Sentiment Classification on Weibo Incidents Using CNN-SVM and Repost Tree

Manshu Tu[1, a],Shengxiang Gao[2,b],Zhe Ji[3,c],Yan Zhang[4,d],and Yonghong Yan[5,e]

[1,4,5]The Key Laboratory of Speech Acoustics and Content Understanding Instituteof Acoustics, Chinese Academy of Sciences, Beijing, China

[2,3]National Computer network Emergency Response technical Team/Coordination Center of China, Beijing

[a]tumanshu@hccl.ioa.ac.cn, [b]gao.shengxiang@cert.org.cn, [c]jz@cert.org.cn,[d]zhangyan@hccl.ioa.ac.cn,[e]yanyonghong@hccl.ioa.ac.cn

Keywords: Sentiment classification. CNN-SVM. Repost tree

Abstract. Sentiment classification on weibo has recently attracted wide attention in research community. Most previous works are focused on weibo comments regarding movies or products. Our study, in contrast, is aimed at gusty incidents on weibo. Comments of the incidents are considered either positive or negative representing attitudes of users towards these incidents. Classifying users' attitudes helps identifying the general opinion of the public. In this paper, we propose an innovative convolutional neural networks based method, termed as CNN-SVM, to classify the incident comments. In addition, according to users' repost actions, we propose a new data structure, repost tree, for dealing with ambiguity in the comments. Extensive experiments demonstrate that the CNN-SVMmethod effectively improves the accuracy of incidents sentiment classification. The new data structure shows to be effective on steering the classification results towards real world sentiment tendency.

Introduction

In this paper, we are interested in analyzing people'semotional attitudes (i.e. positive and negative sentiment bias) towards incidents.We want to improve the accuracy of sentiment classification through modifyingprevious methods. We also take the advantage of weibo data features to obtainthe real ratio of sentiment.

Pang et.al examined the effectiveness of machine learning methods for sentimentclassification tasks on movie reviews. Three machine learning techniquesused in their experiments are Naive Bayes, maximum entropy classification andsupport vector machines (SVM). SVM tends to do the best in terms of performances[1]. It is commonlyused to solve small sample, nonlinear and high dimensional pattern recognitionproblems [2].

CNN usingconvolution to the local features of the filter layer was originally invented for thecomputer vision [3]. Kim trained a simple CNN with one layer convolution ontop of word vectors [4]. In contrast to conventional NLP tasks, sentiment analysis onmicroblogging posts is challenging due to the limited text length.

As described above, CNN and SVM are both popular methods for sentimentclassification. However, to the best of our knowledge, their combinedperformance has not been tested in sentiment classification tasks. None of thesefunctions above has considered microblogs reposts. In order to address theseissues, we propose to combine the CNN and SVM together to build a new classificationmodel. We further propose a novel repost tree to take context intoconsideration for sentiment classification tasks. The repost tree can perform alogical operation on sentiment for each node to adjust its sentiment predictionresult. Our experiments show that the CNN-SVM with the repost tree can getresults close to the real world sentiment tendency ratio. The rest of the paperis organized as follows: Section 2 describes in details the model CNN-SVM andrepost tree. Section 3 presents datasets used in our experiments and data preprocessing.Section 4 gives the conclusions of our research and future works.

Copyright © 2016, the Authors. Published by Atlantis Press.
This is an open access article under the CC BY-NC license (http://creativecommons.org/licenses/by-nc/4.0/).

Model

CNN-SVM. Our proposed model is shown in Fig. 1. It is inspired by Kim's CNN model[4]. A CNN is used for feature extraction, while a SVM carries out the sentimentclassification. The model works as follows. The input of this model is amatrixS $\in R^{(s\times n)}$. The notation n means that every word in a sentence hasn pre-trained vectors:$[v_1,\cdots,v_i,\cdots,v_n]$. The notations refers to every sentenceconsisting of s words:$[w_1,\cdots,w_i,\cdots,w_s]$. The convolution layer of the CNNhas a number of different types of filters F $\in R^{(m\times n)}$, where m is the width ofthe filter. The result of the ith matrix vectors in a sentence through the filter iscomputed as follows equation:

$$c_i = (S * F)_i = \sum_{ki}(S_{[i-m+1:i,:]} \otimes F)_{ki} (1)$$

Where k is the k-th convolution layer, while⊗is the element-wise multiplicationand $S_{[i-m+1:i,:]}$ is a matrix slice of size m along the columns. Every sentencematrix S gets a feature map c $\in$ $R^{(1\times(s-m+1))}$ through one filter for convolutionoperations.To form a richer representation of the data, one kind filter has pconvolution kernels. Suppose there are t types of filters, then one sentence willget p × t feature maps C: $[c_{f_{11}},\cdots,c_{f_{1p}},\cdots,c_{f_{t1}},\cdots,c_{f_{tp}}]$. Max-pooling is apooling layer to find the maximum value in kernel size. At the max-poolinglayer, every feature map c returns the largest value:c_{pool}: $R^{1\times(s+m-1)} \rightarrow R^{1\times1}$. Now we get the high-dimensional features: D: $[c_{pool_{11}},\cdots,c_{pool_{1p}},\cdots,c_{pool_{tp}}]$. The output of the max-pooling layer is passed to a fully connected soft-maxlayer. It computes the probability of the sentence labels as follows equation, e.g.:

