Information Sciences 477 (2019) 55–64

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Three-way enhanced convolutional neural networks for sentence-level sentiment classification

Yuebing Zhang[a,b], Zhifei Zhang[a,c,*], Duoqian Miao[a,b,*], Jiaqi Wang[a,b]

[a] Department of Computer Science and Technology, Tongji University, Shanghai 201804, China
[b] Key Laboratory of Embedded System and Service Computing, Ministry of Education, Shanghai 201804, China
[c] State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210023, China

ARTICLE INFO

Article history:
Received 7 November 2017
Revised 19 October 2018
Accepted 20 October 2018
Available online 22 October 2018

Keywords:
Three-way decisions
Sentiment classification
Convolutional neural networks

ABSTRACT

Deep neural network models have achieved remarkable results in sentiment classification. Traditional feature-based methods perform slightly worse than deep learning methods in terms of classification accuracy, but they have their own advantages in interpretability and time complexity. To the best of our knowledge, few works study the ensemble of deep learning methods and traditional feature-based methods. Inspired by the methodology of three-way decisions, we proposed a three-way enhanced convolutional neural network model named 3W-CNN. 3W-CNN can be seen as an ensemble method which uses the enhance model to optimize convolutional neural networks (CNN). The enhance model is selected according to the classification accuracy and the difference in classification results compared to CNN. Support vector machine with naive bayes features (NB-SVM) is selected as the enhance model after comparing with several baseline models. However, the performance of NB-SVM is worse than CNN on most of benchmark datasets. To address this issue, we construct a component named confidence divider and design a confidence function to distinguish the classification quality of CNN. NB-SVM is further utilized to reclassify the predictions with weak confidence. The experimental results validated the effectiveness of 3W-CNN and showed three-way decisions could further improve the accuracy of sentiment classification.

1. Introduction

Text sentiment analysis (also known as opinion mining) that analyzes people's opinions and emotions from text is an active research field in natural language processing (NLP) [1]. The existing studies of sentiment classification, as an important part of sentiment analysis, can be mainly grouped into two categories: lexicon-based and corpus-based approaches. Lexicon-based approaches typically use a dictionary of sentiment words and phrases, and incorporate intensification and negation to compute a sentiment score for each text [14]. Corpus-based approaches treat sentiment classification as a special case of text categorization problem, which utilize machine learning methods to extract reasonable features from texts and feed into a classifier to predict the sentiment [29].

The existing studies of corpus-based approaches are dominated by two main directions: traditional feature-based methods and deep learning methods. Traditional feature-based methods extract manually designed features from the text, e.g.,

* Corresponding authors.
E-mail addresses: yuebing_zhang@hotmail.com (Y. Zhang), zhifeizhang@tongji.edu.cn (Z. Zhang), dqmiao@tongji.edu.cn (D. Miao).

https://doi.org/10.1016/j.ins.2018.10.030
0020-0255/

N-gram (unigrams, bigrams, trigrams), Part-of-Speech (POS), term frequency-inverse document frequency (TF-IDF), and then use the features and a classification model to complete the sentiment classification task. The common classification models include naive bayes, support vector machine [29], support vector machine with naive bayes features [39], etc. Deep learning models have achieved remarkable results in computer vision and speech recognition in recent years. Within NLP, many works with deep learning models have focused on learning word embeddings (also known as word vectors) with neural language models [25] and performing composition over these word embeddings for classification [4]. Word embeddings have lower dimension than bag of words (BoW), and semantic information of words is encoded into such a dense representation. Moreover, phrase vectors and document vectors are presented [17] to find a good representation for each phrase or document, respectively. Convolutional neural networks (CNN) are popular deep learning models which utilize layers with convolving filters that are applied to local features. CNN models have been shown to be effective for NLP and have achieved excellent results in semantic parsing [45], search query retrieval [31], sentence modeling [12], and other traditional NLP tasks [4]. Kim [15] trained a simple CNN with one layer of convolution on top of word vectors obtained from an unsupervised neural language model for sentence-level classification, and achieved excellent results on multiple benchmarks. But there is still much room for improvement in accuracy.

Generally speaking, deep learning methods perform slightly better than traditional feature-based methods in terms of classification accuracy on most of sentiment classification tasks, but traditional feature-based methods have advantages in interpretability and time complexity. We wonder whether there is an effective method to combine deep learning methods and traditional feature-based methods in order to improve the overall performance. Three-way decisions can exactly address this issue. The methodology of three-way decisions [42] is widely applied in many theoretic fields, such as management sciences, social judgement theory, fuzzy sets theory [44], shadowed sets theory [48] and knowledge granulation [8]. Three-way decisions are also widely used in numerous application fields, including medical decision making [41], credit scoring [24], government decision, e-mail spam filtering [50], face recognition [18], and clustering analysis [46]. To give a formal description of three-way decisions, Yao [43] presented a general overview on existing three-way decisions researches, and extended the rough sets-based three-way decisions to a much wider frontier, which outlines a unified theory of three-way decisions.

In this paper, we use traditional feature-based methods as the enhance model to enhance the CNN [15], which is a very popular deep learning model in NLP tasks, in the framework of three-way decisions. The test data is divided into two groups on top of output layer in CNN, one group is well classified by CNN, and the other is the opposite. The other group is reclassified by the enhance model to improve the performance. The remainder of this paper is organized as follows. In Section 2, we briefly introduce some related works on sentiment classification. In Section 3, we propose 3W-CNN for sentence-level sentiment classification. In Section 4, we report the experimental results and analysis. Finally, we make a conclusion in Section 5.

