Knowledge-Based Systems 107 (2016) 289–300

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Mining opinion summarizations using convolutional neural networks in Chinese microblogging systems

Qiudan Li[a], Zhipeng Jin[a,b], Can Wang[a,b], Daniel Dajun Zeng[a,b,c,*]

[a] The State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing, 100190 China
[b] University of Chinese Academy of Sciences, Beijing, China
[c] Department of Management Information Systems, University of Arizona, Tucson, AZ 85721, USA

ARTICLE INFO

Article history:
Received 5 October 2015
Revised 9 June 2016
Accepted 13 June 2016
Available online 15 June 2016

Keywords:
Chinese microblogging systems
Hot topics
Convolutional neural network
Opinion summarization
Maximal marginal relevance

ABSTRACT

Chinese microblogging is an increasingly popular social media platform. Accurately summarizing representative opinions from microblogs can increase understanding of the semantics of opinions. The unique challenges of Chinese opinion summarization in microblogging systems are automatic learning of important features and selection of representative sentences. Deep-learning methods can automatically discover multiple levels of representations from raw data instead of requiring manual engineering. However, there have been very few systematic studies on sentiment analysis of Chinese hot topics using deep-learning methods. Based on the latest deep-learning research, in this paper, we propose a convolutional neural network (CNN)-based opinion summarization method for Chinese microblogging systems. The model first applies CNN to automatically mine useful features and perform sentiment analysis; then, by making good use of the obtained sentiment features, the semantic relationships among features are computed according to a hybrid ranking function; and finally, representative opinion sentences that are semantically related to the features are extracted using Maximal Marginal Relevance, which meets "relevant novelty" requirements. Experimental results on two real-world datasets verify the efficacy of the proposed model.

1. Introduction

With the rapid development of Web2.0, microblogging has played an increasingly important role in our daily life. For example, the leading social media platform Sina-weibo[1] provides an unprecedented and simple way for people to create, distribute and discover Chinese-language content, interact with others and stay connected with the world. As of September 30, 2015, it had 212 million monthly active users, a 48% increase over the previous year, and 100 million daily active users, indicating a rise of 30% compared with 2014 [1]. On the platform, users can freely express their opinions and exchange ideas in real time on topics ranging from newly released products to recent events. By making good use of such opinion data, one can gain better insights into public opinions. Corporations can better understand users' information needs about their products and then formulate customer-driven marketing plans. Management departments can track people's reactions to both events and policy decisions, which will lead to better-informed decisions and more effective policy implementation.

However, the large amount of microblogs makes it extremely difficult for users to keep track of public opinions. Therefore, it is crucial to mine and summarize opinions from microblogs automatically. Ideally, we expect the opinion summarization model to automatically mine the important features, which is useful in predicting the polarity of microblogs, and then to identify the succinct and representative opinion sentence to generate summaries with meaningful semantics. For example, when a disaster occurs, people often express their hopes and wishes using sentences such as *praying for your safety, hope you get back safe, expect a miracle!!* or *sending blessings*. This suggests a method that first automatically correlates the important features with positive sentiment and then selects a set of sentences that are the most semantically related to the features as the representative opinions. The discovered features may include punctuation (e.g. *!!*), emoticons, words (e.g. *safe, blessings, miracle*), phrases and sentence structure (e.g. *hope…safe*). Based on these features, we may better understand the semantics of wishes and hopes using the opinion sentence (e.g. *praying for your safety*). The characteristics of emergency and natural disaster events are sudden and urgent, which brings a challenge of understanding the semantics of opinions accurately and responding to public opinions very quickly. Therefore, automatic learning of important features and selection of representative sentences are necessary.

* Corresponding author.
E-mail addresses: qiudan.li@ia.ac.cn (Q. Li), jinzhipeng2013@ia.ac.cn (Z. Jin), wangcan2015@ia.ac.cn (C. Wang), zeng@email.arizona.edu (D.D. Zeng).
[1] http://ir.weibo.com/phoenix.zhtml?c=253076amp;p=irol-homeprofile.

http://dx.doi.org/10.1016/j.knosys.2016.06.017
0950-7051/

Deep-learning methods are representation-learning methods that can automatically discover multiple levels of representations from raw data. They are considered as promising methods for various tasks including topic classification, sentiment analysis, question answering and language translation [2]. The most popular kinds of neural network based models for sentiment analysis are recursive neural network (RNN) [3,4] and convolutional neural network (CNN) [5,6]. Although these models have been applied to movie reviews and Twitter messages to demonstrate good performance, there are still significant gaps; little systematic research has been done on performing sentiment analysis in Chinese microblogging systems using deep learning. Users could gain more insights from the models if they can be applied to other domains and other social media platforms. Thus, in this research, we apply deep-learning methods to Chinese microblogging systems to predict the polarity of microblogs about hot topics.

The aim of the opinion sentence is to help users better understand the semantics of an opinion. Recent research on automatic labeling of topics [7] suggested that summarization algorithms can generate better topic labels. Among those summarization algorithms, Maximal Marginal Relevance (MMR) [8] and TextRank [9] are popular. The former is a relevance-based ranking algorithm, which avoids redundancy in the documents used for generating a summary, while the latter is a graph-based summarization method. Identifying important features that can be used to perform semantic association or construct semantic graphs is the key factor of these models. Inspired by these good ideas, we propose a two-stage method to select the opinion sentence by combining MMR and TR, which takes ranking score of CNN, semantic relevance and conciseness into consideration.

In this paper, we propose a CNN-based opinion summarization model to mine and summarize representative opinions for hot topics, such as products and emergency/disaster events, in Chinese microblogging systems. The model first predicts the polarity of a microblog based on a convolutional network. Then, the important features automatically discovered by CNN for each kind of category are used to construct a feature graph based on TextRank. Finally, sentences that are most semantically related to the top-ranked features are selected as the representative opinions via MMR. To evaluate the effectiveness of the model, we perform experiments on two real-world datasets from Chinese microblogging systems. Experimental results verify the efficacy of the proposed model.

This paper makes the following contributions. 1) The model provides a unified framework that enables accurate prediction of sentiment polarity and summarization of opinions for hot topics in Chinese microblogging systems. 2) The convolutional neural network in our model not only improves the accuracy of sentiment prediction but also identifies important sentiment features automatically. 3) The selected opinions take both semantic relevance and novelty into consideration by integrating MMR with CNN and TextRank, which can better understand the semantics of opinion based on both the feature importance and the feature relationships. 4) The experimental results on real data from Chinese microblogging systems show that the model can help keep track of public opinions for hot events accurately.

The remainder of this paper is organized as follows. We begin with a discussion of related work in the areas of opinion mining in social media, opinion summary generation techniques in Section 2. Then, the proposed framework is described in Section 3. A detailed description of the opinion summarization model is introduced in Section 4. Section 5 presents an empirical study as well as evaluation results and discussion. Finally, we conclude this paper in Section 6.

