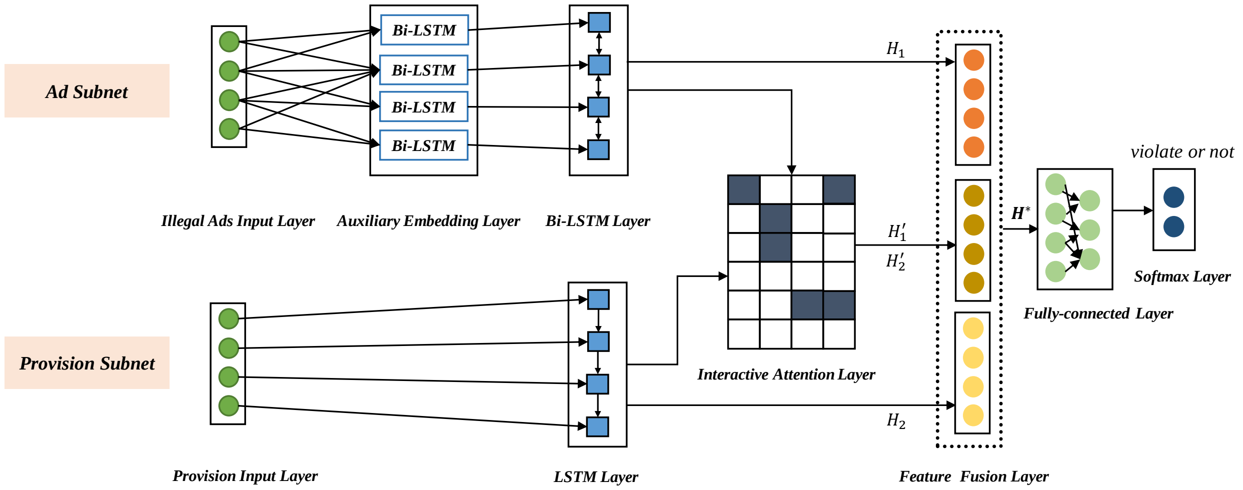
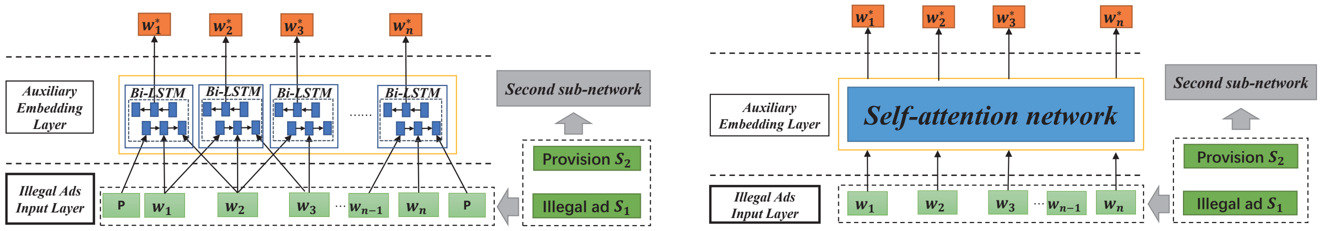


Figure 6. The overall architecture of the attention-based parallel LSTM network (IAD-Net).



(a) The D-BiLSTM as auxiliary embedding layer (k=1).

(b) Self-attention auxiliary embedding layer.

Figure 7. Different auxiliary embedding layer strategies. Subfigure (a) is the proposed D-BiLSTM strcture when window size is 1. The output hidden states can be regarded as strengthening the semantics of original words. Subfigure (b) is the adopted self-attention mechanism as auxiliary embedding layer

embedding layer with special structure compared to Provision Subnet, which is proved to achieve better result as shown in experimental results. Besides, we resort to Bi-LSTM and utilize semantic information both from the left and the right to increase the semantic richness of short ads. Provision Subnet takes sequential data of provision as input. For the limited quantity and semantic of provision, we use LSTM in Provision Subnet instead and remove the auxiliary embedding layer to prevent over-fitting. The encoded information of two sub-networks will be fused by the proposed interactive attention mechanism afterwards and generate the distributed representations H^* in feature fusion layer for classification. A fully-connected connection layer and softmax layer are used to predict labels$\in\{0,1\}$, where *0* means that input ad does not violate the input provision and *1* means the opposite.

