Active Learning for Spam Email Classification

Zheng Chen†
School of Information & Software Engineering, University of Electronic Science & Technology of China
zchen@uestc.edu.cn

Ruiwen Tao
School of Information & Software Engineering, University of Electronic Science & Technology of China
Taorw@std.uestc.edu.cn

Xiaoyang Wu†
National Key Laboratory of Science and Technology on Blind Signal Processing
Chengdu Sichuan China
tisun1017@126.com

Zhimin Wei
National Key Laboratory of Science and Technology on Blind Signal Processing
Chengdu Sichuan China
13882256850@163.com

Xiao Luo
National Key Laboratory of Science and Technology on Blind Signal Processing
Chengdu Sichuan China
ivy_xx@163.com

ABSTRACT

Deep learning has yielded state-of-the-art performance on text classification tasks. In this paper, a new neural network based on Long-Short-Term-Memory model is applied to classify spam emails. Using deep learning method to classify spam emails requires large amounts of labeled data. To solve this problem, active learning method is used to reduce labeling cost and increase model adaptability. In this paper, it is found that the new model performs better than standard CNNs and RNNs on email classification task, and active learning methods can match state-of-the-art performance with just 10% of the labeled data.

CCS CONCEPTS

Information systems → Information retrieval → Retrieval tasks and goals → Clustering and classification

KEYWORDS

Spam email classification, active learning, LSTM

ACM Reference format:
Zheng Chen, Ruiwen Tao, Xiaoyang Wu, Zhimin Wei and Xiao Luo. 2019. Active Learning for Spam Email Classification. In *Proceedings of 2019 2nd International Conference on Algorithms, Computing and Artificial Intelligence (ACAI'19). Sanya, China, 5 pages.* https://doi.org/10.1145/3377713.3377789

†Corresponding author: zchen@uestc.edu.cn, tisun1017@126.com

ACAI '19, December 20–22, 2019, Sanya, China

ACM ISBN 978-1-4503-7261-9/19/12…$15.00
https://doi.org/10.1145/3377713.3377789

1 Introduction

Deeping learning methods applied to the task of text classification have successively advanced the state-of-the-art. Text CNN [1] applies conversational neural network to text recognition and achieves 93.6% of the results on TREC Question Answering Dataset[1]. And Text RNN [2] uses recurrent neural network to capture context information in text, which improves the context-free situation in CNN. In the task of spam email classification, a new model we proposed will be used to extract feature information of both email content and subject. Extracting both feature information respectively and contacting them together can get a better advantage than using only the email content.

Meanwhile, the size of the training samples has a crucial impact on the training results of those neural network models (the larger the data volume, the more obvious the improvement of the model effect). And the advantage of deep learning diminished when dataset is small.

It is especially expensive and tough to annotate a large amount of unlabeled data. But training a model needs substantial amounts of labeled data. In such a situation, active learning provided a promising approach to efficiently select samples. Active learning methods can accelerate the training process of the neural network and reduces the cost of manual labeling.

Active learning is a common method for text recognition, and its essence is a kind of selection way, using classifier and filter to pick unlabeled samples for training. The most commonly used approach is Uncertain Sampling [3], in which the model selects examples for which current predication is least confidence. Other approaches like Query Committee [4] and Density-Weighted [5], use different criterion to measure the representation of the example. To use classifier to filter examples, it is assumed that the

[1] https://trec.nist.gov/data/qamain

classifier has absolute accuracy in the process of filtering, which demands that the classifier has excellent performance on the training dataset.

In this paper, a new model is designed to classify spam email according to the characteristics of the email, and active learning methods will be practiced based on the new model. The email classifier will be trained on a labeled dataset (training dataset). After the training, active learning method will be used to select a certain number of emails from the labeled dataset (training dataset) and the unlabeled dataset (test dataset) to form a new dataset. Finally, the new dataset will be used to train the classifier. A small number of samples will be selected for each screening, so the training speed of the classifier is very fast. The unlabeled dataset (test dataset) is chosen to adapt the model to a new data distribution, and the training dataset is chosen to keep the model from forgetting the data distribution of the previous samples.

In summary, in the task of spam email classification, the deep neural network model is valued as the email classifier, and the pool-based active learning method is used as the email filter, which realizes the rapid migration training of the new model in the case of substantial amounts of email samples.

2 Related Work

Email categorization is a classic text recognition problem. Traditional machine learning methods such as Naive Bayes [6] and SVM[7] [8] are widely applied in various text recognition tasks. The wide appliance of word representation also has advanced the development of text classification in deep learning. Researchers have achieved good results in text categorization by using Neural Network Language Model and CNN, RNN methods. In this paper, a new architecture of model is provided to classify the email. Compared with the classical RNN, CNN text classification model, the new model has improved the effect about nearly 1.5%.