2. Related work

We introduce the related work on opinion mining in social media and opinion summary generation techniques in this section.

2.1. Opinion mining in social media

Most opinion mining methods in social media can be roughly divided into supervised [10–14] and unsupervised methods [15, 16]. The supervised learning methods first design a feature extractor that transforms the raw data into a feature vector; then a standard classifier such as support vector machine (SVM) or naïve Bayes (NB) is trained from manually labeled training data. Barbosa and Feng [10] performed sentiment classification on tweets using a two-stage SVM classifier that includes features selection and different label sources combination. The traditional unsupervised methods often rely on a predefined sentiment lexicon to predict the sentiment polarity of a document. However, in social media, due to its distinct features such as short length, new expressions, fast-evolving patterns and semantic ambiguity in different domains, it is difficult and time consuming to obtain labels and define a sentiment lexicon to include words from different domains [15]. Thus, some efforts have been made to explore the emotional signals in sentiment analysis. Hu et al. [15] incorporates both post- and word-level sentiment-related contextual information into a unified unsupervised framework. Brody and Diakopoulos [17] verified that the length of a word is strongly associated with its sentiment. Li et al. [18] employed the sentiment knowledge learned from one domain to perform sentiment analysis on another domain. Miao et al. [19] proposed a fine-grained opinion mining method which automatically mines product features and opinions from multiple review sources. Zhang et al. [20] proposed a rule-based approach that addresses the unique challenges posed by Chinese sentiment analysis. Zhou and Chaovalit [21] aimed to enhance polarity mining with ontology by providing detailed topic-specific information. Narock et al. [22] proposed an algorithm that enhances the computation of semantic similarity with polarity mining techniques. Zhu et al. [16] introduced another kind of unsupervised method that establishes the duality between sentiment clustering and co-clustering of a tripartite graph using a unified tri-clustering framework. Zhao et al. [23] performed sentiment analysis for Chinese microblogging systems employing an emoticon-based method. In addition, Ravi and Ravi [24] presented a comprehensive review of the research from 2002 to 2014 on various aspects of sentiment analysis including tasks, approaches and applications. The authors found that some of the intelligent techniques such as random forest, online learning algorithms, radial basis function neural network (RBFNN), etc., have not been exploited exhaustively, and there is much work still to be done in this promising area. Serrano-Guerrero et al. [25] conducted a comprehensive assessment of 15 web services which include functionalities related to sentiment analysis and uncovered their capabilities under different circumstances. For example, AlchemyAPI and Semantria are found to be the most common tools, which classify both short and long texts and predict their corresponding polarity ratings. For longer texts, the classification performance of the SentimentAnalyzer is good especially when there are no neutral documents. These useful findings will provide enough information for users to decide the most appropriate tool for their interests.

Recently, deep learning has been a hot research topic successfully applied to sentiment analysis. The key aspect of deep learning is that it automatically learns features from raw data using a general-purpose learning procedure, instead of features designed by human engineers [2]. Because little engineering by hand is required, it can easily discover interesting patterns from large-scale social media data. Among many deep-learning methods, recursive

neural network (RNN) and convolutional network (CNN) are very popular. Socher et al. [3] proposed a matrix-vector recursive neural network model that learns compositional vector representations for phrases and sentences of arbitrary length. Socher et al. [4] proposed a Recursive Neural Tensor Network (RNTN) model, which first represents a phrase using word vectors and a parse tree and then computes vectors for higher nodes in the tree. Liang et al. [26] proposed a recursive neural network (RNN) and polarity transition-based approach to perform sentiment analysis for Chinese microblogging systems that explores the feasibility of predicting polarity for Chinese weibo by deep learning. Regarding convolutional networks for sentiment analysis, dos Santos and Gatti [5] proposed Character to Sentence Convolutional Neural Network (CharSCNN) that applies convolutional layers to extract character-level and sentence-level features. Compared with recursive neural network (RNN), the method does not need any input about the syntactic structure of the sentence and can easily explore the richness of word embeddings produced by unsupervised pre-training [27]. Kim [6] described a series of experiments with convolutional neural networks built on top of word2vec for sentence-level classification tasks. The results show that learning task-specific vectors through fine-tuning offers further gains in performance. A general architecture was further designed to allow for the use of both task-specific and static vectors. Kalchbrenner et al. [28] presented a dynamic convolutional neural network (DCNN) for semantic modeling of sentences, which uses the dynamic k-max pooling operator as a nonlinear subsampling function. The induced feature graph can capture word relations of varying size. The model achieves good performance on Twitter sentiment prediction. However, there have been very few systematic studies on sentiment analysis of Chinese content using deep-learning methods.

2.2. Opinion summary generation techniques

For effective opinion summary generation, opinion summarization, automatic labeling, topic discovery and association mining are the fundamental approaches widely used.

Most opinion summarization work focuses on mining opinion summaries from product reviews [29,30], movie reviews [31] or hotel reviews [32]. They analyzed sentiment on fine-grained features or aspects of a product. Meng et al. [33] studied the problem of opinion summarization for entities in Twitter and suggested a unified optimization framework that generates an opinion summary by integrating three dimensions such as topic, opinion and insight, as well as other factors (e.g. redundancy and language styles). Based on the characteristics of hot event data in Chinese microblogging systems, this paper mainly focuses on understanding the semantics of opinion from the perspective of overall sentiment.

To better understand the semantics of a topic, some automatic labeling methods have been proposed. These methods aim to select a set of words that best describe the semantics of the terms involved in a topic. Recent research [7] suggest that summarization algorithms, such as MMR [8] and TextRank [9], can generate better topic labels. Aletras and Stevenson [34] introduced an unsupervised graph-based method, which uses PageRank to weigh the words in the graph and score the candidate labels. Chang et al. [35] proposed a Twitter context summarization approach, leveraging pairwise and global user influence models to improve text-based summarization.

In addition to automatic labeling methods, topic discovery and association mining are often useful to provide insights into the semantics of events. In [36], we proposed a popular topic detection approach based on a user interest-based model. In [37], dynamic association among Twitter topics was further identified. Zhang et al. [38] proposed a unified framework that combines the author-topic model with social network analysis to discover user and topic communities simultaneously in Twitter. The mined topics can better interpret the community.

Inspired by the above ideas and based on the ranking score of features obtained by CNN, MMR and TR are further integrated to select opinion sentences, which can be used to better understand the semantics of opinions.

