DOI : https://doi.org/10.56452/7-3-107

ISSN: 0974-5823 Vol. 7 No. 3 March, 2022

International Journal of Mechanical Engineering

Embedding of word vectors with emotion and semantic information by semi-supervised deep learning method: enhancement of sentiment analysis

Rakesh Patra

Asst. Professor, Department of Comp. Sc. & Info. Tech., Graphic Era Hill University, Dehradun, Uttarakhand India 248002

Abstract:

Maximum all the information learning methods are work on text data sets for that they collect word vector and ignore the emotic information. In todays word emotion and feelings are as important as words. So, to improve the sentiment dataset categorization, herein research we use deep neural network method for embedding the emotion and word vector together and then perform the sentiment analysis. And shows the improvement in the text or word categorization ability and furthermore efficiently elucidatethe difficulty of lengthy text emotion categorization sample selection in semi-supervised learning technique. In this paper, active learning and adaptive deep confidence network are combined to construct a new semi-supervised learning method-active adaptive deep belief network (ADN), and it is successfully applied to sentiment classification tasks.

Keywords: learning methods, text data sets, emotic information, sentiment analysis, semi-supervised

---

I.INTRODUCTION

There are multiple machine learning methods are developed by researchers time to time to solve the lacking natural languages word and speech tagging i.e. Hidden Markov Mode[1], Maximum Entropy Model[2] and Conditional Random Fields (CRFs)[3], Joint model, Multi-layer Perceptron (MLP), CBOW and Skip-gram method, C&W model.

Compared with the above sentiment analysis methods, which are based on unsupervised, in

DOI : https://doi.org/10.56452/7-3-107

the face of supervised methods, a large number of manual labeling corpus problems are required. In this paper, we propose an active deep Belief Networks model for text sentiment classification. The model has three aspects. First, ADN uses the network structure of the adaptive deep confidence network for semi-supervised learning, continuing the superior learning and data abstraction capabilities of the deep network architecture; Second, ADN uses active learning to determine the most difficult to distinguish data that needs to be labeled, and then selects the labeled data and all unlabeled data to train the deep network architecture; Third, ADN can conduct semi-supervised learning and active learning on the same deep architecture, perform multiple iterations in the active learning process to gradually improve the abstraction and data classification capabilities.

II. Experimental Data and Evaluation Indicators

The data of this experiment comes from the COAE 2015 task, which has 133201 microblog statements, which contain a large number of interfering sentences. It is divided into four different data sets for evaluation, including books (BOO), audio products (DVD), kitchen appliances (KIT) and electronic products (ELE). Each data set involves 1000 derogatory and 1000 derogatory comment documents.

The system uses the accuracy and loss function (Loss Function, LF) as the evaluation index of the experiment. The formula for calculating the correct rate is shown in (12).

$$accuracy = \frac{The\ number\ of\ correct\ samples\ in\ the\ sentiment\ classification}{Test\ corpus\ total\ sample\ number} \quad (12)$$

Compared with the exponential loss function used by ADN, the loss function is the DBN classical mean square error loss function and the hinge loss function in SVM.

III. Experimental Setup

A different amount of nodes are used of the ADN deep architecture of each hidden layer. In the greedy unsupervised learning phase, the number of iterations used to train each layer is 20, the learning rate is 0.1, the initial impulse is 0.5, and after 5 iterations, the impulse becomes 0.9. The supervised learning phase uses a conjugate gradient descent method. For the four data sets, the ADN structure used in the experiment is 50-50-200-2, indicating that the number of nodes in the hidden layer is 50, 50, and 200, respectively, and the number of nodes in the output layer is 2. The number of nodes of input layers is the same as the number of dimensions of the input data in each data set.

All documents in this experimental data set were trained to be unlabeled, including 1000 training data and 1000 test data. Then, 2000 documents were randomly divided into 10 parts and tested using a cross-validation method. Of the 10 randomly divided, training data were

DOI : https://doi.org/10.56452/7-3-107

randomly selected to be 100 documents, and test data were the remaining 100 documents. and the remaining 100 documents were used as test data. Due to the randomness of the selected label data, all classifier experiments were repeated 100 times.

This section has carried out experimental design and analysis from the following four aspects, including:

(1) ADS's sentiment classification performance compared to classical semi-supervised learning methods in the field of sentiment classification;

(2) The performance of active learning in ADN compared with semi-supervised learning methods DBN and ADBN;

(3) Classification performance of deep network structures when different loss functions are used;

(4) Classification accuracy of ADN when using different numbers of annotated documents.

Iv Analysis of Experimental Results

(1) Comparison of ADN classification performance

The classification performance of AND is compared with six representative classifiers:Spectral (Semi-supervised spectral learning),TSVM (Transductive SVM), Active (Active learning), MECH (Mine the Easy Classify the Hard),DBN and ADBN.