2. Related work

Sentiment classification is a fundamental and important study area in sentiment analysis. It hammers at detecting the sentiment polarity of a sentence or a document [38] based on its textual content. Sentiment classification has wide applications, such as product ranking [23] and product sales forecasting [6]. Taking a panoramic view of this area, there are two main directions for sentiment classification: lexicon-based approaches and corpus-based approaches. Lexicon-based approaches typically use sentiment dictionary, intensification and negation to compute a sentiment score for each text. Sentiment words and phrases are marked with sentiment polar and sentiment strength in sentiment dictionary. There are two kinds of sentiment dictionaries according to universality. One is a universal sentiment dictionary which is applicable to almost all fields; the other is a domain sentiment dictionary which is applicable to specific fields. Turney [37] proposed a simple but representative lexicon-based method to classify reviews into recommended or not recommended. The classification of a review is predicted by the average semantic orientation of its phrases, and the semantic orientation of a phrase is calculated by the mutual information between the given phrase and the word “excellent” minus the mutual information between the given phrase and the word “poor”. Ding et al. [5] applied negation words (e.g., not, never, cannot) and contrary words (e.g., but) to improve the performance of lexicon-based method. Thelwall et al. [36] developed *SentiStrength* with sentiment lexicon and linguistic rules for detecting sentiment strength of tweets. In lexicon-based sentiment classification, the problem of contextual polarity is a major cause for classification error. Cho et al. [2] presented a data-driven method of adapting sentiment lexicon to diverse domains. In essence, the method compared the positive/negative review’s dictionary word occurrence ratio with the positive/negative review ratio itself, in order to determine which sentiment words to be removed and which sentiment words’ polarity to be switched.

As for corpus-based approaches, Pang et al. [29] pioneered to treat sentiment classification as a special case of text categorization and applied three machine learning methods (naive bayes, maximum entropy and support vector machines). Support vector machines with bag-of-word features achieved the best performance. Many studies are inspired by this idea and focus on designing effective features to obtain a better performance on sentiment classification. Wang et al. [39] presented support vector machine with naive bayes features (NB-SVM) for sentiment classification on movie and product reviews, which is a SVM variant using NB log-count ratios as feature values. Katz et al. [13] presented *ConSent*, a novel context-based approach which was effective both in noiseless and noisy text. Salvador et al. [7] used meta-learning to combine and enrich several baseline methods (bag of words, n-grams, lexical resource-based classifier) aiming at cross-domain polarity classification. Mohammad et al. [9] combined active learning and self-training to handle cross-lingual sentiment classification.

In recent years, deep learning models have been widely used in NLP. In the beginning, neural network models were used to learn distributed vector representations of word, paragraph [25], and document [17]. Tang et al. [35] proposed sentiment-specific word embeddings for sentiment classification. Meanwhile, some word embeddings for specific tasks like Twitter sentiment classification were proposed [30]. Neural networks were further applied to sentiment classification. Kim [15] trained a simple CNN for sentence classification, and Kalchbrenner [12] subsequently applied dynamic k-max pooling over time to generalize the original max pooling in traditional CNN. Recurrent neural networks (RNN) [11] and some extensions, such as bidirectional recurrent neural networks (BRNN) [16] and gates recurrent neural networks (GRNN) [34], were applied to sentiment classification. In addition to the commonly used neural networks in computer vision, Zhao et al. [49] proposed a self-adaptive hierarchical sentence model which is dedicated to text classification.

Three-way decisions, which give the inspiration for this paper, are a new interpretation of rules in rough set theory [42]. The applications of three-way decisions are widely concerned in many fields. Yu et al. proposed an active three-way clustering method via low-rank matrices for multi-view data [46]. Zhou et al. discussed the applications of three-way decisions in e-mail spam filtering [50]. Liu et al. applied three-way decisions into investment decision [20] and government decision. Li et al. introduced three-way decisions to face recognition and proposed a cost-sensitive sequential three-way decisions method for face recognition [18]. Li et al. proposed a DNN-based sequential granular feature extraction method and a cost-sensitive sequential three-way decisions strategy [19]. Zhang and Min combined three-way decisions with random forest, and applied three-way decisions into recommender systems [47]. Min et al. proposed a frequent pattern discovery algorithm for a new type of pattern by dividing the alphabet into strong, medium, and weak parts, which is inspired by the methodology of three-way decisions and protein tri-partition [26]. The above works enrich the theoretic foundation of three-way decisions, and indicate that three-way decisions are applicable for many practical decision problems.

3. The proposed approach

We proposed a three-way enhanced CNN method (3W-CNN) for sentiment classification. In three-way decisions theory, decision results are divided into three parts, namely, accept, reject and delay decision. In binary sentiment classification, these three parts correspond to positive, negative and boundary, respectively. The instances in boundary region need more information to be classified into positive or negative class. In order to divide the classification results of CNN into positive, negative and boundary region, we construct a confidence divider. Confidence divider can measure the confidence of the classification results of CNN and divide the classification results into three parts by the confidence values. The results with strong confidence follow the predictions of CNN, i.e., positive or negative. The results with weak confidence are divided into boundary region. Afterwards, we use another classification model to reclassify the instances in boundary region. It is worth mentioning that the classification result of "another classification model" (hereinafter referred to as enhance model) needs to be enough different from that of CNN, which guarantees the classification performance of enhance model on boundary region can be better than that of CNN. Finally, we combine the results of CNN and enhance model to output the final classification results.

3.1. Structure

As shown in Fig. 1, 3W-CNN adds a confidence divider and an enhance model compared to the original CNN. The classification process of 3W-CNN is described as follows. Firstly, the entire training data is used to train CNN and enhance model respectively. Secondly, the test data is classified by CNN. Next, the confidence divider divides the classification result of CNN into two parts, definite and uncertain (can also be considered as three parts, i.e., positive region, negative region and boundary region). The instances with weak confidence, which are named boundary data, will be reclassified by the enhance model. The prediction labels of other instances with high confidence values follow the prediction of CNN. Finally, the final classification results of the entire test data are derived from reclassification results of the enhance model and reasonably classification results of CNN.