Traditional active learning methods, such as Lest Confidence [9], Max Margin [10], select the most inaccurate samples. Other active learning methods, such as Density-Weighted method, add representative measures on the base of uncertain sampling. These methods [11] work well on some tasks [12] [13], but not on others [13] [14]. Moreover, the Density-Weighted method is not fully applicable to deep learning tasks. How to combine deep learning task with active learning algorithm and achieve effective results is the future work. In this paper, some traditional active learning algorithms will be practiced on the new spam email classification model and their effectiveness will be tested.

3 Our Model

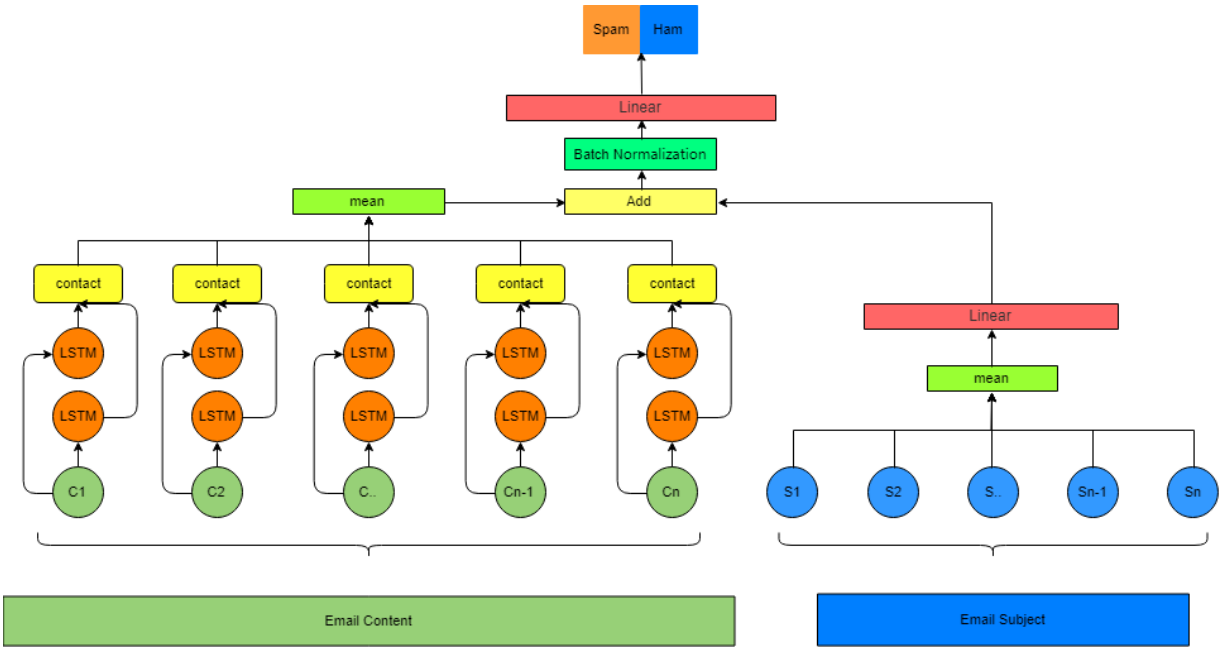

Figure 1. LSTM MC model for spam email classification

3.1 Model Description

To meet the needs of the spam email classification task, a model is designed based on LSTM [16]. Each email has content information and subject information, which are modeled separately by the new model. A character-size-based encoder is

adopted in the model to embedding Chinese emails, which reduces the amount of embedding in the embedding layer and makes the vector information more compact.

Before embedding text information, the length of text padding has to be determined. The padding length of content we set is 1000. Similarly, the subject of each email is filled or intercepted, and the padding length is the maximum of the subject length of all email samples, which is 25.

Embedding Layer: The built-in embedding layer is used for the character-level encoder, without using the pre-trained word representation for training. Because of the large amount of sample data in the task of spam email classification, the model with a higher fitting degree can be obtained by using built-in word embedding layer. But using built-in embedding layer can lead to the over-fitting problem, so dropout and batch normalization [16] methods are used to preventing over-fitting in the model.

Feature extraction layer: The goal of feature extraction layer is to extract the word embeddings of the email into a vector. For different parts of the email, different methods can be applied to extract it. For the content of the email, two-way LSTM [1] is used to extract features. For the subject of the email, we use pooling and linear layer to extract features, so that the subject of the email can be better mapped to the content of the email. Email content uses LSTM to extract features and subject uses linear layer to extract features.