The parallel sub-networks structure converts the multi-label classification of illegal ads into several two-category classification of matching or not between ads and provisions, and improves the accuracy of the final classification.

4.2 Auxiliary Embedding Layer

Illegal ads have intractable linguistic characteristics, such as semantic sparsity, shortness, and large scale, that may impede most of the hidden information discovered from large scale data collections [20]. Hence, we propose the disconnected Bi-LSTM (*D-BiLSTM*), which is based on disconnected recurrent neural network structure [27], as an auxiliary embedding layer in Ad Subnet to mitigate semantic sparsity, enhancing the semantic richness and incorporate local information near the central into word representation vector as well as keeping long-term dependencies. Furthermore, we use different structures as auxiliary embedding layer for comparative analysis including DRNN [27] and self-attention mechanism [26].

The output w_i^* of auxiliary embedding layer, which can be regarded as an auxiliary embedding of the i-th word, is supposed to be related to the words around it, and that's exactly what our D-BiLSTM do. Our D-BiLSTM considers the consecutive k words after the current word, which means the representation vector of the current word are much more context-aware and will be fine-tuned according to phrase in window size k . The representation vector calculated as formula (3) and Figure 7(a) illustrates the architecture of D-BiLSTM as auxiliary embedding layer.

$$\overrightarrow{w}_i^* = LSTM(w_{i-k}, w_{i-k+1}, ..., w_i) \tag{1}$$

$$\overleftarrow{w}_i^* = LSTM(w_{i+k}, w_{i+k-1}, ..., w_i) \tag{2}$$

$$w_i^* = [\overrightarrow{w}_i^* \oplus \overleftarrow{w}_i^*] \tag{3}$$

here $\oplus$ represents an element-wise sum symbol operation, k is the window size and the input length of Bi-LSTM unit is $2k+1$. We pad zero operation vectors in case words around the current word have not covered k windows.

The self-attention mechanism ignores the distance among words to capture the long-term dependencies, which means the outputs of self-attention module can be regarded as context-aware representation vectors. We leverage self-attention as an alternative to auxiliary embedding layer for its context-aware characteristic, moreover, it will save time on choosing hyper-parameter carefully. Figure 7(b) illustrates the structure of self-attention auxiliary embedding layer.

4.3 Interactive Attention Mechanism

The dependencies over long distance in Bi-LSTM will be impaired by fixing width of hidden states [24, 33], besides, the encoded information of ads and provisions need to interact in a rational way. Therefore, we propose the interactive attention mechanism in parallel subnetworks to alleviate this disadvantage. It can also locate the illegal expressions of illegal ads and illegal elements of violated provisions for visual interpretation of classification. Specifically, we have two encoded sentences $h^1 = (h_1^1, ..., h_n^1)$ and $h^2 = (h_1^2, ..., h_m^2)$, where h^1 is the ad and h^2 is the provision. Let $h \in \mathbb{R}^{H \times N}$ be a matrix consisting of Bi-LSTM's hidden states h_i^1,$i \in n$. Here H is the size of hidden states and N is the length of input sequence. The distributed representation H_1 of Ad Subnet is calculated as follows:

$$\begin{aligned} M &= \mathit{tanh}(h) & (4)\\ \alpha &= \mathit{softmax}(w^T M) & (5)\\ H &= \mathit{tanh}(h\alpha^T) & (6)\\ h_n^* &= [\overrightarrow{h}_n \oplus \overleftarrow{h}_n] & (7)\\ H_1 &= [H; h_n^*] & (8) \end{aligned}$$

where $M \in \mathbb{R}^{H \times N}$, $\alpha \in \mathbb{R}^N$. w, H and $h_n^* \in \mathbb{R}^H$, w^T is a transpose. $H_1 \in \mathbb{R}^{2H}$ is concatenated by H and h_n^*. We use the final states $\overrightarrow{h}_n$ and $\overleftarrow{h}_n$ of Bi-LSTM to calculate h_n^* and concatenate h_n^* to distributed representation H_1, which is inspired by [29] which is proved to work better in practice.