The CNN structure in this paper is the same as that of Kim [15] with one layer of convolution and one layer of max-pooling, which is shown in Fig. 2. We use pre-trained word vectors obtained from an unsupervised neural language model as inputs. These vectors were trained by Mikolov et al. [25] on 100 billion words of Google News and are publicly available[1].

3.2. Confidence divider

The classifiers need to provide not only the prediction labels but also the confidence values for the prediction labels. The confidence values indicate the confidence of the classifier to correctly classify the current input. We define a confidence function as the mapping from prediction results to their confidence values. Confidence function is a key to confidence divider because it is the basis for dividing the instances.

Confidence function can be defined as follows:

$$\tilde{CF}(label_i^p | input_i) = \begin{cases} \beta_i & label_i^p = label_i^g \\ \gamma_i & label_i^p \neq label_i^g \end{cases} \quad (1)$$

[1] http://code.google.com/p/word2vec/.

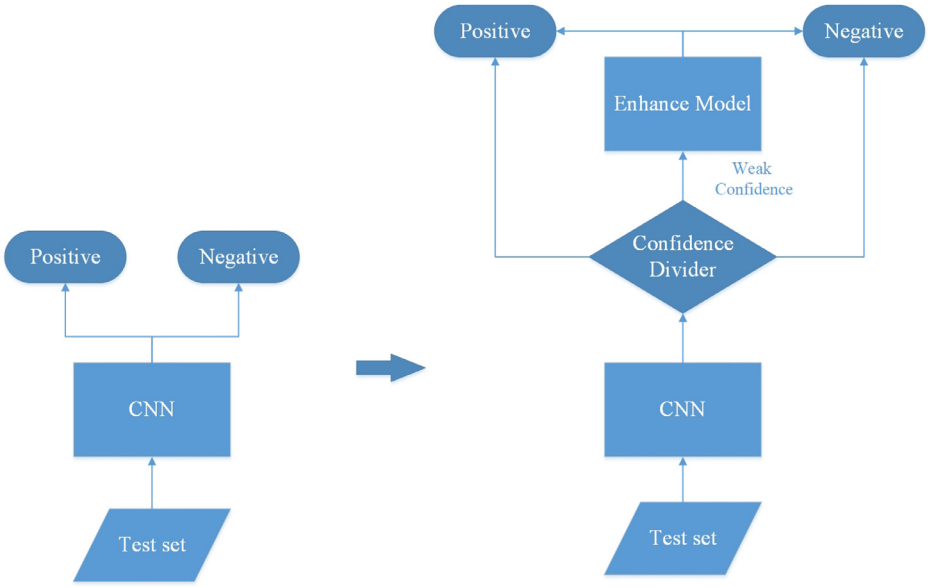

Fig. 1. From CNN to 3W-CNN.

where $input_i$ denotes the i-th input instance, $label_i^p$ denotes the prediction label of $input_i$, and $label_i^g$ denotes the golden label of $input_i$, β_i and γ_i range from 0 to 1. Theoretically, a perfect confidence function makes the confidence values of the correctly classified inputs to be 1 (namely $\beta_i = 1$), and the confidence values of misclassified inputs to be 0 (namely $\gamma_i = 0$). However, designing a perfect confidence function is almost impossible. We can try to construct a reasonable and effective confidence function to approximate the perfect one. The higher β_i, the lower γ_i, the better confidence function.

There is a simple and intuitive constructor method for confidence function in neural network models. Taking binary classification as an example, the output layers of most of neural network models for binary classification are softmax layers with two neurons. Let o_1 and o_2 denote the output values of the softmax layer, and o_1, o_2 can be regarded as the scores of the two categories respectively. The most commonly used prediction method of neural network models is to select the category with a higher score in the output layer. Therefore, the intuitive constructor method of confidence function in neural network models is defined as follows:

$$CF(o_1, o_2) = |o_1 - o_2| \tag{2}$$

The definition of confidence function in Eq. (2) is intuitive and reasonable. The higher $|o_1 - o_2|$, the more confident the model is to classify the current input into the category with higher score. In experiment part, we can see that Eq. (2) is simple but effective.

3.3. Enhance model

The enhance model, as shown in Fig. 1, is also an important component of 3W-CNN. We use the enhance model to reclassify the boundary data from CNN. Many classification models can be options for the enhance model, e.g. recurrent neural networks (RNN), support vector machine (SVM), multinomial naive bayes (MNB) or support vector machine with naive bayes features (NB-SVM). In this section, we discuss how to select a suitable enhance model for CNN.

To illustrate the conditions that a suitable enhance model for CNN should meet, we assume that the two models are independent in classification. Let P_1, P_2 denote the classification accuracy of the CNN and the enhance model on entire test data, respectively. Let α denote the ratio of boundary data to entire test data, E_1^b denote the error rate of CNN on boundary data, c denote the ratio of instances in the boundary data whose prediction label is changed by the enhance model, r denote the ratio of instances in changed boundary data whose prediction label is changed to be correct. D denotes the difference between the CNN and the enhanced model in classification result, defined by $D = \frac{N_d}{N}$, where N is the number of instances in test set and N_d is the number of instances in which the CNN and the enhanced model give different prediction labels. Based on $c \propto D$ and $r \propto P_2 - (1 - E_1^b)$, the classification accuracy of 3W-CNN on test data $P_{1,2}$ can be computed as follows:

$$\begin{aligned} P_{1,2} &= \frac{P_1 N - E_1^b \alpha N + \alpha N c r}{N} \\ &= P_1 - E_1^b \alpha + \alpha c r \\ &= P_1 + \alpha (c r - E_1^b) \end{aligned}$$