$$p(y = j|C_{pool}, B) = softmax_j(C_{pool}W + B) = \frac{e^{C_{pool}w_j+b_j}}{\sum_{k=1}^{K} e^{C_{pool}w_k+b_k}} (2)$$

Where w_k and b_k are the weight vectors and bias of k-th class. The *softmax-layer*compares the predicted labels and the real labels to fine-tune the CNNmodel. When the CNN accuracy remains stable, the training data that havehigh-dimensional vectors is sent to the SVM model. Then we train the SVMmodel until its best accuracy. Retaining these parameters we have trained, thensend the test data to the model for classifying.

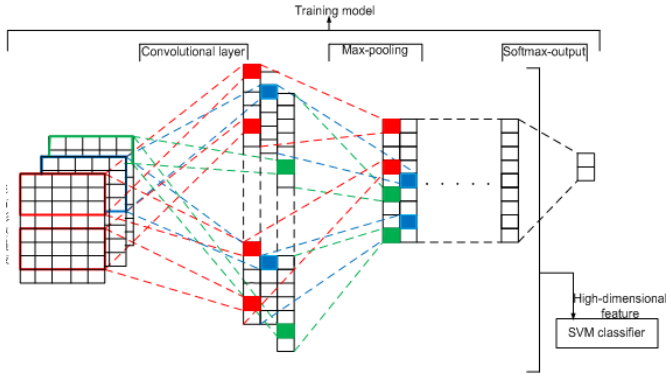

Fig. 1. In this figure, there are three different filters. At the end of the convolutional layer, every columnrepresents a sentence go through a kind of filter. Every kind feature map has *p* featuremaps.These feature maps go through max-pooling layer changing to *p*high-dimensionalfeatures.

Fig. 2. The different sizes of the circles represent different hierarchies. The number inside oroutside the circle is the sentiment result using a classification algorithm

Repost Tree. Reposting a microblog is a common behavior in microblogging. Users reposta microblog to air their opinion about the microblog they have forwarded.The repost can express different types of opinions. For example, when an incidentoccurred, one user comments with a negative message. Then another userreposts this message and writes a positive sentence, expressing a supporting attitude. It means that the second user has a negative attitudetowards this

incident. If we only analysis the first sentence, the classifierwill classify it to be positive, but the real sentiment of this user is negative.

Weibo has a convenient feature for building a repost tree. It shows pathsof each repost. The paths include nicknames and contents like "@+nickname :+content + //". We build repost trees using this structure to adjust sentimentclassification results. Repost trees contain nickname, content and timestamp.The structure of the proposed tree is shown in Fig. 2. We arrange all thenodes according to the number of their child nodes in descending order. Then the repost tree can proceed to its logical operation step.We define the statusof all nodes on a branch as $Branch_{node} = n_1, \cdots, n_j, \cdots, n_n$ the value of n_i iseither +1 or -1. The logical operation result for $Node_n$ is regarded as the nodereal sentiment tendency. It is computed as follows:

$$Node_n = \prod_{i=1}^{N} n_i (3)$$

Experiments

Datasets and preprocessing. Our model is tested in three datasets. All of them are extracted from SinaWeibo.MN is used for this data abbreviation.WZ is used for this data abbreviation.Paris for this data abbreviation.

Dataset	Training	Test
MN	1600	400
WZ	23120	5780
Paris	480	120

Table 1. Datasets

Data	SVM	RBM	DBN	CNN	CNN-SVM
MN	84.28%	84.17%	84.25%	91.96%	93.28%
WZ	55.73%	56.23%	54.38%	71.98%	77.56%
Paris	70.16%	56.86%	54.55%	74.85%	70.50%

Table 2. Results

We use the method described in 3.2 to label MZ and Paris data. All the dataare processed using the following steps: deleting URL, word segmentation usingICTCLAS, removing stop words and sentences with number of words less thantwo. The details of these datasets are list in the Table 1. Training datasets andtesting datasets have the same negative examples to positive examples ratio. Thesize of each training set is quadruple to that of its corresponding testing set.