$h_i^1 \in \mathbb{R}^H$ is the hidden state of BiLSTM layer in Ad Subnet and is calculated as formula (3). $h_j^2 \in \mathbb{R}^H$, $j \in m$ is the hidden state of LSTM layer in Provision Subnet. The interactive attention weights e_{ij} is calculated as:

$$e_{ij} = {h_i^1}^T h_j^2 \tag{9}$$

gr *i* and *j* are the index of words in ads and provision respectively. In order to obtain more detailed local relavance on an ad and provisions, we use attention weights e_{ij} to calculate the interactive distributed representations $H_1^{'}$ and $H_2^{'}$ as follows [6]:

$$\tilde{h}_i^1 = \sum_{j=1}^{m} \frac{\exp(e_{ij})}{\sum_{k=1}^{m} \exp(e_{ik})} h_j^2, \forall i \in [1, ..., n] \tag{10}$$

$$\tilde{h}_j^2 = \sum_{i=1}^{n} \frac{\exp(e_{ij})}{\sum_{k=1}^{n} \exp(e_{kj})} h_i^1, \forall j \in [1, ..., m] \tag{11}$$

According to formula (10), the $\tilde{h}_i^1$ represents the weighted summation of h_j^2, which means the contents in provisions that are relevant to the ads will be captured automatically and represented as $\tilde{h}_i^1$. The same is feasible for formula (11).

$$\begin{aligned} H_1^{'} &= [h^1; \tilde{h}^1; h^1 - \tilde{h}^1; h^1 \odot \tilde{h}^1] & (12)\\ H_2^{'} &= [h^2; \tilde{h}^2; h^2 - \tilde{h}^2; h^2 \odot \tilde{h}^2] & (13) \end{aligned}$$

We compute the difference and element-wise multiplication between $\langle h^1, \tilde{h}^1 \rangle$ and $\langle h^2, \tilde{h}^2 \rangle$, which will help sharpen local inference information between an ad and provision [6, 18]. Symbol $\odot$ stands for element-wise multiplication.

4.4 Classification Network

The fully-connected layer takes H^* as input for final classification:

$$H^* = \left[H_1, H_2, H_1^{'}, H_2^{'}\right] \tag{14}$$

Then, the *softmax* layer is followed to predict the label $\hat{y}$ from $\{\mathit{0,1}\}$ for input sequences (S_1, S_2) and takes the outputs of fully-connected layer W^* as input:

$$\begin{aligned} \hat{p}(y|(S_1, S_2)) &= \mathit{softmax}(W_s W^* + b_s) & (15)\\ \hat{y} &= \mathit{arg}\,\max_{y} \hat{p}(y|(S_1, S_2)) & (16) \end{aligned}$$

The loss function $J(\theta)$ to be minimized is the cross-entropy error with L2 regularization.

$$J(\theta) = -\sum_i \sum_j y_i^j \log \hat{p}_i^j + \lambda \parallel \theta \|^2 \tag{17}$$

where i is the index of input sequences (S_1, S_2) and j is the index of true label. $y_i^j \in \mathbb{R}^2$ and $\hat{p}_i^j \in \mathbb{R}^2$ are the ground truth representation and the estimated probability respectively. λ is the hyper-parameter of L2 regularization to control degree of alleviating over-fitting. We adopt the Adam optimizer for training phrase [15].

5 EXPERIMENTAL RESULTS AND ANALYSES

In this section, we present our evaluation results and compare with a number baselines including the state-of-the-art models. Moreover, we give some qualitative analyses of our model in detail.

5.1 Experimental Settings

In our experiments, we pre-train word embeddings in our dataset by word2vec [17]. The dimension of word embeddings is 300. Other hyper-parameters are as follows: mini-batch size is 256, dropout rate [9] is 0.5 in both sub-networks, vocab size in Ad Subnet is 55 and 30 in Provision Subnet, the learning rate is initialized as 0.05 with exponential decay. In addition, we apply an early-stop strategy [3] with 100 epochs. We partition the dataset into three non-overlapped subsets including training set, validation set and test set with a ratio of 4:1:1.

5.2 Comparison with Baselines

We use accuracy and Micro-F1 score as evaluation metrics, and compare our IAD-Net with several baseline models. As Table 3 shows, the SVM+BTM model is one of the state-of-the-art models for short text topic modeling and Seed LDA model shows great competitiveness on this issue. The results are presented in Table 3.