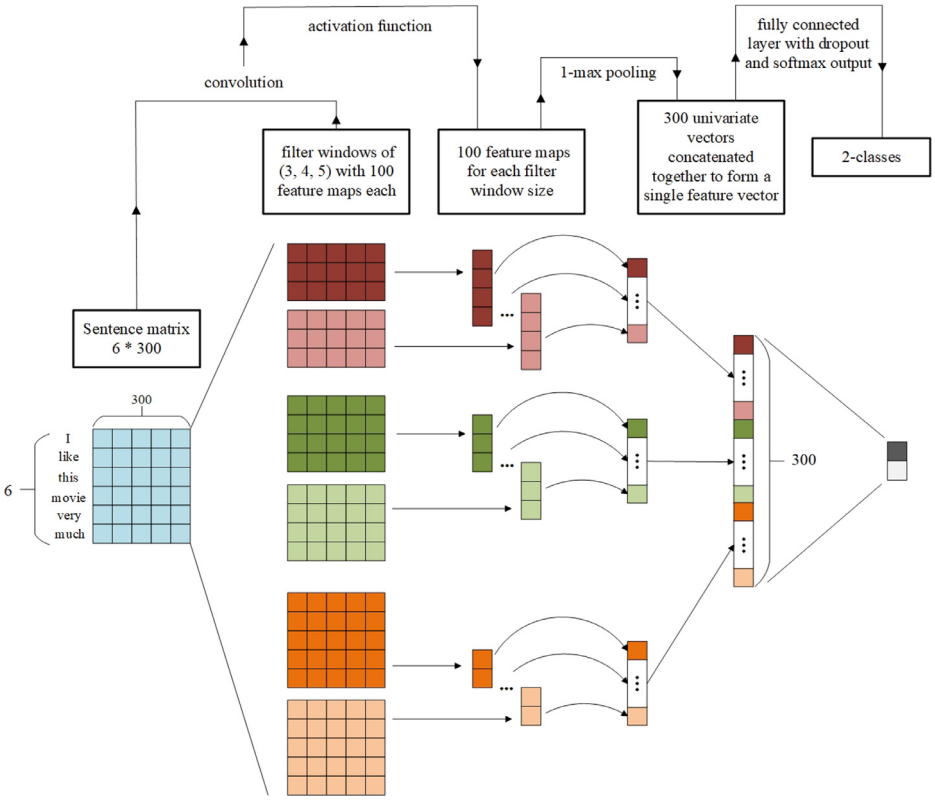

Fig. 2. Structure of CNN for Sentence Classification in 3W-CNN.

$$\propto P_1 + \alpha D(P_2 - (1 - E_1^b)) \tag{3}$$

We can see the enhance model affects $P_{1,2}$ with D and P_2. P_2 needs to be higher than $1 - E_1^b$ in order to guarantee a good $P_{1,2}$, which means the accuracy of the enhance model on entire test data needs to be higher than the accuracy of the CNN on boundary data. Moreover, P_2, D the higher, $P_{1,2}$ the better.

In order to explain why $P_2 \geq 1 - E_1^b$, we let $1 - E_2^b$ denote the classification accuracy of the enhance model on boundary data, $1 - E_1^r$ and $1 - E_2^r$ respectively denote the classification accuracy of the CNN and the enhance model on the rest data (data outside the boundary data). Obviously, $1 - E_1^b \leq P_1 \leq 1 - E_1^r$. Since the CNN and the enhance model are independent in classification result, we have $1 - E_2^b = P_2 = 1 - E_2^r$, which means it is possible that $P_2 \geq 1 - E_1^b$ when $P_2 \leq P_1$. Actually, some instances are easy to predict so that they can be correctly classified by both the CNN and the enhance model, while some instances are hard to predict so that they are misclassified by both. Thus, in fact, the CNN and the enhance model are not completely independent in classification result. But the enhance model still has difference with the CNN in classification result, which means $1 - E_2^b \leq P_2$ but $P_2 - (1 - E_2^b) \leq P_1 - (1 - E_1^b)$. It is also proved by the experiment results in Table 5. As shown in Table 5, when α is set to 0.1, we get $1 - E_2^b \geq 1 - E_1^b$.

In experiment part, some baseline methods, including SVM, MNB, NB-SVM, RNN and BRNN, are compared as the enhance model, eventually NB-SVM is selected.

4. Experiments

4.1. Datasets and baseline methods

The proposed method are evaluated on four benchmark datasets. Summary statistics of the datasets are in Table 1. We describe each dataset in detail below:

- MR[2]: Movie reviews dataset where each instance is a sentence [28]. The objective is to classify each movie review to either positive or negative.

[2] http://cs.cornell.edu/people/pabo/movie-review-data/.

Table 1
Statistics of the four datasets used in this paper. N counts the number of instances, $dist(+,-)$ lists the class distribution, l represents the average sentence length and *test* is the size of test data, we use 10-fold cross-validation (CV) in this paper.

Dataset	N	$dist(+,-)$	l	*test*
MR	10,662	(0.5, 0.5)	18	CV
CR	3780	(0.64, 0.36)	17	CV
SUBJ	10,000	(0.5, 0.5)	21	CV
MPQA	10,462	(0.32, 0.68)	3	CV

- CR[3]: Annotated customer reviews of 14 products obtained from Amazon [10]. The task is to classify each customer review into positive or negative.
- SUBJ: Subjectivity dataset where the task is to classify a sentence as being subjective or objective [27].
- MPQA[4]: Phrase-level opinion polarity detection subtask of the MPQA dataset [40].

3W-CNN is compared with the following baseline methods.