In email content, the vector of character i is C_i, the length of the content padding is a fixed value C_L, and the embedding dim of each character in content is E. Positive-order text is embedded as C_E^s and reverse-order text is embedded as C_E^r. Input them into different LSTM networks respectively. The outputs of two LSTM are C_H^s and C_H^r, and H is the number of the hidden units in LSTM cell. Combine C_H^s , C_H^r and C_E^s together, we will get the text output C_o . Finally, averaging the output C_o , the content representation V_c is obtained. The shape of V_c is [(H+E+H), 1]. V_c is the feature vector of email content extracted by the content feature extraction layer of the model.

In the subject of the email, the vector of character i is S_i, the length of subject padding is a fixed value S_L, and the embedding length of each character in subject is E. The word embedding of subject represents as S_E, and then averaging S_E to obtain S_M. Finally, the vector V_s is obtained by passing S_M through Leak Relu activation function and a linear layer. Vector V_s has the same shape as vector V_c. V_s is the feature vector of email subject extracted by the subject feature extraction layer of the model.

The feature vector V_e of the email is obtained by adding the subject feature vector V_s and the content feature vector V_c. V_e is the overall feature representation of the email.

Linear Classification Layer: Classification is based on the overall feature of the email representing V_e. After Leak RELU activation function and batch normalization, V_e is output to a linear layer to classification.

To sum up, the input of the email classification model is two text sequences in the email, and the output is the result of the email classification. As described above, the main purpose of the model is to classify mails, and the core structure of the model is LSTM, so we called the model LSTM based Email Classifier (LSTM-MC). The LSTM-MC model is shown in Figure 1.

3.2 Active Learning

A lot of data is needed in email recognition under deep learning model, since under such method, manual labeling by experts is needed. But the labeling cost increases with the increase of data scale. Active learning method tries to solve this problem. By selecting a certain number of samples and training only these samples, the model achieves the accuracy of training all samples.

Two active learning algorithms are used for email classification model, which both are uncertain sampling, including Least Confidence and Max Entropy. These two methods are practiced on the email classification model.

3.2.1 Selection Method. Least Confidence method is practiced to select the most inaccurate samples. The largest probability y_i assigned by model is the smallest of all samples. For email classification tasks, y_i has only two states, y_0 and y_1. $\emptyset^l$ is the information measure of the sample. The larger the $\emptyset^l$ is, the more information the sample contains, and the less confidence the model has in the result of sample judgment.

$$\emptyset^l = 1 - max_2 P(y_1, y_0 | x_i) \quad (1)$$

After obtaining the uncertainty measure $\emptyset^l$ of sample x_i, the pool-based data filtering method is chosen to select N samples with the largest $\emptyset^l$ from all data samples as the result of LC method.

$$x(1 \dots n) = max_n(\emptyset^l) \quad (2)$$

Max Entropy (ME) method is a kind of uncertain screening method, which takes Entropy[18] as the measure criterion to select some samples with the greatest entropy. For sample x_i, the uncertainty measure $\emptyset^e$ is the entropy of the classification result.

$$\emptyset^e = -\sum_t P_\theta(y_t \vee x_i) log P_\theta(y_t \vee x_i) \quad (3)$$

After obtaining the uncertainty measure of sample $\emptyset^e$ the pool-based data filtering method is used to select N samples with the largest $\emptyset^e$ from all data samples as the screening result of ME active learning method.

$$x(1 \dots n) = max_n(\emptyset^e) \quad (4)$$

3.2.2 Practice of Active Learning Algorithms. In the task of email classification, the data set is divided into labeled sample set L and unlabeled sample set U. According to the active learning strategy, a certain number of samples are selected from the unmarked dataset for annotation, and the model is trained by it. Finally, a better performance model is obtained.

Input: labeled sample set L, unlabeled sample set U, model M.
Output: Model M with better performance, marked dataset L^f.

Step 1: Neural network model M is trained according to dataset L.

Step 2: Using M and active learning algorithm, U^n and L^n were screened from unmarked dataset U and marked dataset L, respectively. The size of U^n and L^n is N.
Step 3: Label U^n and add L^n to get a new labeled dataset L^i.
Step 4: uses L^i to train M.
Step 5: Add the screening result U^n to L, and get the final labeled dataset L^f, as the next labeled dataset L.

The algorithm implementation process is shown in Figure 2.