Table 3 and Figure 8 demonstrate that our IAD-Net outperforms all the baseline models in most provisions. Moreover, we can observe that: (1) The shallow models have relatively poor performance due to a lack of mechanism to process sequential data and can't capture the location information of the word pairs. (2) Models based on semantic analysis such as Seed LDA pays more attention to the inner semantic structure of the text to obtain a more expressive logic structure,

Table 3. Comparative results of accuracy and Micro-F1 scores with baseline models on IAD dataset. Our IAD-Net outperforms all the baseline models and has a 4% improvement beyond the best baseline model.

Model	Acc.	F1
Naive Bayes [38]	0.683	0.683
Logistic Regression [7]	0.745	0.745
Random Forest [31]	0.751	0.751
Seed-guided LDA [12]	0.846	0.846
SVM+BTM [32]	0.859	0.859
LSTM [16]	0.791	0.791
RNN+Attention [19]	0.823	0.823
TextCNN [14]	0.853	0.853
Attention-BiLSTM [37]	0.867	0.867
CNN+TEWE [21]	0.875	0.875
IAD-Net (ours)	**0.915**	**0.915**

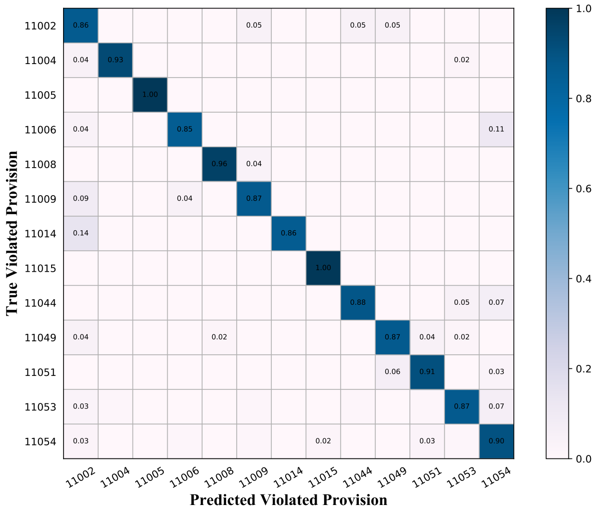


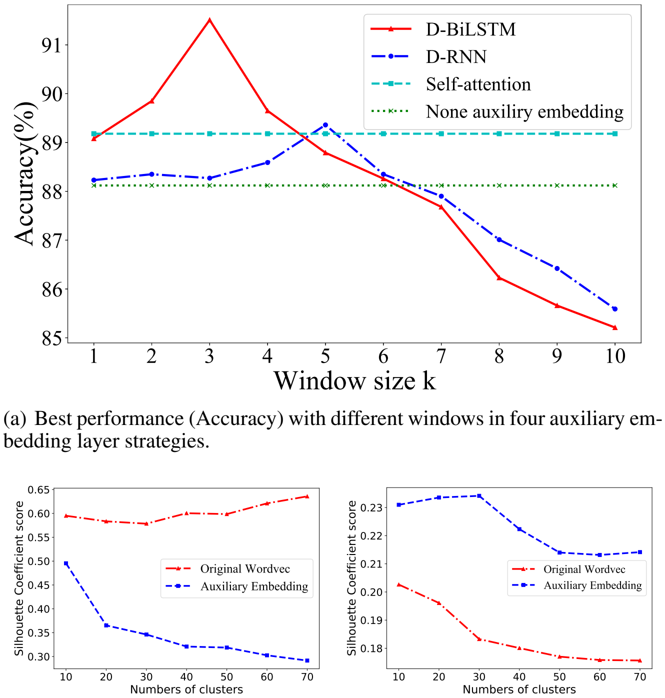
Figure 8. The confusion matrix of IAD-Net.

which contributes to a better perception of data distribution. (3) The neural network-based models achieve better performance by considering the sequential information, especially those with attention mechanism. (4) All baselines, especially the AttBiLSTM, have not utilized the provision text containing crucial information, and thus achieves worse accuracy, which demonstrates the parallel subnet Two processing provision text does have a positive effect on the classification task. (5) Learning the long-term dependencies across sequences, capturing the relations of word pairs with attention mechanism and joint learning of illegal ads and provisions are critical aspects.