- NB-SVM and MNB. Naive Bayes SVM and Multinomial Naive Bayes with uni and bigram features [39].
- RAE and MV-RecNN. Recursive AutoEncoder [33] and Matrix-vector Recursive Neural Network [32].
- CNN [15]. Convolutional Neural Network for sentence modeling.
- Paragraph-Vec. Paragraph Vector [17] is an unsupervised model to learn distributed representations of words and paragraphs. We use a public implementation[5] and apply logistic regression on top of the pre-trained paragraph vectors for prediction.
- cBoW. Continuous Bag-of-Words model. We use max pooling as the global pooling mechanism to compose a phrase/sentence vector from a set of word vectors.
- RNN, BRNN. Recurrent Neural Networks [11] and Bidirectional Recurrent Neural Networks [16].
- GrConv. Gated Recursive Convolutional Neural Network [3].

4.2. Performance measures and statistical tests

To evaluate the performance of the proposed method, classification accuracy and time complexity are calculated. Further, the statistical tests are used to verify the significance of the experimental results. Wilcoxon test is used in this paper as per the studies of Liu [21,22]. The process of Wilcoxon test is described in the following. Let Acc^b_M and $Acc^b_{M'}$ respectively denote the accuracy of methods M and M' on dataset b, and b denotes MR, CR, SUBJ or MPQA. In this study, M and M' denote the different methods used for sentiment classification, including the above baseline methods and the proposed method in this work. Let $d^b_{MM'}$ denote the difference between Acc^b_M and $Acc^b_{M'}$, i.e.,

$$d^b_{MM'} = Acc^b_M - Acc^b_{M'} \tag{4}$$

The experiment is constructed on four datasets, thus there are four $d^b_{MM'}$ for methods M and M', i.e.,$d^1_{MM'}, d^2_{MM'}, \cdots, d^4_{MM'}$. Let $|d^{(1)}_{MM'}| \le |d^{(2)}_{MM'}| \le |d^{(3)}_{MM'}| \le |d^{(4)}_{MM'}|$ denote the ranking of the absolute values of $d^1_{MM'}, d^2_{MM'}, \cdots, d^4_{MM'}$ from smallest to greatest. Let $p^b_{MM'}$ denote the ranking position of $d^b_{MM'}$, i.e., $p^b_{MM'} \in \{1,2,3,4\}$, and $p^b_{MM'} \neq p^{b'}_{MM'}$, if $b \neq b'$. Let $I^b_{MM'}$ denote an indication variable of $d^b_{MM'}$. As per the Wilcoxon test, the value of $I^b_{MM'}$ can be calculated as follows:

$$I^b_{MM'} = \begin{cases} 1 & d^b_{MM'} > 0 \\ 0.5 & d^b_{MM'} = 0 \\ 0 & d^b_{MM'} < 0 \end{cases} \tag{5}$$

Let $R^+_{MM'}$ denote a score indicating the degree that method M outperforms method M', and $R^-_{MM'}$ denote a score indicating the degree that method M' outperforms method M. As per the Wilcoxon test, $R^+_{MM'}$ and $R^-_{MM'}$ can be calculated as follows:

$$\begin{aligned} R^+_{MM'} &= \sum_{b=1}^{4} p^b_{MM'} I^b_{MM'}, \\ R^-_{MM'} &= \sum_{b=1}^{4} p^b_{MM'} (1 - I^b_{MM'}) \end{aligned} \tag{6}$$

$R^+_{MM'}$ and $R^-_{MM'}$ are also used as performance measures in our experiments.

[3] http://cs.uic.edu/~liub/FBS/sentiment-analysis.html.
[4] http://www.cs.pitt.edu/mpqa.
[5] http://github.com/mesnilgr/iclr15.

Table 2
Classification accuracy of the baseline methods and the proposed 3W-CNN.

Model	MR	CR	SUBJ	MPQA
MNB	79.0	80.0	93.6	86.3
RAE	77.7	-	-	86.4
MV-RecNN	79.0	-	-	-
PV	74.8	78.1	90.5	74.2
cBoW	77.2	79.9	91.3	86.4
RNN	77.2	82.3	93.7	90.1
BRNN	**82.3**	82.6	**94.2**	**90.3**
GrConv	76.3	81.3	89.5	84.5
NB-SVM	79.4	81.8	93.2	86.3
CNN	81.5	85.0	93.4	89.6
3W-CNN	**82.3**	**85.8**	93.5	**90.3**

Table 3
Wilcoxon test results for baseline methods and 3W-CNN.

Models	R^+	R^-
3W-CNN vs. BRNN	5.5	4.5
3W-CNN vs. CNN	10	0
3W-CNN vs. RNN	9	1
3W-CNN vs. NBSVM	10	0
CNN vs. BRNN	4	6
CNN vs. RNN	7	3
CNN vs. NBSVM	10	0
BRNN vs. RNN	10	0
BRNN vs. NBSVM	10	0
RNN vs. NBSVM	7	3

4.3. Experimental settings

In CNN we use: filter windows of 3, 4, 5 with 100 feature maps each, dropout rate of 0.5, l_2 constraint of 3, and mini-batch size of 50.

The parameter α, which is defined in Section 3.3, is set to 0.1, and it will be explained in Section 4.4. We do not perform any dataset-specific tuning other than early stopping on test sets. For datasets without standard test set we use 10-fold cross-validation to test. Training is done through stochastic gradient descent over shuffled mini-batches.

We use the publicly available word2vec vectors that were trained on 100 billion words from Google News. The vectors have dimensionality of 300 and were trained using the continuous bag-of-words architecture [25].