For the Active, MECH, and ADN methods, an active and a negative document is selected from the initial document annotation set for active learning, and 100 documents are actively selected in the training set for manual labeling and added to the annotation data set to train the classifier. Each time, the 20 most uncertain documents are selected for manual labeling, and the newly labeled documents are added to the label data set, and then the labeled and unlabeled documents are used to further train the classifier. After actively learning iterations 5 times, 100 annotation documents were used for training.

For the Spectral, TSVM, ADBN, and DBN, 100 annotation documents are randomly selected.

For the Active, AND and MECH, by active learning 100 annotation documents are selected, the first two of which are the only two selected randomly., and only the first two annotation documents are randomly selected.

DOI : https://doi.org/10.56452/7-3-107

The test accuracy rate when using 100 labeled data training on 4 data sets is shown in Table I.

Table I.Test the correctness of the training on the data set

	Data set / Classifier	KIT	ELE	BOO	DVD
Correct rate/%	Spectral	63.7	57.7	55.8	56.2
	TSVM	65.5	62.9	58.7	57.3
	Active	68.1	63.3	58.6	58.0
	MECH	74.1	70.6	62.1	62.7
	DBN	72.6	73.6	64.3	66.7
	ADBN	75.0	75.0	66.0	67.9
	ADN	77.5	76.8	69.0	71.6

The table I shows that ADN has a higher test accuracy rate on the four data sets than the previous five methods. The reason can be classified into the following two points:

First and foremost, a deep network structure is used by ADN to map documents belonging to different categories to different areas of Euclidean space, and the information that other classifiers cannot obtain can be successfully extracted.

Second, an exponential loss function is used by ADN's global optimization to maximize and differentiate the separability of an annotated document.

(2) Active learning experiment results

100 documents are randomly selected from the training set for labeling, and then the ADBN method is used for training and testing. The design of ADN is the same as before. In order to eliminate the random factors present in the experiment, all experiments were repeated 100 times. The test accuracy of DBN, ADBN and ADN on 4 data sets is shown in Figure 1.

DOI : https://doi.org/10.56452/7-3-107

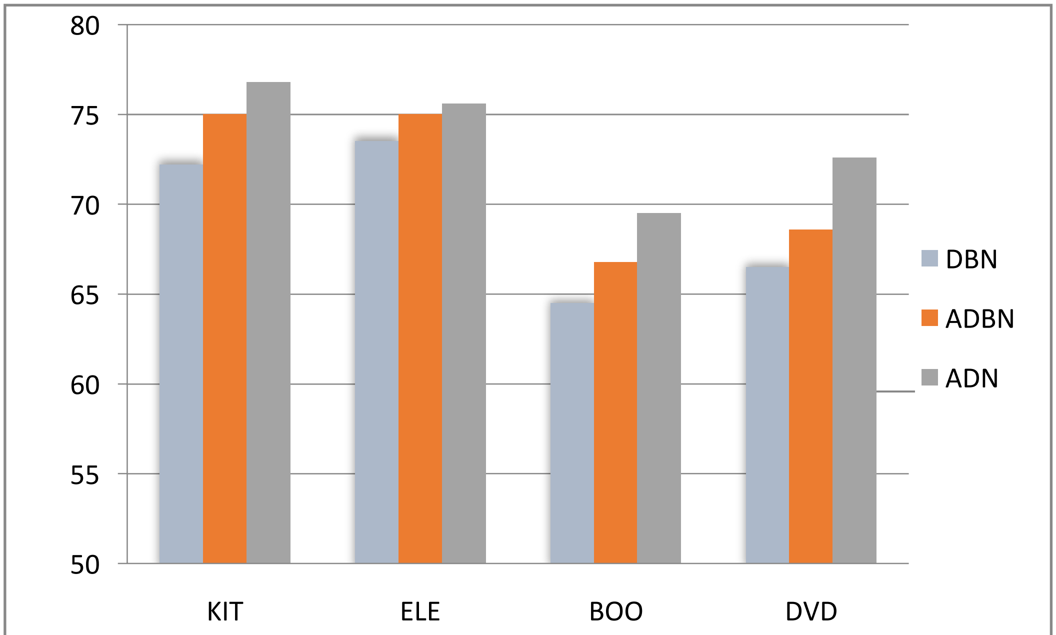

Fig. 1 Test accuracy of DBN, ADBN, and ADN on four datasets

Experiments show that the accuracy rate of ADBN on 4 data sets has increased by 2% to 3% compared with DBN. The main reason is that the exponential loss function is used in the ADBN deep network structure, and the effect is better than the mean square loss function used by DBN, the next experiment will be verified and compared. Compared with ADBN, ADN has an average correct rate of 74.8% on the dataset, which also proves that the use of active learning methods is effective.