5.3 Analysis on Auxiliary Embedding Layer

We conduct experiments to investigate the influence of different auxiliary embedding layer strategies including DRNN, D-BiLSTM, self-attention and no auxiliary embedding layer. Besides, We analyze the input and output distribution of the BiLSTM layer in AD Subnet to study the impact of the auxiliary embedding layer.

Figure 9(a) demonstrates that the D-BiLSTM outperforms the others. We can observe that: (1) The optimal window size of D-BiLTSM is 3 and it actually takes 7 words as a unit input sequence. Compared to the optimal window size 5 of DRNN, D-BiLSTM only uses one more word but has an improvement of 2.25% in accuracy, which proves that consecutive information from the backward direction of D-BiLSTM improves accuracy effectively. (2) As window size increases after optimal k, D-BiLSTM shows a slightly faster downward trend than DRNN, which may be caused by the words actually used to grow faster in D-BiLSTM and the average length of ads is limited, resulting in increasing semantic interference from padded zero vectors. (3) While the self-attention layer has the lowest accuracy, there is still a 1% improvement compared to no auxiliary embedding layer, and no requirement for empirical parameter adjustment makes it achieve better generalization performance.



(a) Best performance (Accuracy) with different windows in four auxiliary embedding layer strategies.

(b) Clustering result of the input of Bi-LSTM layer.

(c) Clustering result of the output of Bi-LSTM layer.

Figure 9. Analysis on auxiliary embedding layer. Subfigure (a) is the comparasion of different auxiliary embedding layer. Subfigures (b) and (c) are the clustering analysis of auxiliary embedding layer. PCA [30] is applied to reduce dimension before clustering.

Words with similar semantics tend to be clustered into the same category and hidden states in auxiliary embedding layer and Bi-LSTM layer also have clustering tendencies [10]. To further explore the impact of auxiliary embedding layer, we use K-means [8] to cluster the input and output of Bi-LSTM layer with or without auxiliary embedding layer (D-BiLSTM and parameters are set optimally) and use silhouette coefficient to evaluate the clustering effect. As Figure 9(b) shows, the clustering effect is worse when there exists auxiliary embedding layer, which indicates the distribution of auxiliary embeddings is more scattered and auxiliary embedding layer helps to enhance semantic richness (w_i^*). Figure 9(c) shows that the auxiliary embedding layer improves the aggregation of hidden states and helps Bi-LSTM layer to extract more generalized and stable features for final prediction (h_i^*).

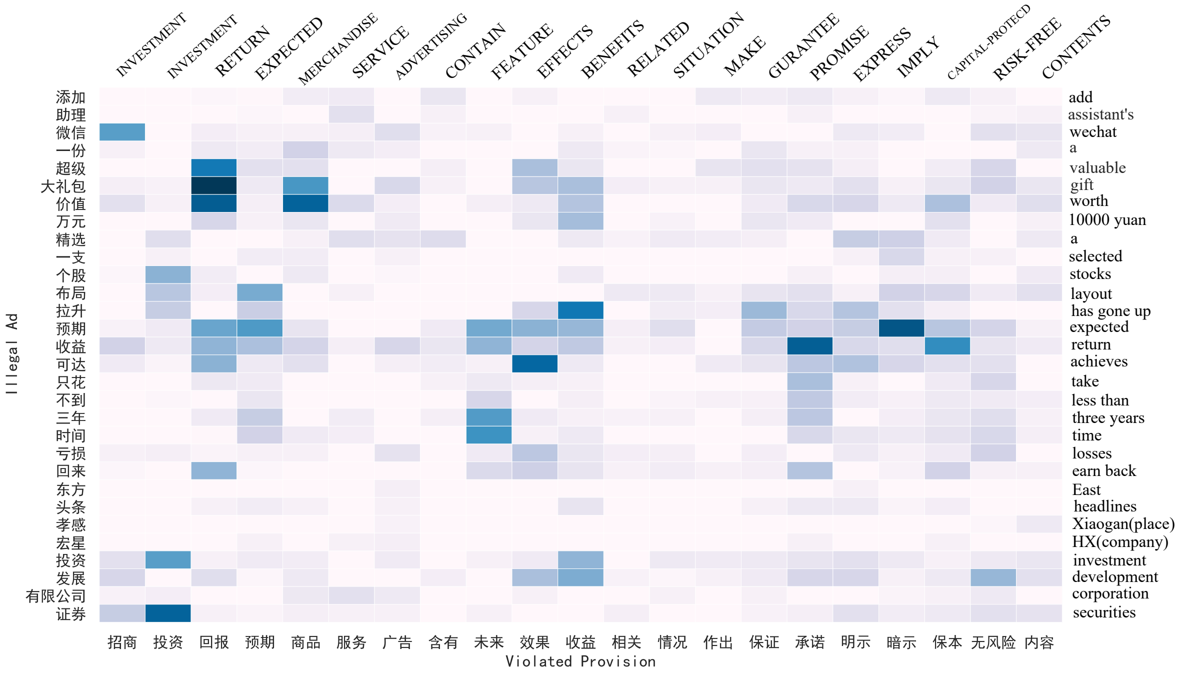
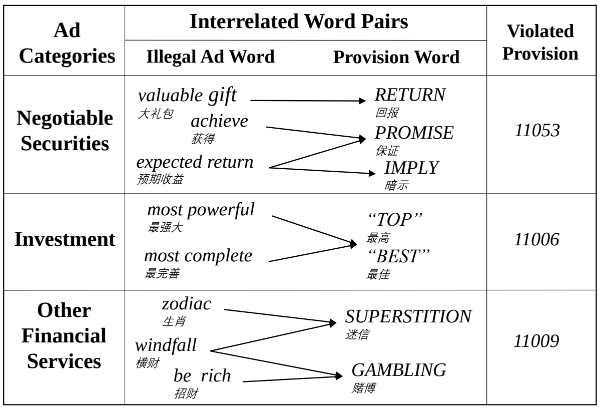