4.4. Results and discussions

The classification accuracy of the baseline methods and the proposed 3W-CNN are shown in Table 2. 3W-CNN achieves comparable results with BRNN and outperforms other baseline methods by a large margin. 3W-CNN achieves the best results on CR, MR and MPQA datasets, and slightly worse result than BRNN on SUBJ dataset. According to Table 2 and Table 3, we can draw the conclusion that, in terms of classification accuracy, the ranking of RNN, BRNN, NB-SVM, CNN and 3W-CNN is 3W-CNN > BRNN > CNN > RNN > NB-SVM, where " > " denotes "outperform".

3W-CNN is a combination of CNN and NB-SVM, and 3W-CNN achieves higher accuracy than both CNN and NB-SVM on all the datasets as shown in Table 2, which indicates that the combination mechanism is effective. If CNN or NB-SVM achieves a good result on a dataset, then 3W-CNN can achieve a better result. For example, CNN can achieve 85.0% on CR dataset which was the best, and 3W-CNN can achieve 85.8%. On the other hand, 3W-CNN is also limited by CNN and NB-SVM. 3W-CNN can not perform very well if CNN and NB-SVM do not achieve good results.

Now we study how α affects the performance of 3W-CNN and explain why α is set to 0.1. As illustrated in Fig. 3 which is experimented on MR dataset, as α increases, the classification accuracy of both CNN and NB-SVM on boundary data generally improves, but CNN improves faster. When α reaches a value between 0.4 and 0.5, the classification accuracy of CNN is equal to that of NB-SVM on boundary data, and it is exactly when the classification accuracy of CNN is equal to that of 3W-CNN on entire test data. 3W-CNN performs best when α equals 0.1.

In order to verify the effectiveness of the confidence function defined in Eq. (2), the classification accuracy of CNN on the boundary data which is selected by the confidence function ($\alpha = 0.1$) is illustrated in Table 4. We can see that the classification accuracy of CNN on the boundary data is much lower than that on the rest data, which proves that the con-

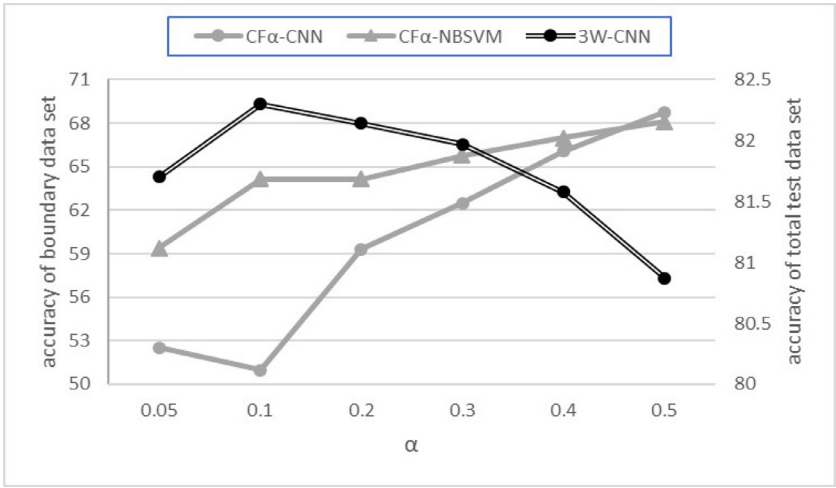

Fig. 3. How α affects the accuracy of CNN and NB-SVM on boundary data and the accuracy of 3W-CNN on entire test data. $CF_\alpha - CNN$ and $CF_\alpha - NBSVM$ denote the accuracy of CNN and NB-SVM on boundary data, respectively. $3W - CNN$ denotes the accuracy of 3W-CNN on entire test data.

Table 4
Classification accuracy of CNN on the boundary data and the rest data. The boundary data is selected by the confidence function defined in Eq. (2) with alpha of 0.1. Acc_b denotes the accuracy on the boundary data, Acc_r denotes the accuracy on the rest data.

CNN	MR	CR	SUBJ	MPQA
Acc_b	54.5	53.2	64.7	57.3
Acc_r	84.4	88.2	96.3	93.1

Table 5
The values of D when several baseline models are used as the enhance model.

D	MR	CR	SUBJ	MPQA
SVM	27.4	39.0	20.7	29.9
MNB	12.6	17.2	8.1	11.4
NB-SVM	16.3	19.8	7.1	11.9
RNN	10.8	9.2	8.9	9.8
BRNN	7.8	9.5	8.7	10.0

fidence function has the ability to distinguish the quality of the classification result. The classification accuracy of CNN on the boundary data is very low so that we can use the enhance model to improve the classification performance of this part.

The values of D (defined in Section 3.3) when several baseline models are used as the enhance model are shown in Table 5. Table 5 shows that the value of D is relatively low when neural network models are used as the enhance model. The reason is that the model structure and classification process of the neural network models (e.g. RNN, BRNN) are similar to those of CNN, while those of the non-neural network models (e.g. SVM, MNB, NB-SVM) are much different. As described in Section 3.3, the higher D, the better enhance model. Therefore, we use non-neural network models as the enhance model.

The classification accuracy of several non-neural network models on the boundary data and the rest data is shown in Table 6. In terms of the classification accuracy on the boundary data, the enhance model needs to be high, at least higher than the CNN. Therefore, we compare the results in Table 6 and select NB-SVM as the enhance model. We also study if D affects the selection of the enhance model, and find that the difference between the Acc_b of NB-SVM and CNN is positively correlated with the D, which means the higher D, the better 3W-CNN performs on the boundary data. Now we explain why 3W-CNN performs not so well on SUBJ dataset. When NB-SVM is used as the enhance model, the value of D on SUBJ dataset is low (7.1%), so that the promotion of Acc_b is not obvious (68.8% vs 64.7%).

