Generating Behavior Features for Cold-Start Spam Review Detection

Xiaoya Tang, Tieyun Qian(✉), and Zhenni You

School of Computer Science, Wuhan University, Wuhan, Hubei, China
{xiaoyatang,qty,znyou}@whu.edu.cn

Abstract. Existing studies on spam detection show that behavior features are effective in distinguishing spam and legitimate reviews. However, it usually takes a long time to collect such features and is hard to apply them to cold-start spam review detection tasks. In this paper, we exploit the generative adversarial network for addressing this problem. The key idea is to generate *synthetic behavior features* (SBFs) for new users from their *easily accessible features* (EAFs). We conduct extensive experiments on two Yelp datasets. Experimental results demonstrate that our proposed framework significantly outperforms the state-of-the-art methods.

Keywords: Spam review detection · Cold-start problem · Generative adversarial network

1 Introduction

The cold-start spam review detection problem, i.e., *a review is just posted by a new reviewer*, is critical in preventing the damage of spams in their early stage. Two embedding models [6,7] show much better performance than the traditional methods. Nevertheless, the inherent problem, i.e., *lack of effective behavior features for new users who post just one review*, remains unsolved.

To solve this problem, we generate the synthetic behavior features (SBFs) for new users who actually do not have such features. Specifically, we first extract six real behavior features (RBFs) for regular users and three types of easily accessible features (EAFs) which exist for both regular users and new user. Secondly, taking these EAFs as the input, we generate SBFs in the generator of the GAN and use the discriminator of the GAN to train the generator. The trained GAN is finally applied to new users to get SBFs which are actually not yet observed for these new users. We conduct extensive evaluations on two real world Yelp datasets. Results demonstrate that our model significantly outperforms the state-of-the-art baseline methods.

2 Real Behavior and Easily Accessible Features

Real Behavior Features (RBFs). We choose six types of real behavior features including activity window (AW) [3,4], maximum number of reviews

G. Li et al. (Eds.): DASFAA 2019, LNCS 11448, pp. 324–328, 2019.
https://doi.org/10.1007/978-3-030-18590-9_38

(MNR) [4,5], percentage of positive reviews (PR) [4,5], review count (RC) [3,4], reviewer deviation (RD) [1,4], and maximum content similarity (MCS) [1,5].

Easily Accessible Features (EAFs). We choose three types of features which can easily accessible for both regular and new reviewers including text features (TF), rating features (RF), and attribute features (AF). All these EAFs are converted into 100-dimension embeddings. We using the convolutional neural network (CNN) like those in [6,7] to get TF from the review text. Moreover, we discretize the deviation of rating score of this review from the average ratings on the same product and the deviation of registered timestamp from posting timestamp to get RF and AF. If the value falls into an interval, the corresponding dimension is set to 1 and other dimensions are set to 0.

3 Our Proposed Model

We aim to exploit generative adversarial network (GAN) [2] to generate the behavior features which are effective in spam detection but new users lack such information in the cold-start scenario. The architecture of our behavior feature generating (bfGAN) model is shown in Fig. 1. The left part in Fig. 1 is the generator which is used to generate SBFs from the input EAFs. The right part is the discriminator which will make a discrimination between SBFs and RBFs using a classifier to guide the training.

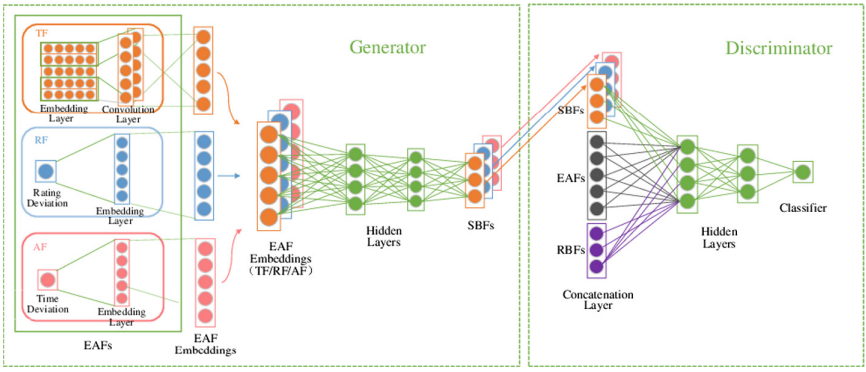

Fig. 1. Architecture of our bfGAN model

Generator. The generator contains six layers. The first three layers are used to do normalization and get EAFs. We then use three non-linear hidden layers to transform EAFs into SBFs.

The generator includes a task loss $\mathcal{L}_t$ to misjudge the discriminator and a closeness loss $\mathcal{L}_c$ to encourage the generator to generate SBFs that have similar

distribution as RBFs. We use the cross entropy loss to define the overall loss for the generator as:

$$\mathcal{L}_G = \mathcal{L}_t + \mathcal{L}_c = \min_{\Theta_G} J(D(EAF_+ \oplus SBF), 1) + J(RBF, SBF), \tag{1}$$

where D is the discriminative function activated by *tanh*, EAF_+ is the positive EAF matching RBF or SBF and $\oplus$ denotes the concatenation of two embeddings.

Discriminator. The discriminator should judge the (EAF+, RBF) pairs from the realistic training data as real and the (EAF+, SBF) pairs from the generator as fake. Hence we define two loss functions $J(D(EAF_+ \oplus RBF), 1)$ and $J(D(EAF_+ \oplus SBF), 0)$ for this purpose.