Figure 10. Attention visualization for determining the illegal expressions and illegal elements.



Ad Categories	Interrelated Word Pairs		Violated Provision
	Illegal Ad Word	Provision Word	
Negotiable Securities	*valuable gift* 大礼包 *achieve* 获得 *expected return* 预期收益	*RETURN* 回报 *PROMISE* 保证 *IMPLY* 暗示	*11053*
Investment	*most powerful* 最强大 *most complete* 最完善	*“TOP”* 最高 *“BEST”* 最佳	*11006*
Other Financial Services	*zodiac* 生肖 *windfall* 横财 *be rich* 招财	*SUPERSTITION* 迷信 *GAMBLING* 赌博	*11009*

Figure 11. Key word pairs between illegal ads and provisions illustrated by the intention weight.

5.4 Analysis on Interactive Attention Mechanism

Interactive attention mechanism in IAD-Net is used to capture important semantic information in both sub-networks and build implicit mapping between illegal ads and provisions. To further verify the effectiveness of IAD-Net, we explore the interpretability and usability of IAD-Net by analyzing the attention weights.

We visualize the weights of interactive attention layer and Figure 10 illustrates how the interactive attention mechanism reflects the relevance of word pairs between illegal ads and provisions. The darker color indicates that the corresponding word is more indicative. The illegal ad in Figure 10 was classified correctly by IAD-Net. In general, there are two subsequences with darker color in illegal ad. The former is “*valuable gift worth more than 10000 yuan*”[5] which is related to “*RETURN*” and “*BENEFIT*” in provision and could be induced for consumers. The latter is “*expected return is up to -, earn back losses in less than three years*” which violated the “*CAPITAL PRESERVATION*” and “*RISK-FREE*” directly. In detail, further analysis finds that interactive attention can establish holistic semantic association from the partial via the interrelated word pairs, so as to help IAD-Net make the final decision. For example, ⟨*gift*,*RETURN*⟩ helps determine the induced content, besides, word pairs like ⟨*achieve*,*PROMISE*⟩, ⟨*achieve*,*FUTURE BENEFITS*⟩, ⟨*three year*,*FUTURE*⟩ and ⟨*expected return*,*IMPLY*⟩ with large weights locally indicate the key words that lead to illegality and its corresponding violated content in provision correctly. Figure 11 demonstrates more interrelated word pairs of different illegal ads and provisions determined by the interactive attention mechanism. The interactive attention mechanism builds implict mapping between ads and provisions, and then determine the consistency of ads and provisions to find out the violated provision precisely.