3W-CNN improves the classification accuracy compared to CNN, but 3W-CNN needs to train two models and perform prediction process twice. It seems that 3W-CNN performs worse than CNN in terms of time complexity. However, as shown in Table 7, there is only a minimal increase both in the training time and the predicting time of 3W-CNN. The reason is that NB-SVM, as the enhance model, can complete the training and predicting process in much shorter time than CNN.

Table 6
Classification accuracy of the non-neural network models ($\alpha = 0.1$). Acc_b denotes the accuracy on the boundary data. Acc_r denotes the accuracy on the rest data.

$Acc_b \backslash Acc_r$	MR	CR	SUBJ	MPQA
SVM	50.1\66.2	46.8\65.0	53.2\74.6	55.0\71.5
MNB	59.7\81.2	60.8\82.1	69.9\96.2	60.7\89.1
NB-SVM	61.3\81.1	63.7\83.3	68.8\94.6	62.3\88.7
CNN	54.5\84.4	53.2\88.2	64.7\96.3	57.3\93.1

Table 7
Time complexity of 3W-CNN compared with CNN and NB-SVM on MR dataset. It is the average value of ten experiments on the computer with one GeForce GTX 1080 Ti.

Model	Training time	Predicting time
CNN	33.67s	10.25s
NB-SVM	0.32s	0.09s
3W-CNN	34.15s	10.41s

5. Conclusion

Inspired by the methodology of three-way decisions, we proposed a sentiment classification model 3W-CNN which is optimized from CNN with NB-SVM. We design a confidence function to divide the outputs of CNN into two parts, the boundary data and the rest data. NB-SVM is further used to improve the classification performance of the boundary data. The experimental results show that 3W-CNN has a good performance on four benchmark datasets. In fact, 3W-CNN can be regarded as an ensemble framework for any two models, if you can find an effective confidence function to construct the confidence divider. Meanwhile, the two models also need to have complementary property on classification ability. In the future work, we will try to find other effective confidence functions, and apply the framework on other models to study its effectiveness and applicability.

Acknowledgments

The work described in this paper has been supported by the National Key R&D Program of China (Grant No. 213), the Natural Science Foundation of China (Grant No. 61673301), the Major Project of Ministry of Public Security (Grant No. 20170004), the Open Research Funds of State Key Laboratory for Novel Software Technology (Grant No. KFKT2017B22).

References

[1] P.F. Brown, P.V. deSouza, R.L. Mercer, V.J.D. Pietra, J.C. Lai, Class-based n-gram models of natural language, Computational Linguistics 18 (4) (1992) 467–479.
[2] H. Cho, S. Kim, J. Lee, J.-S. Lee, Data-driven integration of multiple sentiment dictionaries for lexicon-based sentiment classification of product reviews, Knowl Based Syst 71 (2014) 61–71.
[3] K. Cho, B. Van Merriënboer, D. Bahdanau, Y. Bengio, On the properties of neural machine translation: encoder-decoder approaches, Syntax, Semantics and Structure in Statistical Translation (2014).
[4] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, P. Kuksa, Natural language processing (almost) from scratch, Journal of Machine Learning Research 12 (Aug) (2011) 2493–2537.
[5] X.W. Ding, B. Liu, P.S. Yu, A holistic lexicon-based approach to opinion mining, in: Proceedings of the 2008 International Conference on Web Search and Data Mining, 2008, pp. 231–240.
[6] Z.P. Fan, Y.J. Che, Z.Y. Chen, Product sales forecasting using online reviews and historical sales data: a method combining the bass model and sentiment analysis, J Bus Res 74 (2017) 90–100.
[7] M. Franco-Salvador, F.L. Cruz, J.A. Troyano, P. Rosso, Cross-domain polarity classification using a knowledge-enhanced meta-classifier, Knowl Based Syst 86 (2015) 46–56.
[8] H. Fujita, T.R. Li, Y.Y. Yao, Advances in three-way decisions and granular computing, Knowl Based Syst 91 (2016) 1–3.
[9] M.S. Hajmohammadi, R. Ibrahim, A. Selamat, H. Fujita, Combination of active learning and self-training for cross-lingual sentiment classification with density analysis of unlabelled samples, Inf Sci (Ny) 317 (C) (2015) 67–77.
[10] M. Hu, B. Liu, Mining and summarizing customer reviews, in: Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2004, pp. 168–177.
[11] O. Irsoy, C. Cardie, Opinion mining with deep recurrent neural networks., in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014, pp. 720–728.
[12] N. Kalchbrenner, E. Grefenstette, P. Blunsom, D. Kartsaklis, A convolutional neural network for modelling sentences, in: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, 2014, pp. 212–217.
[13] G. Katz, N. Ofek, B. Shapira, Consent: context-based sentiment analysis, Knowl Based Syst 84 (2015) 162–178.
[14] S.M. Kim, E. Hovy, Automatic detection of opinion bearing words and sentences, in: Companion Volume to the Proceedings of the International Joint Conference on Natural Language Processing (IJCNLP), 8, 2005.
[15] Y. Kim, Convolutional neural networks for sentence classification, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014, pp. 1746–1751.