Most of the existing approaches discussed above either perform sentiment analysis on movie review and Twitter messages or report summarization results from raw data. Little work has been done on mining opinion summarizations from hot event data in Chinese microblogging systems. Applying the deep-learning methods to hot event data in Chinese microblogging systems can both extend the application area of novel methods and increase understanding of these events. One distinct feature of hot social events in Chinese microblogging data is that it often contains information like punctuation, emoticons, words, phrases and sentence structure. Identifying such information is essential for understanding the semantics of an opinion. Our framework focuses on extracting representative opinions for hot topics in Chinese microblogging systems based on the accurate prediction of polarity using deep-learning methods. Thus our model is able to further discover meaningful and representative opinions, which distinguishes our work from the existing sentiment analysis and summarization works.

3. A framework for automatic representative opinion summary generation

This section first introduces the formal definition of mining and summarizing representative opinions in Chinese microblogging systems, then presents the proposed framework.

Given a set of microblogs on a hot event, the framework aims to produce a representative opinion summary, denoted as $O = \{O_{Pos}, O_{Neg}\}$, where O_{Pos} and O_{Neg} represent the summary for positive sentiment and negative sentiment, respectively. Using O_{Pos} as an example, it is a $\langle\{f\}, \{p\}\rangle$pair, where {f} is a set of useful features for identifying the positive polarity of the microblogs, and {p} is a set of representative opinion sentences semantically related to the features and extracted from positive microblogs. The representation of O_{Neg} is similar to that of O_{Pos}.

Fig. 1 depicts the proposed framework of automatically generating an opinion summary from a collection of microblogs related to a hot event. It consists of three stages: data representation, polarity prediction and feature relationship mining and opinion sentence extraction.

First, it's necessary to produce an effective representation for each microblog. Word2vec is a deep-learning-inspired method that attempts to understand meaning and semantic relationships among words. It learns vector representations of words using continuous bag-of-words (CBOW) and Skip-gram. Therefore, at the data representation stage, due to the good performance of capturing syntactic and semantic information [27], we adopt word2vec to learn domain-specific vectors for word and sentence.

At the sentiment prediction stage, given the low dimensional representation of sentences or raw sentences without feature engineering, CNN [6] is applied to automatically mine useful features and perform sentiment analysis on the Chinese microblogging system. The model first uses a convolution operation to produce feature maps. Then the resolution of the feature maps is reduced by a pooling operation, and finally, the obtained useful local features are fed to a fully connected softmax layer to predict the sentiment label of the microblog.

Based on the prediction results, we build positive and negative datasets by selecting the posts and features together with assigned sentiment labels. Making good use of the automatically learned

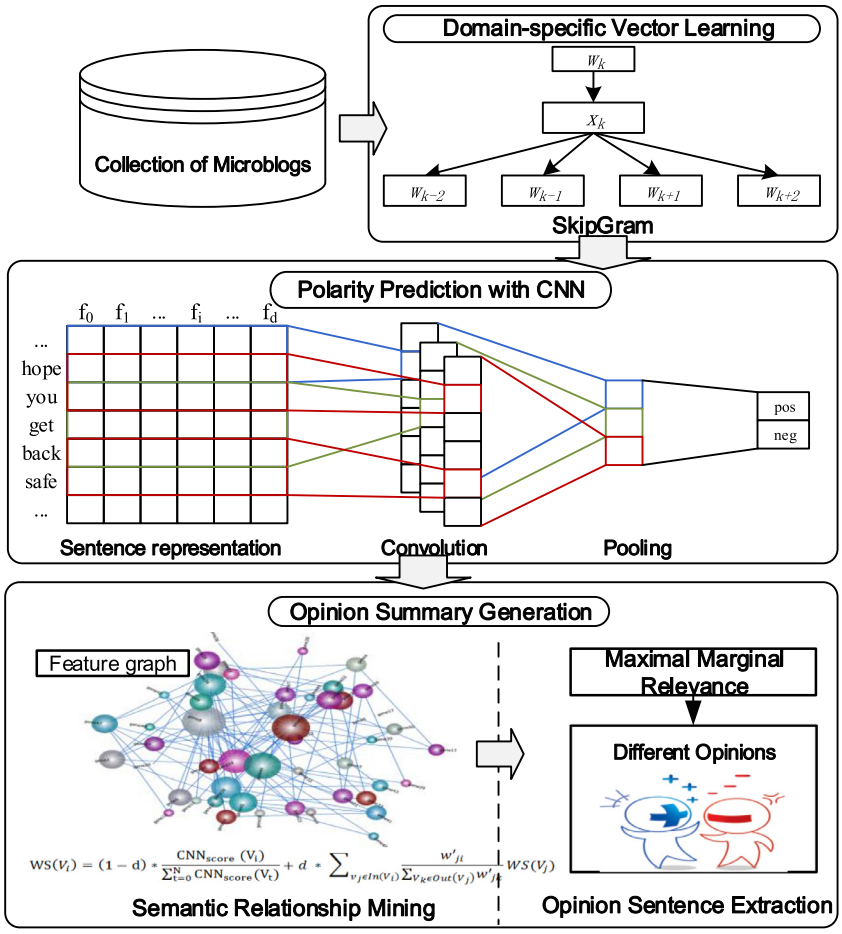

Fig. 1. A framework for automatic opinion summary generation in Chinese microblogging systems.

features, the semantic relationships among features are then computed according to a hybrid ranking function. Finally, representative opinion sentences that are semantically related to the features are extracted using the MMR approach. Below, we illustrate the details of the proposed framework.

4. Mining and summarizing representative opinions in Chinese microblogging systems

4.1. Learning domain-specific vector representation for word and sentence

It has been demonstrated that lower dimensional vector representation can capture syntactic and semantic information [27]. By mapping word vectors into a vector space, semantically similar words will have similar vector representation. The skip-gram model, which is a state-of-the-art word-embedding method, is used to learn the dense word vector representation. Given a sequence of training words $w_1, w_2, \ldots, w_N$ in the domain-specific corpus, the model aims to maximize the average log probability:

$$\frac{1}{N}\sum_{m=1}^{N}\left[\sum_{j=-k_1}^{k_1}\log P\left(w_{m+j}|w_m\right)\right]$$

where k_1 is the size of the training window and $P(w_{m+j}|w_m)$ denotes the probability of correctly predicting the word w_{m+j}, in which w_m represents the middle word in the training window.

Using this method, the word w_i corresponds to a k-dimensional word vector $x_i \in R^k$. Then, a vector representation of a sentence consisting of n words is as follows:

$$x_{1:n} = x_1 \oplus x_2 \oplus \ldots \oplus x_n$$

where $\oplus$ is the concatenation operator. $x_{i:i+j}$ refers to the concatenation of words $x_i, x_{i+1}, \ldots, x_{i+j}$.

4.2. Polarity prediction

Borrowing the sentence classification approach for Twitter data using CNN [6], we automatically learn the important features for identifying the polarity of a microblog and then perform sentiment analysis on hot events in Chinese microblogging systems. This consists of convolution operation, pooling operation and label prediction.