6 CONCLUSION

In this paper, we have constructed and released a large-scale dataset called IAD for the classification of violated provision of the illegal ads on Internet. We further proposed a novel model called IAD-Net for illegal ads classification, which consists of two parallel subnetworks to process illegal ads and provisions simultaneously, an auxiliary embedding layer and an interactive attention mechanism to alleviate the issues caused by legal-related advertising text and enhance the interpretability of IAD-Net. Experimental study on IAD-Net based on IAD dataset demonstrates the effectiveness of the proposed IAD-Net over previous methods. We hope our proposed IAD dataset and IAD-Net will boost the research in this area.

[5] The lowercase\uppercase italic text represents content in ad\provision respectively. The same for the word pair ⟨-,-⟩ below.

ACKNOWLEDGEMENTS

This work was supported by the Fundamental Research Funds for the Central Universities and National Program for Special Support of Top-Notch Young Professionals.

REFERENCES

[1] David M Blei, Andrew Y Ng, and Michael I Jordan, 'Latent dirichlet allocation', *Journal of machine Learning research*, **3**(Jan), 993–1022, (2003).

[2] Juan Manuel Cabrera, Hugo Jair Escalante, and Manuel Montes-y Gómez, 'Distributional term representations for short-text categorization', in *International Conference on Intelligent Text Processing and Computational Linguistics*, pp. 335–346. Springer, (2013).

[3] Rich Caruana, Steve Lawrence, and C Lee Giles, 'Overfitting in neural nets: Backpropagation, conjugate gradient, and early stopping', in *Advances in neural information processing systems*, pp. 402–408, (2001).

[4] Peggy E Chaudhry, 'The looming shadow of illicit trade on the internet', *Business Horizons*, **60**(1), 77–89, (2017).

[5] Jiming Chen, Kang Hu, Qi Wang, Yuyi Sun, Zhiguo Shi, and Shibo He, 'Narrowband internet of things: Implementations and applications', *IEEE Internet of Things Journal*, **4**(6), 2309–2314, (2017).

[6] Qian Chen, Xiaodan Zhu, Zhenhua Ling, Si Wei, Hui Jiang, and Diana Inkpen, 'Enhanced lstm for natural language inference', *arXiv preprint arXiv:1609.06038*, (2016).

[7] Weiwei Cheng and Eyke Hüllermeier, 'Combining instance-based learning and logistic regression for multilabel classification', *Machine Learning*, **76**(2-3), 211–225, (2009).

[8] John A Hartigan and Manchek A Wong, 'Algorithm as 136: A k-means clustering algorithm', *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, **28**(1), 100–108, (1979).

[9] Geoffrey E Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R Salakhutdinov, 'Improving neural networks by preventing co-adaptation of feature detectors', *arXiv preprint arXiv:1207.0580*, (2012).

[10] Bo-Jian Hou and Zhi-Hua Zhou, 'Learning with interpretable structure from rnn', *arXiv preprint arXiv:1810.10708*, (2018).

[11] Cisco Visual Networking Index, 'Forecast and methodology, 2015–2020', *White paper*, 1–41, (2016).

[12] Jagadeesh Jagarlamudi, Hal Daumé III, and Raghavendra Udupa, 'Incorporating lexical priors into topic models', in *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 204–213. Association for Computational Linguistics, (2012).

[13] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov, 'Bag of tricks for efficient text classification', *arXiv preprint arXiv:1607.01759*, (2016).

[14] Yoon Kim, 'Convolutional neural networks for sentence classification', *arXiv preprint arXiv:1408.5882*, (2014).

[15] Diederik P Kingma and Jimmy Ba, 'Adam: A method for stochastic optimization', *arXiv preprint arXiv:1412.6980*, (2014).

[16] Ji Young Lee and Franck Dernoncourt, 'Sequential short-text classification with recurrent and convolutional neural networks', *arXiv preprint arXiv:1603.03827*, (2016).

[17] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean, 'Distributed representations of words and phrases and their compositionality', in *Advances in neural information processing systems*, pp. 3111–3119, (2013).

[18] Lili Mou, Rui Men, Ge Li, Yan Xu, Lu Zhang, Rui Yan, and Zhi Jin, 'Natural language inference by tree-based convolution and heuristic matching', *arXiv preprint arXiv:1512.08422*, (2015).

[19] Daniel Ortega and Ngoc Thang Vu, 'Neural-based context representation learning for dialog act classification', *arXiv preprint arXiv:1708.02561*, (2017).