(3) Loss function effect

The classical DBN network structure uses the mean square error loss function, and the ADN network structure uses the exponential loss function of the adaptive deep confidence network. The hinge loss function in SVM is another loss function frequently used. The experiment verifies the performance of these three loss functions in the sentiment classification problem, which is shown in Figure 2.

DOI : https://doi.org/10.56452/7-3-107

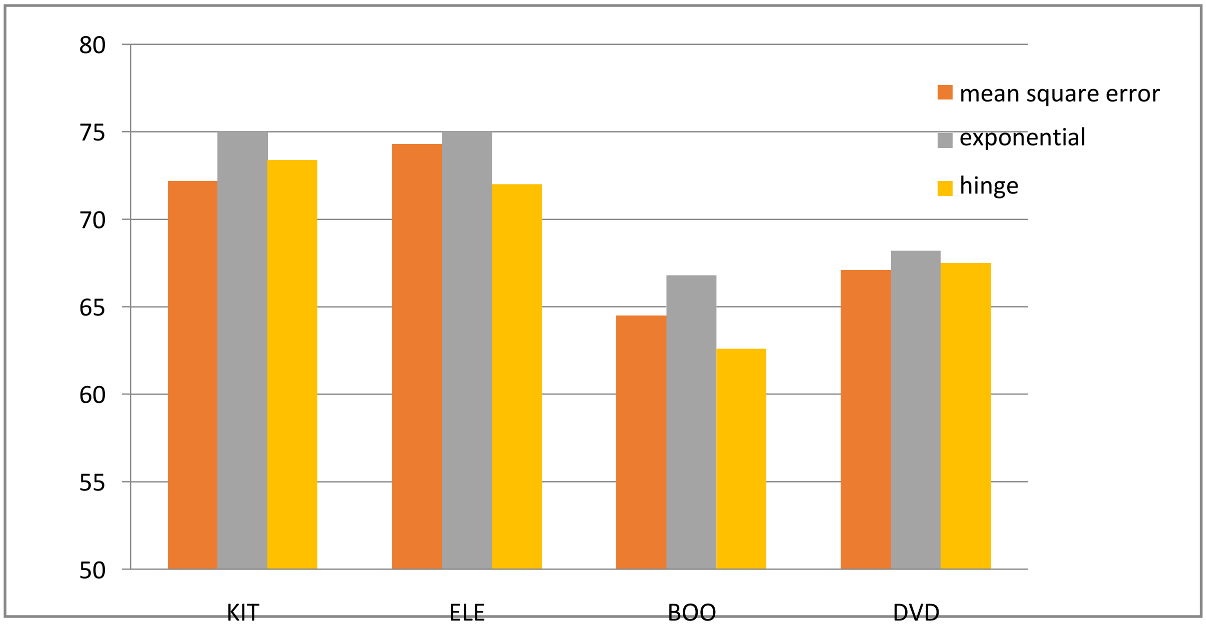

Fig. 2 The correct test rate for different loss functions

The experimental comparison results indicate that the exponential loss function has the best performance on the four review data sets. The performance of the mean square error loss function and the hinge loss function are similar, which proves the correctness of the ADN architecture using the exponential loss function.

(4) Effect of different quantity label sets

In order to further verify the ADN using different number of annotation data performance, this chapter conducted a series of experiments using different numbers of annotated documents on four data sets. The results are shown in Figure 3.