4.2.1. Convolution operation

Based on the above vector representation of a sentence, the one-dimensional convolution operation performs dot product between the filter of the convolution $wf \in R^{hk}$ (h is the window size) and each h-gram in the sentence $x_{1:\,n}$ to obtain another sequence; specifically, the filter wf is applied to each possible window of words in the sentence $\{x_{1:h},\ x_{2:h+1}, \ldots, x_{n-h+1:\,n}\}$ to produce a feature map $c = [c_1, c_2, \ldots, c_{n-h+1}]$, $c \epsilon R^{n-h+1}$. Each element c_i in c can be calculated as:

$$c_i = \sigma\left(wf \cdot x_{i:i+h-1} + b\right)$$

where $b \in R$ is a bias term. According to the range of the index i, the type of convolution operation can be classified into narrow and wide; the former requires that $n \geq h$, while the latter does

not have requirements on n or h. For the latter one, out-of-range input values x_i where $i < 1$ or $i > n$ are set to be zero. The trained weights in the filter wf can be used as a linguistic feature detector to recognize a specific class of phrases. The lengths of these phrases are not larger than h, where h is the width of the filter. Applying the weights wf in a wide convolution ensures that all weights in the filter reach the entire sentence, including the words at the margins. This is particularly important when h is set to a relatively large value such as 8 or 10. In addition, a wide convolution guarantees that the application of the filter h to the input sentence n always produces a valid non-empty result c, independently of the width h and the sentence length n. Therefore, we pad zeros at the beginning and end of the sentence.

In the following equation, hyperbolic tangent, sigmoid and ReLU (rectified linear unit) can be used as a nonlinear function σ. Compared with sigmoid and tanh functions, ReLU doesn't have a gradient vanishing problem, which can be chosen as the activation function to train deep networks efficiently. It is defined as:

$$f(t) = \max(0,\ t)$$

4.2.2. Pooling operation

The pooling operation aims to reduce the resolution of feature maps by applying a pooling function to several units in a local region of a size determined by a parameter called pooling size. The pooling operation units will serve as generalizations over the features obtained from convolution. These generalizations are invariant to small variations in location, since they are spatially localized in frequency. A max approach is used as the pooling function to each convolution feature map independently and captures the most useful local features:

$$\hat{c} = \max\{c\}$$

$\hat{c}$ is a feature corresponding to the particular filter. This pooling scheme naturally deals with variable sentence lengths. The model uses multiple filters (with varying window sizes) to produce multiple features. This fixed-sized global feature vector can then be fed to the classical affine network layers.

4.2.3. Label prediction

The features produced by the pooling operation form the penultimate layer $\hat{U} = \{\hat{c}_1, \hat{c}_2, \ldots, \hat{c}_p\}$ (p is the number of filters). They are then passed to a fully connected softmax layer whose output is the probability distribution over labels:

$$P\left(y = j|\hat{U}\right) = \frac{\exp\left(\hat{U} \cdot \hat{w}_j\right)}{\sum_{j=1}^{K} \exp\left(\hat{U} \cdot \hat{w}_j\right)}$$

where K is the number of labels and $\hat{w}$ denotes the weight matrix of the fully connected layer.

4.3. Opinion summary generation

When the sentiment label is assigned to each post, positive and negative datasets are built by selecting the posts and features together with the associated labels. Then, a set of features that best describe each dataset are identified by Semantic Relationship Mining, and finally, representative opinion sentences are selected using the MMR criterion, which strives to reduce redundancy while maintaining semantic relevance to the features.

4.3.1. Semantic relationship mining

Selection of representative features for positive and negative data can provide useful cues for accurate opinion sentence extraction. By performing the max pooling operation on all feature maps, each word in the specific phrase feature receives a vote, which is normalized as:

$$S_w = \frac{V_w - min}{max - min}$$

where V_w is the number of votes word w gets, max and min are the maximum number and minimum number of votes in the sentence containing word w. By normalization, the score of each word in the sentence lies between [0,1]. For the collection of positive and negative data obtained by CNN classification, we compute the cumulative score of each feature word, respectively, which indicates the sentiment importance of the feature word. Then, a feature graph-based ranking function, which combines the importance of features mined by CNN [6] and TextRank [9], is developed to further identify important features and their associations. The intuition behind the model is that when one feature links to another one, it means that the feature casts a vote for the other feature; then, the importance of a feature depends on the number of votes a feature gets and the importance of the feature casting the vote. Formally, let $G = (V, E)$ be a feature graph, where V and E are the set of vertices represented by features and the set of edges represented by association between features, respectively. $In(V_i)$ is the set of features that points to feature V_i, $Out(V_i)$ is the set of features that feature V_i points to, co-occurrence relation is used to express the association of features and the association strength between feature V_i and V_j is denoted by w'_{ij}. The score of a feature V_i is computed as follows:

$$WS(V_i) = (1-d) * \frac{CNN_{score}(V_i)}{\sum_{t=0}^{N} CNN_{score}(V_t)} + d * \sum_{v_j \in In(V_i)} \frac{w'_{ji}}{\sum_{V_k \in Out(V_j)} w'_{jk}} WS(V_j)$$

where d is a damping factor that can be set between 0 and 1. $CNN_{SCORE}(V_i)$ is the score value computed by CNN and N is the number of vertices in the graph.

Through the above iterative process, the features that are identified as important ones by CNN and also recommended by the related and highly influential features are selected as important features. These important mined related features tend to help further extract representative opinion sentences.

4.3.2. Opinion sentence extraction

Once we obtain important features for each sentiment category, we extract the representative opinion sentences to generate an opinion summary. Sentences are selected according to a combined criterion of relevance and novelty, where the former measures the semantic relevance between the sentence and a given feature set, and the latter measures the dissimilarity between the opinion sentence being considered and the previously selected one. MMR criterion [8] provides an efficient and unified way to consider these two factors and thus is adopted to extract the opinion sentence. It is defined as follows:

$$MMR = Arg \max_{D_i \in R \setminus S} \left[\lambda Sim_1(D_i, F) - (1-\lambda) \max_{d_j \in S} Sim_2\left(D_i, D_j\right) \right]$$

where D is a sentence collection, F is a feature set, S is the subset of already selected sentences in D and R\S is the set of unselected sentences in D. Given D and F, the opinion sentence extraction model will select opinion sentences with high marginal relevance; namely, the selected sentence is relevant to the feature set F and contains minimal similarity to previously selected sentences in S. Sim_1 is the similarity metric between a sentence and a feature, while Sim_2 is the similarity metric between two sentences.

Table 1
Description of emergency events-related data.