Another source of error in the discriminator may come from the unrealistic behavior features. To separate two sources of error, we add a third type of input consisting of RBFs with mismatched EAFs, which the discriminator should learn to score as fake. We formally define a loss function $J(D(RBF \oplus EAF_-), 0)$ to achieve this. The overall function loss $\mathcal{L}_D$ for the discriminator is as follows.

$$\begin{aligned} \mathcal{L}_D = \min_{\Theta_D} \; & J(D(EAF_+ \oplus RBF), 1) + \frac{1}{2} J(D(EAF_+ \oplus SBF), 0) \\ & + \frac{1}{2} J(D(RBF \oplus EAF_-), 0), \end{aligned} \tag{2}$$

4 Experiments

Experimental Settings. We verify the effectiveness of our proposed model on Hotel and Restaurant datasets of Yelp. We use Adam algorithm to train GAN. The learning rate is 0.00001. The number of iterations for TF, RF, and AF is 300, 21, 19 on Hotel, and 13, 67, 13 on Restaurant, respectively. We stop iteration when the network becomes stable or the loss in the generator becomes the lowest. In the generator network, the number of neurons in two hidden layers is set to 64 and 32, and tanh is adopted as activation function. The last layer is the mapping layer for generating SBFs with 6 neurons. In the discriminator network, the number of neurons in two hidden layers is the same with the generator's. The final classification layer has 1 neuron and uses sigmoid as activation function.

Comparison with Baselines. The comparison results between our model and nine state-of-the-art methods are shown in Table 1.

It is clear that our proposed bfGAN achieves the best performance in terms of F1 and Accuracy on both Hotel and Restaurant datasets. Our bfGAN model combines the traditional approaches in finding effective real behavior features and the recent advances in deep learning to generate synthetic behavior features to simulate the real ones.

Compared to the state-of-the-art method AEDA, our model uses less information and does not integrate domain adaption into the framework, but it reaches

Table 1. Comparison with baselines.

Features	Hotel				Restaurant			
	P	R	F1	Acc	P	R	F1	Acc
LF [3]	54.5	71.1	61.7	55.9	53.8	80.8	64.6	55.8
Supervised-CNN [6]	61.2	51.7	56.1	59.5	56.9	58.8	57.8	57.1
LF+BF [3]	63.4	52.6	57.5	61.1	58.1	61.2	59.6	58.5
BF_EditSim+LF [6]	55.3	69.7	61.6	56.6	53.9	82.2	65.1	56.0
BF_W2Vsim+W2V [6]	58.4	65.9	61.9	59.5	56.3	73.4	63.7	58.2
RE* [6]	62.1	68.3	65.1	63.3	58.4	75.1	65.7	60.8
RE+RRE+PRE* [6]	63.6	71.2	67.2	65.3	59.0	78.8	67.5	62.0
AE [7]	76.7	74.2	75.4	75.8	80.3	66.2	72.6	75.0
AEDA [7]	83.9	74.2	78.7	80.0	82.4	65.1	72.8	75.6
bfGAN(ours)	81.2	85.7	**83.4**	**83.0**	76.7	73.4	**75.1**	**75.7**

*Denotes that the model is trained on the labeled data and a large number of unlabeled data.

an improvement of 6.0% and 3.2% of F1 over AEDA on Hotel and Restaurant, respectively. This clearly demonstrates that our bfGAN model can generate highly effective SBFs whose distribution is close to that of RBFs, and hence it achieves significantly better performance than baseline methods.

5 Conclusion

In this paper, we propose a novel bfGAN model for cold-start spam review detection. To address the problem of lacking effective real behavior features (RBFs) for new users in cold-start scenario, we design a GAN framework to generate synthetic behavior features (SBFs) using EAFs and further present a new implementation of GAN by incorporating an extra loss to explicitly guide SBFs to be close to RBFs. We conduct extensive experiments on two Yelp datasets. Results demonstrate that our bfGAN model significantly outperforms the state-of-the-art baselines.

Acknowledgments. The work described in this paper has been supported in part by the NSFC projects (61572376).

References

1. Fei, G., Mukherjee, A., Liu, B., Hsu, M., Castellanos, M., Ghosh, R.: Exploiting burstiness in reviews for review spammer detection. In: ICWSM, pp. 175–184 (2013)
2. Goodfellow, I., et al.: Generative adversarial nets. In: NIPS, pp. 2672–2680 (2014)
3. Mukherjee, A., Venkataraman, V., Liu, B., Glance, N.: Fake review detection: classification and analysis of real and pseudo reviews. Technical Report UIC-CS-2013-03, University of Illinois at Chicago, Technical report (2013)

4. Mukherjee, A., Venkataraman, V., Liu, B., Glance, N.S.: What yelp fake review filter might be doing? In: ICWSM (2013)
5. Rayana, S., Akoglu, L.: Collective opinion spam detection: bridging review networks and metadata. In: KDD, pp. 985–994 (2015)
6. Wang, X., Liu, K., Zhao, J.: Handling cold-start problem in review spam detection by jointly embedding texts and behaviors. In: ACL, pp. 366–376 (2017)
7. You, Z., Qian, T., Liu, B.: An attribute enhanced domain adaptive model for cold-start spam review detection. In: COLING, pp. 1884–1895 (2018)