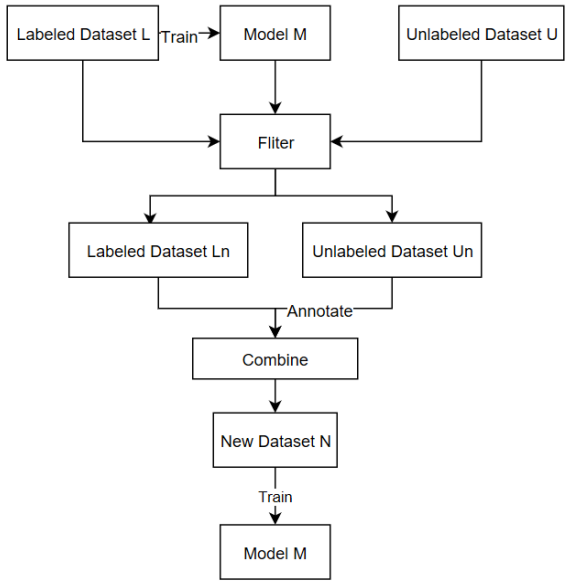

Figure 2. Implementation of Active Learning Method

4 Experiment

4.1 Dataset

To verify the effectiveness of the email classification model LSTM-MC and the active learning method, TREC Spam Dataset[2] is used. The sample size of TREC dataset is 98680. The TREC data set is divided into training set, validation set and test set. The size of training set is 57 720, the validation set is 6900, and the test set is 34360. The test set is used as the evaluation dataset to evaluate the performance of each model. The standard F1 value is used as the score of each model.

4.2 Email Classify Model Performance

In this paper, the email classification model combines content information and subject information according to the characteristics of email. The LSTM-MC is compared with other widely used and well-performing text classification models such as Text CNN[1] and Text RNN[2]. It is guaranteed as much as possible that these three different models have same size parameters to reduce the influence of hyperparameters on the model. Batch training is used to train each model and the batch is set as 128. RMSprop gradient descent algorithm is used for all three models, and the dimension of word vector is unified to 100. In the Text CNN model, the number of convolution channels is 100, and there are 10 convolution kernels of different sizes. LSTM-MC and Text RNN both use LSM to extract features, so the LSTM hidden units are unified into 1000. LSTM-MC, Text RNN and Text CNN are trained with training set, and the model effect is evaluated with test set. The performance is shown in Table 1.

Table 1. Comparation of text classification models

Model	Precision	Recall	F1 score
LSTM-MC	0.9554	0.9897	0.9722
Text RNN	0.9234	0.9859	0.9536
Text CNN	0.9345	0.9894	0.9612

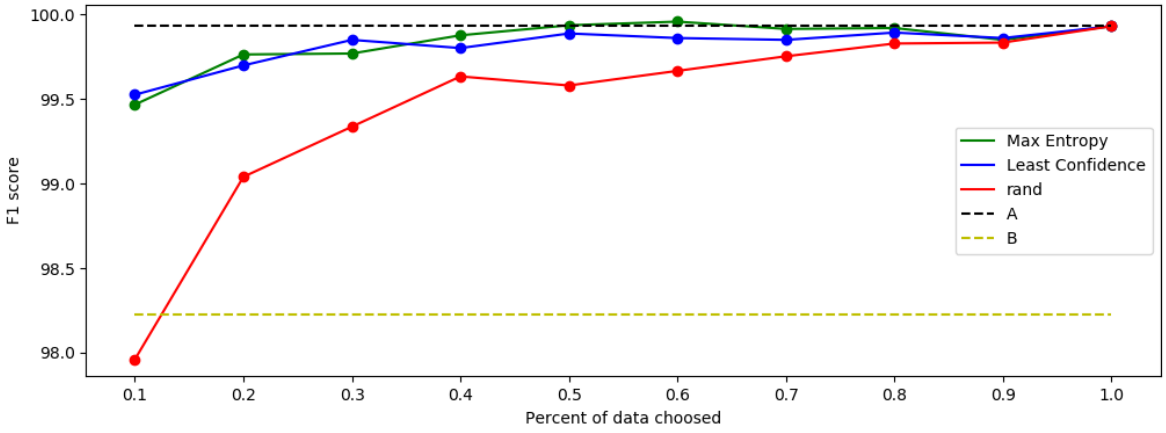

Figure 3. Comparation of active learning methods

[2] https://trec.nist.gov/data/spam.html

Table 1 shows the Precision, Recall, and F1 score of the three models on the test set after 20 training sessions in the training set.

4.3 Active Learning Performance

Active learning is effective in reducing labeling costs and accelerating training time. It is hoped that some samples can be selected by active learning method, and only those selected samples will be trained, so that the effect of training all samples with the model can achieved.