The name of the topic	Time period	#of microblogs with sentiment	#of positive microblogs	#of negative microblogs
亚航失联 (AirAsia plane lost)	2014.12.28–2015.01.12	4207	2214	1993
台湾空难 (Taiwan plane crash)	2015.02.04–2015.02.19	877	521	356
德国空难 (Germanwings air crash)	2015.03.24–2015.04.07	304	110	194

Table 2
Description of natural disaster events-related data.

The name of the topic	Time period	#of microblogs with sentiment	#of positive microblogs	#of negative microblogs
康定地震 (Kangding earthquake)	2014.11.22–2015.12.07	1021	849	172
鲁甸地震 (Ludian earthquake)	2014.08.03–2014.08.18	1301	1031	270
尼泊尔地震 (Nepal earthquake)	2015.04.25–2015.05.01	157	135	22

5. Experimental analysis

In this section, we conduct comprehensive and systematic analyses to evaluate the proposed automatic representative opinion summary generation model. The process of collecting the dataset and the evaluation metrics are presented first. Then, we explain the purpose of our experiments in detail. Next, the performance of our model is compared with the results of three other models; results of the comparison verify the efficacy of the proposed model. Finally, a case study and visual analytics of sentiment is used to further illustrate the summarization results of the proposed model.

5.1. Dataset description

To thoroughly examine the performance of the proposed model from different perspectives, two kinds of real-world Chinese microblogging datasets are used to conduct our experiments. 1) The first dataset comes from COAE2014, which is the 6th Chinese Opinion Analysis Evaluation, and involves product-related topics such as cars, cell phones, jewelry, and milk from a Chinese microblogging platform. Three annotators manually annotated the microblogs as positive, negative or neutral. We obtain 9465 microblogs, 5501 positive microblogs and 3964 negative microblogs. 2) We use Sinaweibo API to collect an event-related dataset, which consists of emergency events and natural disaster events. We chose these topics for the following reasons. Social media such as Twitter and Chinese microblogs has been successfully and widely used in many different emergency and disaster scenarios [39]. Accurately understanding public sentiment and representative opinions towards these events is especially important for improved emergency or disaster management. After obtaining the data, ensemble imbalanced classification and a knowledge expansion-based spam filtering algorithm [40] are used to filter out noisy data such as advertisements, robot-generated microblogs, or off-topic microblogs. Furthermore, we use the above mentioned annotation method to annotate the 5388 emergency events-related microblogs and 2479 disaster event-related microblogs. The statistics of the datasets are shown in Tables 1 and 2.

5.2. Evaluation metrics

Accuracy, precision, recall and AUC measures, which are commonly adopted evaluation metrics in opinion mining [25], are utilized in our experiments to evaluate the performance of the proposed model. Let TP, FP, TN and FN refer to the number of predictions falling into TruePositive, FalsePositive, TrueNegative and FalseNegative categories in the confusion matrix. Then, the accuracy is defined as:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

For positive sentiment, the precision and recall are defined as follows:

$$Pos.Pre = \frac{TP}{TP + FP},\ Pos.Rec = \frac{TP}{TP + FN}$$

Similarly, for negative sentiment, the precision and recall are defined as follows:

$$Neg.Pre = \frac{TN}{TN + FN},\ Neg.Rec = \frac{TN}{TN + FP}$$

To balance the performance of two sentiment types, we average the results to obtain the overall precision and recall:

$$Precision = \frac{Pos.Pre + Neg.Pre}{2},\ Recall = \frac{Pos.Rec + Neg.Rec}{2}$$

AUC [41], which stands for "area under the ROC curve," refers to the amount of area under the receiver operating characteristic curve. It is a value between 0.5 and 1. In classification problems, AUC's threshold independence makes it uniquely qualified for model selection. The model with a higher AUC will be the better one regardless of the threshold setting.

5.3. Purpose of our experiments

The focus of the experiments is to test the efficacy of each component in the proposed opinion mining and summarizing model. The following questions are addressed.

- Question 1: Does CNN accurately predict the sentiment polarity for product or emergency/disaster-related events in Chinese Microblogging Systems?
- Question 2: In practical application scenarios, it is usually necessary to make a quick and accurate prediction for the current event based on historical events in the same category. Does the features' self-learning ability in CNN help provide such an accurate prediction among different events?
- Question 3: Is the Semantic Relationship Mining approach effective or intuitively explanatory for mining associations among features and identifying important features?
- Question 4: Do the identified important features provide cues that help extract more understandable opinion sentences?
- Question 5: Does the model extract semantically relevant and not redundant opinions, which helps better understand the public's opinions?

To answer these questions, we proceed as follows. For the first question, we conduct a set of experiments on the COAE2014 data to evaluate the CNN model performance for product data. For the second question, for emergency events, we predict the sentiment about the GermanWings air crash based on the AirAsia lost plane event and Taiwan plane crash event. For natural disaster events, the sentiment about the Nepal earthquake is predicted based on

Table 3
Comparison of different sentiment prediction models for COAE2014 data.

Model	Accuracy	Precision	Recall	AUC
CNN	**86.01**	**85.77**	**85.77**	**0.936**
SVM	84.34	84.02	83.71	0.912
Random forest	83.90	83.70	83.05	0.921
Logistics regression	83.08	82.68	82.49	0.899

Table 4
Prediction results for emergency events.

Model	Accuracy	Precision	Recall	AUC
CNN	**89.47**	**89.79**	**87.23**	**0.936**
SVM	86.84	87.18	83.99	0.909
Random forest	87.83	87.78	85.54	0.929
Logistics regression	82.57	81.04	81.62	0.836

the Kangding earthquake and Ludian earthquake. In addition, comparison between static term vector and non-static word vector is presented, which further explains the accurate prediction. For the third, fourth and fifth questions, a case study is presented to demonstrate how the mined features affect the results of summarization and why the extracted opinions are important in a real scenario. Finally, visual analytics of sentiment is used to show how the proposed model can offer users a vivid representation to help them better understand the semantics of public opinion and, more importantly, how it will enable users to navigate through the data, interact with the system and provide timely feedback.

5.4. Experimental results and discussion

In this section, to examine the usefulness and effectiveness of CNN for sentiment prediction in Chinese Weibo, we compare it with three traditional classification methods: support vector machines (SVM), logistic regression and random forest. Unigram and bigram are used to construct the feature space for the comparison methods. All the baselines are implemented based on WEKA [42], which is a widely used tool for data mining tasks. The best results are shown in bold in the Tables 3-6.

In all the experiments, some parameter settings of the proposed model are as follows. During the word vector training phase, we set the length of the word vector to 300 dimensions and set the window size to 5. During the CNN training phase, three groups of convolution templates with filter window sizes of 3, 4 and 5 are used; the number of each template is 100. In addition, the batch size of training samples each time is 50, and the ReLU method is used as the activation function. During semantic association mining, parameter d is set to 0.5 and the window size to 4. Finally, in the opinion summary generation stage, λ is set to 0.6.