[16] S.W. Lai, L.H. Xu, K. Liu, J. Zhao, Recurrent convolutional neural networks for text classification, in: Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, 2015, pp. 2267–2273.
[17] Q. Le, T. Mikolov, Distributed representations of sentences and documents, in: Proceedings of the 31st International Conference on Machine Learning (ICML-14), 2014, pp. 1188–1196.
[18] H.X. Li, L.B. Zhang, B. Huang, X.Z. Zhou, Sequential three-way decision and granulation for cost-sensitive face recognition, Knowl Based Syst 91 (2016) 241–251.
[19] H.X. Li, L.B. Zhang, X.Z. Zhou, B. Huang, Cost-sensitive sequential three-way decision modeling using a deep neural network, Int. J. Approximate Reasoning 85 (2017) 68–78.
[20] D. Liu, Y.Y. Yao, T. Li, Three-way investment decisions with decision-theoretic rough sets, International Journal of Computational Intelligence Systems 4 (1) (2011) 66–74.
[21] Y. Liu, J.W. Bi, Z.P. Fan, A method for multi-class sentiment classification based on an improved one-vs-one (OVO) strategy and the support vector machine (SVM) algorithm, Inf Sci (Ny) 394–395 (2017) 38–52.
[22] Y. Liu, J.W. Bi, Z.P. Fan, Multi-class sentiment classification: the experimental comparisons of feature selection and machine learning algorithms, Expert Syst Appl 80 (2017) 323–339.
[23] Y. Liu, J.W. Bi, Z.P. Fan, Ranking products through online reviews: a method based on sentiment analysis technique and intuitionistic fuzzy set theory, Information Fusion 36 (2017) 149–161.
[24] S. Maldonado, G. Peters, R. Weber, Credit scoring using three-way decisions with probabilistic rough sets, Inf Sci (Ny) (2018), doi:10.1016/j.ins.2018.08.001.
[25] T. Mikolov, I. Sutskever, K. Chen, G.S. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, in: Advances in Neural Information Processing Systems, 2013, pp. 3111–3119.
[26] F. Min, Z.H. Zhang, Z.W. Jie, S.R. Ping, Frequent pattern discovery with tri-partition alphabets, Inf Sci (Ny) (2018), doi:10.1016/j.ins.2018.04.013.
[27] B. Pang, L. Lee, A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts, in: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, 2004, p. 271.
[28] B. Pang, L. Lee, Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales, in: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, 2005, pp. 115–124.
[29] B. Pang, L. Lee, S. Vaithyanathan, Thumbs up?: Sentiment classification using machine learning techniques, in: Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing (EMNLP '02), 2002, pp. 79–86.
[30] Y.F. Ren, R. Wang, D.H. Ji, A topic-enhanced word embedding for twitter sentiment classification, Inf Sci (Ny) 369 (2016) 188–198.
[31] Y.L. Shen, X.D. He, J.F. Gao, L. Deng, G. Mesnil, Learning semantic representations using convolutional neural networks for web search, in: Proceedings of the 23rd International Conference on World Wide Web, ACM, 2014, pp. 373–374.
[32] R. Socher, B. Huval, C.D. Manning, A.Y. Ng, Semantic compositionality through recursive matrix-vector spaces, in: Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, 2012, pp. 1201–1211.
[33] R. Socher, J. Pennington, E.H. Huang, A.Y. Ng, C.D. Manning, Semi-supervised recursive autoencoders for predicting sentiment distributions, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2011, pp. 151–161.
[34] D.Y. Tang, B. Qin, T. Liu, Document modeling with gated recurrent neural network for sentiment classification., in: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2015, pp. 1422–1432.
[35] D.Y. Tang, F.R. Wei, B. Qin, N. Yang, T. Liu, M. Zhou, Sentiment embeddings with applications to sentiment analysis, IEEE Trans Knowl Data Eng 28 (2) (2016) 496–509.
[36] M. Thelwall, K. Buckley, G. Paltoglou, Sentiment strength detection for the social web, J Assoc Inf Sci Technol 63 (1) (2012) 163–173.
[37] P.D. Turney, Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews, in: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, 2002, pp. 417–424.
[38] L. Wang, F.J. Ren, D.Q. Miao, Multi-label emotion recognition of weblog sentence based on bayesian networks, IEEJ Trans. Electr. Electron. Eng. 11 (2) (2016) 178–184.
[39] S. Wang, C.D. Manning, Baselines and bigrams: Simple, good sentiment and topic classification, in: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, 2012, pp. 90–94.
[40] J. Wiebe, T. Wilson, C. Cardie, Annotating expressions of opinions and emotions in language, Lang Resour Eval 39 (2) (2005) 165–210.
[41] J.T. Yao, N. Azam, Web-based medical decision support systems for three-way medical decision making with game-theoretic rough sets, IEEE Trans. Fuzzy Syst. 23 (1) (2015) 3–15.
[42] Y.Y. Yao, Three-way decisions with probabilistic rough sets, Inf Sci (Ny) 180 (3) (2010) 341–353.
[43] Y.Y. Yao, An outline of a theory of three-way decisions, in: Rough Sets and Current Trends in Computing, Springer, 2012, pp. 1–17.
[44] Y.Y. Yao, S. Wang, X.F. Deng, Constructing shadowed sets and three-way approximations of fuzzy sets, Inf Sci (Ny) 412–413 (Supplement C) (2017) 132–153.
[45] S.W.T. Yih, X. He, C. Meek, Semantic parsing for single-relation question answering, in: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, 2014, pp. 643–648.
[46] H. Yu, X.C. Wang, G.Y. Wang, X.H. Zeng, An active three-way clustering method via low-rank matrices for multi-view data, Inf Sci (Ny) (2018), doi:10.1016/j.ins.2018.03.009.
[47] H.R. Zhang, F. Min, Three-way recommender systems based on random forests, Knowl Based Syst 91 (2016) 275–286.
[48] Y. Zhang, J.T. Yao, Game theoretic approach to shadowed sets: a three-way tradeoff perspective, Inf Sci (Ny) (2018), doi:10.1016/j.ins.2018.07.058.
[49] H. Zhao, Z.D. Lu, P. Poupart, Self-adaptive hierarchical sentence model, in: Proceedings of the 24th International Conference on Artificial Intelligence, 2015, pp. 4069–4076.
[50] B. Zhou, Y.Y. Yao, J.G. Luo, Cost-sensitive three-way email spam filtering, J Intell Inf Syst 42 (1) (2014) 19–45.