DOI : https://doi.org/10.56452/7-3-107

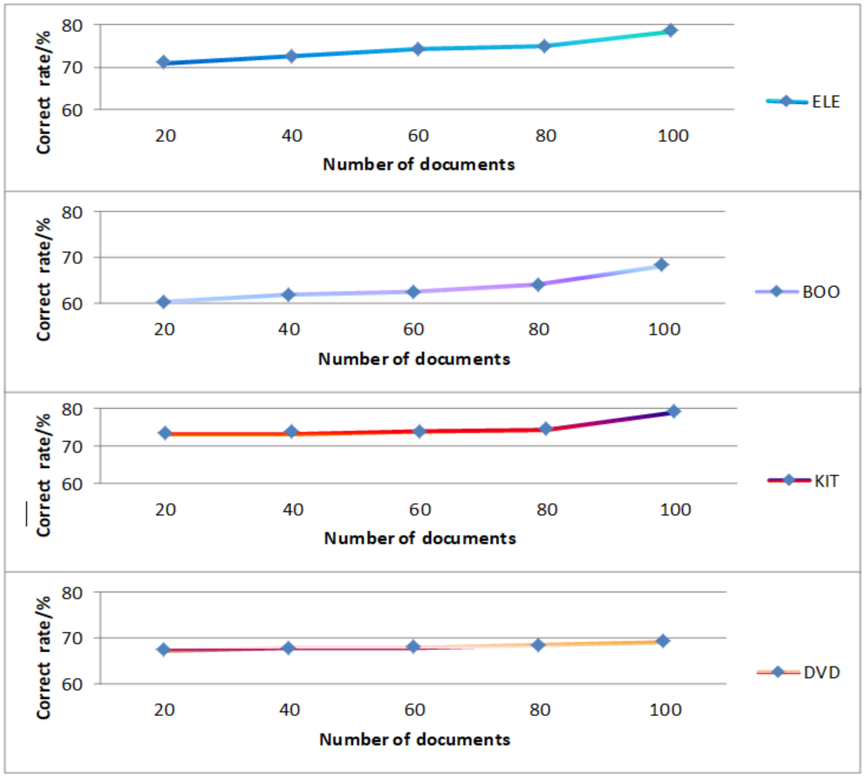

Fig. 3 Different number of labeled documents for test accuracy

Each experimental result of the ADN method is repeated and the results are averaged. An average of 73% accuracy is achieved with only 20 annotation documents. For most sentiment classification data sets, the test accuracy rate increases slowly as the annotation document increases.

CONCLUSION

In this paper, active learning and adaptive deep confidence network are combined to construct a new semi-supervised learning method-active adaptive deep belief network (ADN), and it is successfully applied to sentiment classification tasks. ADN can select the appropriate training documents for manual labeling, and can extract effective information from a large number of unlabeled documents improving the robustness of the classifier. The performance of ADN was compared with other deep learning methods. The experimental results indicate that ADN has better performance.

Through unsupervised learning and supervised learning interval training, better able to adjust the depth of the architecture parameters, improve the depth of the network structure of abstraction and classification ability. After the active learning iteration is complete, all the annotated and unlabeled documents are used to further optimize the deep architecture. Since the ADN method uses the same deep network structure to actively select the annotation dataset and

DOI : https://doi.org/10.56452/7-3-107

classify the review documents, the use of different models in document data selection and classification avoids the existence of markup document alienation. More importantly, iteratively trains the parameters of the deep structure when actively selecting the annotated data set, further improving the classification performance of ADN.

REFERENCES

[1] Si j, Mukherjee A, Liu B, et al. (2014)Exploiting Social Relations and Sentiment for StockPrediction[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing(EMNLP). Doha, Qatar:Association for Computational Linguistics: 1139-1145.

[2] Li Y, Li W, Sun F, et al. (2015) Component-Enhanced Chinese Character Embeddings. Computer Science.

[3] Kai Y, Lei J, Yuqiang C, et al. (2013) Deep learning: yesterday, today and tomorrow. Journal of computer Research and Development, 50(9): 1799-1804.

[4] Epistemological View: Data Ethics, Privacy Trust on Digital Platform, Harsh, R., Acharya, G., Chaudhary, S., 2018 IEEE International Conference on System, Computation, Automation and Networking, ICSCA 2018, 2018, 8541166

[5] Wilson T. Wiebe J., Hoffmann P. (2005)Recognizing contextual polarity in phrase-level sentiment analysis. HLT/EMNLP.

[6] Enhance the Data Security in Cloud Computing by Text Steganography, Sanghi, A., Chaudhary, S., Dave, M., Lecture Notes in Networks and Systemsthis link is disabled, 2018, 18, pp. 241–248

[7] Bengio Y, Schwenk H, Sen´ecal J S, et al. (2003) A Neural Probabilistic Language Model. Journal of Machine Learning Research, 3(6):1137–1155.

[8] Mikolov T, Chen K, Corrado G, et al. (2013) Efficient Estimation of Word Representations in Vector Space. Computer Science.

[9] Mikolov T, Sutskever I, Chen K, et al. (2013) Distributed Representations of Words andPhrases and their Compositionality. Advances in Neural Information Processing Systems, 26:3111–3119.

[10] Collobert R, Weston J. (2008) A Unified Architecture for Natural Language Processing: Deep Neural Networks with Multitask Learning[C]//Proceedings of the 25th International Conference on Machine Learning. New York, NY, USA: ACM, ICML'08.

[11] Tang D, Wei F, Yang N, et al. (2014) Learning Sentiment-Specific Word Embedding for Twitter Sentiment Classification[C]// Meeting of the Association for Computational Linguistics:1555-1565.

[12] Maas A L, Daly R E, Pham P T, et al. (2011) Learning word vectors for sentiment analysis[C]// Meeting of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics:142-150.