5.4.1. Evaluation of sentiment prediction for COAE2014 data

We apply the CNN method to the COAE 2014 data. Table 3 shows the performance comparison of CNN over three traditional classification methods.

We use 10-fold cross validation as a means to avoid bias in the experiments. This involves dividing the data into 10 groups, iteratively using combinations of nine distinct groups to learn the model and the remaining group to validate the performance of the model, and averaging the predictive accuracy score.

As can be seen from Table 3, CNN performs better than the traditional classification methods in all four evaluation measures. The method can predict the sentiment polarity with an accuracy of 86% on product-related Chinese microblogging data. The improved performance suggests that, on one hand, word2vec obtains the low dimensional vector representation of each word, which makes the input sentence vector of CNN more meaningful; on the other hand, the features' self-learning ability in CNN helps automatically identify the effective features for sentiment prediction.

In order to further investigate if the mined feature words are meaningful for sentiment prediction, the number of features in the regions selected by each convolution template during the max pooling process is accumulated and we obtain the feature distribution of each microblog. Fig. 2 shows the feature distribution of a positive microblog and a negative microblog.

From the left part of Fig. 2, we can observe that for the positive microblog 三星尹恩惠用的手机, 顿时觉得好高端 (*Samsung's mobile phone used by Yun Eunhea, immediately feel very advanced*), the important features identified by CNN are *very* and *advanced*, which can help judge the sentiment polarity of the microblog. For the negative microblog 最怕用三星的手机, 电池很不给力! (*The most afraid of using Samsung's mobile phone, the battery is not very good*), *not, good* and *!* are selected as the important features, and CNN correctly predicts the label of the microblog as negative by successfully identifying the negative word *not*.

5.4.2. Evaluation of sentiment prediction for emergency and disaster event data

Emergency and natural disaster events are naturally characterized as sudden and urgent. In addition, social media has become a platform for people to express and spread their opinions about such events in real time, which brings a challenge of responding to public opinions very quickly. Therefore, it is extremely important to provide a quick and accurate prediction for the current event based on historical events of the same category. For emergency events, we predict the sentiment about the GermanWings air crash based on the AirAsia lost plane event and Taiwan plane crash event. For natural disaster events, sentiment about the Nepal earthquake is predicted based on the Kangding earthquake and Ludian earthquake. The prediction results are shown in Tables 4 and 5.

Table 4 shows that sentiment prediction using CNN yields the best results in all four measures on emergency events. For natural disaster events, CNN still performs far better than SVM and logistic

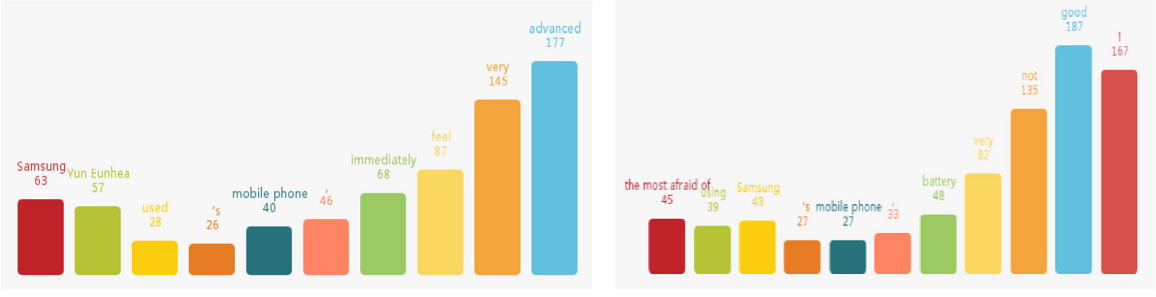

Fig. 2. The feature distribution of a positive microblog and a negative microblog.

Table 5
Prediction results for natural disaster events.

Model	Accuracy	Precision	Recall	AUC
CNN	**89.81**	78.95	**94.06**	**0.970**
SVM	87.26	74.31	81.18	0.926
Random forest	89.80	**80.48**	73.15	0.907
Logistics regression	87.26	73.57	73.57	0.928

Table 6
Prediction results with static vectors and non-static vectors.

		Accuracy	Precision	Recall	AUC
Emergency events	Static	87.83	87.10	86.33	0.921
	Non-static	**89.47**	**87.23**	**89.79**	**0.936**
Natural disaster events	Static	84.08	70.40	79.33	0.918
	Non-static	**89.81**	**78.95**	**94.06**	**0.970**

regression. Compared with CNN, random forest achieves better results on precision, nearly equal results on accuracy, but worse results on recall and AUC. The results on both events datasets confirm that the sentiment label of the newly happened event data can be predicted by using past events of the same category as training data. In addition, traditional classification methods using n-grams as features can only extract features of words, while CNN can find more specific sentence patterns, which makes CNN achieve better performance.

5.4.3. Prediction with static vectors and non-static vectors

The input of CNN consists of sentence vectors, which are composed of word vectors trained by word2vec. However, these word vectors are static and not adapted to the domain-specific task since they are learned by unsupervised learning. Non-static word vectors refer to the fine-tuning vectors that are updated based on the loss error of classification during the training, which can capture more meaningful representations for domain-specific data [6]. Both static word vectors and non-static word vectors are used to evaluate their effects on four evaluation measures in our experiments. The results are shown in Table 6.

It is interesting to note that non-static vectors appear to have a positive effect on prediction results; results on all four measures are improved significantly. This is because we use the supervised learning method to update the word vector and obtain a more meaningful word vector representation, similar to sentiment orientation

In addition, we conduct experiments to further test the effectiveness of the learned static vectors and non-static vectors. We choose some seed words, use cosine similarity to measure the similarity between two word vectors, and obtain the most similar words for each seed word. Table 7 summarizes the top 4 similar words of the chosen seed words.

As shown in Table 7, the non-static word vectors can reveal more semantic information. By dynamically updating the word vectors during the training process, semantically related words are clustered together. If we take an emoticon *[one's face is covered with tears]* as an example, based on the representation of static vectors, only *[tears]* is found to be relevant to it; however, non-static vectors identify that heartache, *[sad]* and *[tears]* are all related to the original emoticon expressing sad feelings. This example further confirms that sentiment prediction accuracy can be improved by effectively discovering latent semantic representation of words, which is useful for emotion recognition. Moreover, the non-static word vectors also provide a valuable resource for automatically building a domain-specific sentiment knowledge base.

Table 7
Semantically related words discovered by static vectors and non-static vectors.