The active learning algorithm include least confidence and max entropy. The purpose of the test set is to verify the effectiveness of the active learning algorithm. LSTM-MC is used as the classifier of active learning method. At the beginning, the F1 value of LSTM-MC on the test set was 98.229.

LC and ME methods are treated as practical algorithms of active learning, and the results are compared with random selection. Select different number of samples in the test set for training, and then detect the F1 value of the model in the test set. The experimental results are shown in Figure 3.

Figure 3 shows the performance of the model with different proportion of test set samples. A is the result of training with all data, and B is the result of the model on the test set at the beginning of the experiment.

When the data size is ten percent of all data, the F1 scores of the model on the test set are 99.52 and 99.46 after using LC and ME active learning methods, which are much higher than the F1 value of the model at the beginning of the experiment, which is 98.22. And the F1 value of random initialization method is only 97.96.

5 Conclusion

In this paper, a new model is established to classify e-mails, and some active learning algorithms are implemented on this model. Experiments show that the new model has better performance than the classical CNN RNN model in the task of email classification, and the active learning method also has a good performance in this task.

ACKNOWLEDGMENTS

Financial support for this study was provided by the Science and Technology Planning Project of Sichuan Province, China (Grant No. 2017JY0080).

REFERENCES

[1] Yoon Kim.2014.Convolutional neural networks for sentence classification. arXiv preprintarXiv:1408.5882.
[2] Lai, S.; Xu, L.; Liu, K.; and Zhao, J. 2015. Recur-rent convolutional neural networks for text classification. InAAAI,2267–2273.
[3] Lewis, D. D. , & Gale, W. A. . (1994). A sequential algorithm for training text classifiers. Proc International Acm-sigir Conference on Research & Development in Information Retrieval.
[4] Seung, H. S. . (1992). Query by Committee. Workshop on Computational Learning Theory. ACM.
[5] Settles, M. Craven, and S. Ray. Multiple-instance active learning. In Advances in Neural Information Processing Systems (NIPS), volume 20, pages 1289–1296. MIT Press, 2008b.
[6] Androutsopoulos, I., Paliouras, G., Karkaletsis, V., Sakkis, G., Spyropoulos, C.D., Stamatopoulos, P.: Learning to filter spam email: A comparison of a Naive Bayesian and a memory-based approach. In Zaragoza, H., Gallinari, P., Rajman, M., eds.: Proc. Workshop on Machine Learning and Textual Information Access, 4th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD 2000), Lyon, France (2000) 1–13
[7] Thorsten Joachims. 1998.Text cate-gorization with support vector machines: Learning with many relevant features. Springer.
[8] Lee, S. W, & Verri, A. (2002). [Lecture Notes in Computer Science] Pattern Recognition with Support Vector Machines Volume 2388 || Applications of Support Vector Machines for Pattern Recognition: A Survey. (Vol.10.1007/3-540-45665-1, pp.213-236). Springer Berlin Heidelberg.
[9] Culotta, A. , & Mccallum, A. . (2005). Reducing labeling effort for structured prediction tasks. National Conference on Artificial Intelligence. AAAI Press..
[10] Scheffer, T., Decomain, C., & Wrobel, S. (2001). Active Hidden Markov Models for Information Extraction. International Conference on Advances in Intelligent Data Analysis..
[11] Burr Settles. Active learning literature survey. University of Wisconsin, Madison, 52(55-66):11, 2010.
[12] Tomanek, K., & Olsson, F. (2009). A Web Survey on the Use of Active Learning to Support Annotation of Text Data. Naacl Hlt Workshop on Active Learning for Natural Language Processing.
[13] Shen, Y. , Yun, H. , Lipton, Z. C. , Kronrod, Y. , & Anandkumar, A. . (2017). Deep active learning for named entity recognition.
[14] Gasperin, C. . (2009). Active Learning for Anaphora Resolution. Naacl Hlt Workshop on Active Learning for Natural Language Processing. Association for Computational Linguistics.
[15] Schein, A. I. , & Ungar, L. H. . (2007). Active learning for logistic regression: an evaluation. Machine Learning, 68(3), 235-265.
[16] Graves, A. . (2012). Long Short-Term Memory. Supervised Sequence Labelling with Recurrent Neural Networks. Springer Berlin Heidelberg.
[17] Ioffe, S. , & Szegedy, C. . (2015). Batch normalization: accelerating deep network training by reducing internal covariate shift. International Conference on International Conference on Machine Learning. JMLR.org.
[18] Shannon, C. E. . (1948). A mathematical theory of communication. The Bell System Technical Journal, 27.