Distance Supervision and Feature. Emoticons Users nowadays like to air their opinion with emoticons. Weibocome with many emoticons. Some of them show obvious emotional tendency,therefore we use these emoticons as labels Go et al.[5]. In weibo, the emoticonsare encoded in the form of "[Chinese word]". We convert obvious emoticonsto "+1" or "-1", where "+1" represents positive sentiment and "-1" representsnegative sentiment.We use the publicly available word2vec vectors that were trained on threebillions weibo sentences. The vectors have a dimensionality of 200, containing1043631 words [6].

Other Methods for Comparing. We use SVM as a baseline.The Deep Belief Network (DBN) and Restricted Boltzmann Machine (RBM)also be used in our experiments [7]. We implement the DBN and RBM modelin[8] for comparison. We also include a simple CNN model in our experiments.A five-fold cross-validation are conducted in all models except the dictionarymethod.

Performance Results Table. 2 shows the accuracy of CNN-SVM in comparison with the othermethods. We first compare the result of these three datasets, we find that allmethods have better performance in MN dataset than that in WZ and Parisdataset. The reason is that the MN dataset is a standard dataset, all the sentencesare integrated and coherent.

In these three datasets, RBM and DBN have similar poor accuracies. Onone hand DBN consists of RBMs, so DBN performs poorly when RBM has weakresults. One the other hand, RBM is not very suitable for nonstandard textdata classification. CNN performs with good accuracies in both three datasets.It improves with 7.68% in the first dataset, 15.75% in the second datasets and4.69% in third datasets, compared with the best performance among SVM, RBMand DBN. The proposed CNN-SVM shows higher accuracy than that of CNN. Ithas 1.32% increase in the first dataset and 5.58% increase in the second dataset,2.65% increase in the third dataset compared with CNN. The improvement infirst dataset is less than second and third dataset. This might due to that CNNhas shown a good performance in the first dataset already. The SVM has a betterperformance in the

third dataset than in other two datasets, because SVM isusually suitable for small datasets.
For the repost tree, we construct experiments using the Paris dataset, whichhas 14470 repost microblogs. At the website, (the website reported the People'sDaily survey result of the sentiment tendency on this incident), we take thesurvey result as real world ratio of positive and negative sentiment of the public.As shown in Fig. 3, the prediction with CNN-SVM differs by 20% compared withthe real ratio. However, after applying the repost tree, the prediction results aresteered close to the real ratio. According to the result, CNN-SVM performs thebest when connected to the repost tree. It differs from the real ratio by 7.46%.

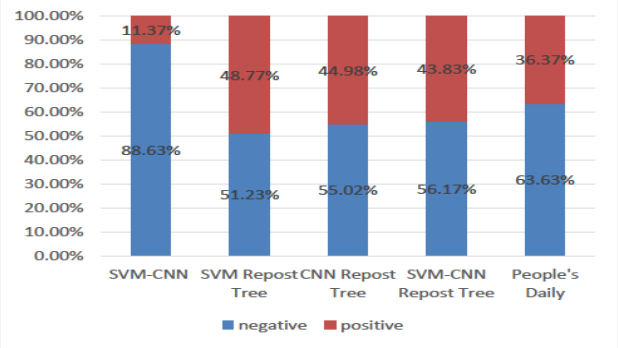

Fig. 3.

Conclusion and Feature Work

In this paper we show that CNN-SVM is a method with better performanceon weibo incident sentiment classification compared to other state-of-the-art approaches.It can improve the accuracy of incidents sentiment classification by alarge margin. Besides, the new data structure, repost tree, can further adjustthe prediction result of CNN-SVM towards the real world sentiment tendency.The repost tree might have ambiguity, we will consider thissituation in our following work.

Acknowledgements

We acknowledge the support of China post doctoral science foundation 2015LH0041.

References

[1] Pang B, Lee L, Vaithyanathan S: Thumbs up? :sentiment classification using machinelearning techniques. Proceedings of Emnlp. 147, 79–86 (2009)

[2] Liu Xia, Lou Reed. Application of SVM in text classification. Computer education 72-74. (2007)L é cun Y, Bottou L, Bengio Y, et al.: Gradient-based learning applied to documentrecognition. Proceedings of Emnlp. 86(11), 2278-2324 (1998)

[2] Kim Y, Kim Y.:Convolutional Neural Networks for Sentence Classification. EprintArxiv.(2014)

[3] Johnson R, Zhang T.:Effective Use of Word Order for Text Categorization withConvolutional Neural Networks. EprintArxiv.(2014).

[4] Iyyer M, Enns P, Boyd-Graber J, et al.:Political Ideology Detection Using Recursive Neural Networks. Meeting of the Association for Computational Linguistics. 2014:1113-1122.

[5] Hinton G E, Osindero S, Teh Y W.:A fast learning algorithm for deep belief nets. Neural Computation,18(7),1527-54,(2006)

[6] Zhou S, Chen Q, Wang X.:Fuzzy deep belief networks for semi-supervised sentiment classification. Neurocomputing,131(9),312–322,(2014)