Seed words	Static vectors	Non-static vectors
希望 hope	有 have	天堂 heaven
	我们 we	[许愿] [wish]
	都 all	安好 well
	祈求 ask for	上帝保佑 God bless!
平安 safe	祈祷 pray	祈祷 pray
	点赞 praise	阿弥陀佛 Amitabha
	好 good	平安无事 safe and sound
	他们 they	哀悼 grieve
[泪流满面] [one's face is covered with tears]	[泪] [tears]	[围观] [surround to watch]
	研究 investigate	心痛 heartache
	总算 finally	[悲伤] [sad]
	分 separate	[泪] [tears]
活着 alive	身边 at one's side	安好 well
	活下去 live	平安无事 safe and sound
	生命 life	点赞 praise
	所有人 everyone	当下 present

5.4.4. A case study

In an emergency and disaster event analysis environment, besides offering users sentiment labels for each post, users often need to further understand the semantics of opinions via representative opinions. Generally speaking, words often provide important cue information, while sentences can provide more complete semantic information. Based on this intuition, we propose a two-step opinion generation method: first, a semantic relationship mining model is used to identify important features; secondly, based on the obtained features, MMR is applied to extract representative sentences to generate an opinion summary. The extracted representative sentences meet "relevant novelty" requirements, namely, maximized semantic relevance and minimized redundancy.

The purpose of this experiment is to use emergency events as a case to test the effectiveness of the opinion summary generation model. As shown in previous sections, CNN can achieve 89.47% prediction accuracy.

Based on the Semantic Relationship Mining model described in Section 4.3, we get the important features of positive and negative emotions. The identified top20 important features are shown in Table 8.

It can be seen from Table 8 that these features can better describe the semantic information for the positive and negative datasets. In order to further evaluate the effectiveness of these features, we compare the proposed hybrid method integrating CNN and TextRank (CNN&TextRank for short) with the following methods: TextRank method, frequency-based and combined frequency-based and TextRank (Frequency&TextRank). Based on the results of accurate sentiment classification of data collections, the frequency-based method selects the features with high term frequency from the obtained positive and negative data. The hybrid pattern of frequency-based and TextRank is similar to that of CNN and TextRank discussed in Section 4.3, replacing the CNN item score with the normalized word frequency score. The results obtained by these three methods are shown in Table 9.

As shown in Table 9, for the positive dataset, the important features mined by the three methods are almost the same; most of the features include terms indicating good wishes such as *pray, safe, cherish* and *strong*. For the negative data, the hybrid approach Frequency&TextRank achieved the best results, because emoticons with strong negative emotional expressions such as *[tears], [sad]* or *[heartbroken]* are ranked higher.

We further compare the two hybrid approaches, CNN&TextRank and Frequency&TextRank. For the positive dataset, CNN&TextRank not only finds 70% of the important features that can also be mined by Frequency&TextRank, but also identifies strong positive

Table 8
Top 20 important features of positive and negative dataset discovered by combining CNN and TextRank.

	Top20 important features
Positive dataset	[蜡烛]([candle]), 安息(rest in peace), 逝者(deceased), 平安(safe), 祈祷(pray), 祈福(bless), 默哀(stand in silent tribute), 一路(all the way), 他们(they), 奇迹(miracle), 珍惜(cherish), 保佑(god bless), 遇难者(victim), 安好(well), 乘客(passenger), 坚强(strong), 哀悼(grieve), 归来(come back home), 活着(alive), [飞机]([plane])
Negative dataset	[泪]([tears]), [悲伤]([sad]), 坐飞机(by plane), 以后(later), [伤心]([heart broken]), [衰]([scared]), [吃惊]([surprised]), 真的(really), 这么(so), 任性(capricious), 卧槽(damn), 可怕(terrible), 出事(accident), 坠毁(crash), [生病]([sick]), 不能(can't), 安全(safety), [震惊]([shocked]), 现在(now), 怎么(what's wrong)

Table 9
Top 20 important features by TextRank, Frequency-based and Method combines Frequency-based and TextRank.

		Important features (top20)
TextRank	Positive dataset	[蜡烛] ([candle]), 逝者(deceased), 安息(rest in peace), 祈祷(pray), 平安(safe), 祈福(bless), 复兴(TransAsia), 珍惜(cherish), 默哀(stand in silent tribute), 世界(world), 乘客(passenger), 大陆(mainland), 人们(people), 遇难者(victim), 搜救(rescue), 活着(alive), 一路(all the way), 遇难(die in an accident), 保佑(god bless), 同胞(compatriot)
	Negative dataset	坐飞机(by plane), 以后(later), 出事(accident), [泪] ([tears]), 不能(can't), 坠毁(crash), 知道(know), 真是(indeed), 中国(China), 马航(MH), 今年(this year), 印尼(Indonesia), [悲伤] ([sad]), 不是(not), 这种(this kind), 航空公司(airline), 飞行员(pilot), 专家(expert), 现在(now), 可能(possible)
Frequency	Positive dataset	[蜡烛] ([candle]), 安息(rest in peace), 逝者(deceased), 希望(hope), 祈祷(pray), 平安(safe), 祈福(bless), 台湾(Taiwan), 亚航(AirAsia), 飞机(plane), 失联(lost), 发生(happen), 他们(they), 空难(air crash), 生命(life), 默哀(stand in silent tribute), 坚强(strong), 奇迹(miracle), 珍惜(cherish), 你们(you)
	Negative dataset	飞机(plane), [泪] ([tears]), 失联(lost), 亚航(AirAsia), 坐飞机(by plane), 怎么(what's wrong), [悲伤] ([sad]), 客机(passenger plane), 为什么(why), 以后(later), 这么(so), 现在(now), [衰]([scared]), [伤心] ([heart broken]), 台湾(Taiwan), 安全(safety), 可怕(terrible), 这样(like this), [吃惊] ([surprised]), 出事(accident)
Combining frequency and TextRank (Frequency&TextRank)	Positive dataset	[蜡烛] ([candle]), 安息(rest in peace), 逝者(deceased), 希望(hope), 祈祷(pray), 平安(safe), 祈福(bless), 默哀(stand in silent tribute), 珍惜(cherish), 坚强(strong), 奇迹(miracle), 复兴(TransAsia), 遇难者(victim), 节哀(condolences), 乘客(passenger), 他们(they), [心]([heart]), 家属(family member), 灾难(disaster), 一路(all the way)
	Negative dataset	[泪] ([tears]), 坐飞机(by plane), [悲伤] ([sad]), 以后(later), 现在(now), 怎么(what's wrong), 出事(accident), [伤心] ([heart broken]), [衰] ([scared]), 真的(really), 可怕(terrible), 今年(this year), 为什么(why), 这么(so), 马航(MH), 坠毁(crash), [吃惊] ([surprised]), 不能(can't), 真是(indeed), 不敢(don't dare)

emotional features such as *god bless, well* or *come back home*. For the negative data, 75% of the important features obtained by CNN&TextRank are consistent with that of Frequency&TextRank. More strong negative emotional emoticons including *[tears], [sad], [heartbroken]* and *[scared]* rank higher and interesting and meaningful negative expressions such as *damn, [sick]* and *[shocked]* are also mined. Therefore, in general, CNN&TextRank is able to identify more significant and meaningful positive and negative features. This may due to the following reasons. On one hand, CNN can obtain the emotional semantic score of each word, which can't be mined by frequency-based methods; on the other hand, the feature graph constructed by TextRank helps identify more important features, enhancing the interpretability of the results.