[20] Xuan-Hieu Phan, Le-Minh Nguyen, and Susumu Horiguchi, 'Learning to classify short and sparse text & web with hidden topics from large-scale data collections', in *Proceedings of the 17th international conference on World Wide Web*, pp. 91–100. ACM, (2008).

[21] Yafeng Ren, Yue Zhang, Meishan Zhang, and Donghong Ji, 'Improving twitter sentiment classification using topic-enriched multi-prototype word embeddings', in *Thirtieth AAAI conference on artificial intelligence*, (2016).

[22] Kazuya Shimura, Jiyi Li, and Fumiyo Fukumoto, 'Hft-cnn: Learning hierarchical category structure for multi-label short text categorization', in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 811–816, (2018).

[23] Dawei Song, Peter Bruza, Zi Huang, and Raymond YK Lau, 'Classifying document titles based on information inference', in *International Symposium on Methodologies for Intelligent Systems*, pp. 297–306. Springer, (2003).

[24] Ming Tan, Cicero dos Santos, Bing Xiang, and Bowen Zhou, 'Lstm-based deep learning models for non-factoid answer selection', *arXiv preprint arXiv:1511.04108*, (2015).

[25] Xu Tan, Yuanchao Shu, Xie Lu, Peng Cheng, and Jiming Chen, 'Characterizing and modeling package dynamics in express shipping service network', in *2014 IEEE International Congress on Big Data*, pp. 144–151. IEEE, (2014).

[26] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin, 'Attention is all you need', in *Advances in neural information processing systems*, pp. 5998–6008, (2017).

[27] Baoxin Wang, 'Disconnected recurrent neural networks for text categorization', in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2311–2320, (2018).

[28] Shaonan Wang, Jiajun Zhang, Chengqing Zong, et al., 'Exploiting word internal structures for generic chinese sentence representation', (2017).

[29] Yequan Wang, Minlie Huang, Li Zhao, et al., 'Attention-based lstm for aspect-level sentiment classification', in *Proceedings of the 2016 conference on empirical methods in natural language processing*, pp. 606–615, (2016).

[30] Svante Wold, Kim Esbensen, and Paul Geladi, 'Principal component analysis', *Chemometrics and intelligent laboratory systems*, **2**(1-3), 37–52, (1987).

[31] Baoxun Xu, Xiufeng Guo, Yunming Ye, and Jiefeng Cheng, 'An improved random forest classifier for text categorization.', *JCP*, **7**(12), 2913–2920, (2012).

[32] Xiaohui Yan, Jiafeng Guo, Yanyan Lan, and Xueqi Cheng, 'A biterm topic model for short texts', in *Proceedings of the 22nd international conference on World Wide Web*, pp. 1445–1456. ACM, (2013).

[33] Zidong Yang, Ji Hu, Yuanchao Shu, Peng Cheng, Jiming Chen, and Thomas Moscibroda, 'Mobility modeling and prediction in bike-sharing systems', in *Proceedings of the 14th Annual International Conference on Mobile Systems, Applications, and Services*, pp. 165–178, (2016).

[34] Sarah Zelikovitz and Haym Hirsh, 'Improving short text classification using unlabeled background knowledge to assess document similarity', in *Proceedings of the seventeenth international conference on machine learning*, volume 2000, pp. 1183–1190, (2000).

[35] Jichuan Zeng, Jing Li, Yan Song, Cuiyun Gao, Michael R Lyu, and Irwin King, 'Topic memory networks for short text classification', *arXiv preprint arXiv:1809.03664*, (2018).

[36] Daochen Zha and Chenliang Li, 'Multi-label dataless text classification with topic modeling', *Knowledge and Information Systems*, 1–24, (2017).

[37] Dongxu Zhang and Dong Wang, 'Relation classification via recurrent neural network', *arXiv preprint arXiv:1508.01006*, (2015).

[38] Wei Zhang and Feng Gao, 'An improvement to naive bayes for text classification', *Procedia Engineering*, **15**, 2160–2164, (2011).

[39] Peng Zhou, Wei Shi, Jun Tian, Zhenyu Qi, Bingchen Li, Hongwei Hao, and Bo Xu, 'Attention-based bidirectional long short-term memory networks for relation classification', in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pp. 207–212, (2016).