We perform some experiments with window sizes of 2, 3, 4 and 5 to examine the impact of different window size on the result, and the experimental results show that the impact is small. This may be due to the characteristics of short text in Sina-weibo. Finally, the window size is set to 4.

Parameter d adjusts the impact of CNN score and TextRank on the final score of the features. Intuitively, decreasing d will increase the influence of the CNN score on the extracted features, and vice versa. The experimental results with the settings $d=0.2$, $d=0.5$ and $d=0.8$ are shown in Table 10. From Table 10, we can get the following findings. For the positive data, the important features mined by the proposed hybrid approach with different settings are basically consistent. For the negative data, when parameter d is set to 0.5 and 0.8, 70% of the features identified by the approach are similar, and strong negative emotional emoticons rank higher when d is set to 0.5. When parameter d is 0.2 and 0.5, 80% of the features identified by the approach are similar, and the ranking orders of the features are basically identical. The observed findings suggest that the quality of important features can be improved by increasing the weight of CNN's score. To balance the effects of CNN and TextRank, d is set to 0.5.

Based on the above identified important features and their ranking scores, each clause in a microblog, separated by punctuation, is given a cumulative score. MMR is then used to obtain the final opinion summary with minimum redundancy. Table 11 shows the results for emergency events.

As can be seen from Table 11, the positive data mainly discusses two aspects. One is mourning for the victims, including *[candle] [candle] [candle], Let deceased rest in peace, stand in silent tribute*; the other is hoping for a miracle to happen, such as *pray, hope for miracle to happen, I wish they come back home alive*. The negative data also talks about two topics, expressing sad feelings, such as *[tears] [tears] [tears]* or *[sad] [sad] [sad]*, or fear and complaints about the event, such as *it's too terrible, accident happens all the time* or *[surprised]*. As we can see, the phrase/clause extracted by the MMR method are more meaningful and diverse, which is useful for gaining insights into the deep semantic information of public opinions.

5.5. Visual analytics of sentiment

Based on the above results for sentiment prediction and opinion summary, visual analytics will show users a vivid representation to help them better understand the semantics of public opinion. More importantly, it will enable users to navigate through the data, interact with the system and provide timely feedback, which also provides an efficient way for the proposed model to learn from users' feedback and thus improve the sentiment analysis performance.

When a user submits a topic term "plane crash," the CNN-based model is first built to predict the polarity of microblogs on this topic. The distribution of polarity between positive sentiment and negative sentiment from December 2014 to April 2015 is computed as shown in Fig. 3; it will provide users an intuitive and overall

Table 10
Top 20 important features mined by CNN&TextRank with different parameter settings d.

d = 0.2	Positive dataset	[蜡烛] ([candle]), 安息(rest in peace), 逝者(deceased), 平安(safe), 祈祷(pray), 祈福(bless), 默哀(stand in silent tribute), 一路(all the way), 奇迹(miracle), 保佑(god bless), 一切(everything), 遇难者(victim), 安好(well), 珍惜(cherish), 坚强(strong), 哀悼(grieve), 归来(come back home), 乘客(passenger), [飞机]([plane]), 活着(alive)
	Negative dataset	[泪] ([tears]), [悲伤] ([sad]), [伤心] ([heart broken]), 坐飞机(by plane), 这么(so), [衰] ([scared]), [吃惊] ([surprised]), 以后(later), 任性(capricious), 怎么(what's wrong), 卧槽(damn), 真的(really), [生病], 可怕(terrible), [震惊] ([shocked]), 还是(still), 不敢(don't dare), 这样(like this), 危险(dangerous), 坠毁(crash)
d = 0.5	Positive dataset	[蜡烛] ([candle]), 安息(rest in peace), 逝者(deceased), 平安(safe), 祈祷(pray), 祈福(bless), 默哀(stand in silent tribute), 一路(all the way), 他们(they), 奇迹(miracle), 遇难者(victim), 保佑(god bless), 珍惜(cherish), 安好(well), 乘客(passenger), 坚强(strong), 哀悼(grieve), 归来(come back home), 活着(alive), [飞机]([plane])
	Negative dataset	[泪] ([tears]), [悲伤] ([sad]), 坐飞机(by plane), 以后(later), [伤心] ([heart broken]), [衰] ([scared]), [吃惊] ([surprised]), 真的(really), 这么(so), 任性(capricious), 卧槽(damn), 可怕(terrible), 出事(accident), 坠毁(crash), [生病]([sick]), 不能(can't), 安全(safety), [震惊] ([shocked]), 现在(now), 怎么(what's wrong)
d = 0.8	Positive dataset	[蜡烛] ([candle]), 安息(rest in peace), 逝者(deceased), 平安(safe), 祈祷(pray), 祈福(bless), 默哀(stand in silent tribute), 珍惜(cherish), 乘客(passenger), 一路(all the way), 遇难者(victim), 奇迹(miracle), 复兴(TransAsia), 人们(people), 保佑(god bless), 家属(family member), 活着(alive), 灾难(disaster), 安好(well), 世界(world)
	Negative dataset	[泪] ([tears]), 坐飞机(by plane), 以后(later), [悲伤] ([sad]), 真的(really), 出事(accident), [衰] ([scared]), [伤心] ([heart broken]), 坠毁(crash), 不能(can't), 真是(indeed), 马航(MH), 今年(this year), 现在(now), 可怕(terrible), [吃惊] ([surprised]), 看到(see), 悲剧(tragedy), 危险(dangerous), 卧槽(damn)

Table 11
Opinion summary for emergency event.

Positive opinion summary		Negative opinion summary	
[蜡烛][蜡烛][蜡烛]	[Candle] [Candle] [Candle]	[泪][泪][泪]	[Tears] [Tears] [Tears]
逝者安息	Let deceased rest in peace	[悲伤][悲伤][悲伤]	[Sad] [Sad] [Sad]
祈祷	Pray	[衰]	[Scared]
默哀	Stand in silent tribute	[吃惊]	[Surprised]
一路走好	All the way walk good	卧槽	Damn
保佑	God bless	太可怕了	It's too terrible
珍惜生命	Cherish life	我真的真的痛心	I'm really really saddened
希望出现奇迹	Hope for miracle to happen	老是出事	Accident happens all the time
[飞机][飞机][飞机]	[Plane][Plane][Plane]	能不能安全点	Can't be safe?
祝他们还活着回来	I wish they come back home alive	坠毁	Crash

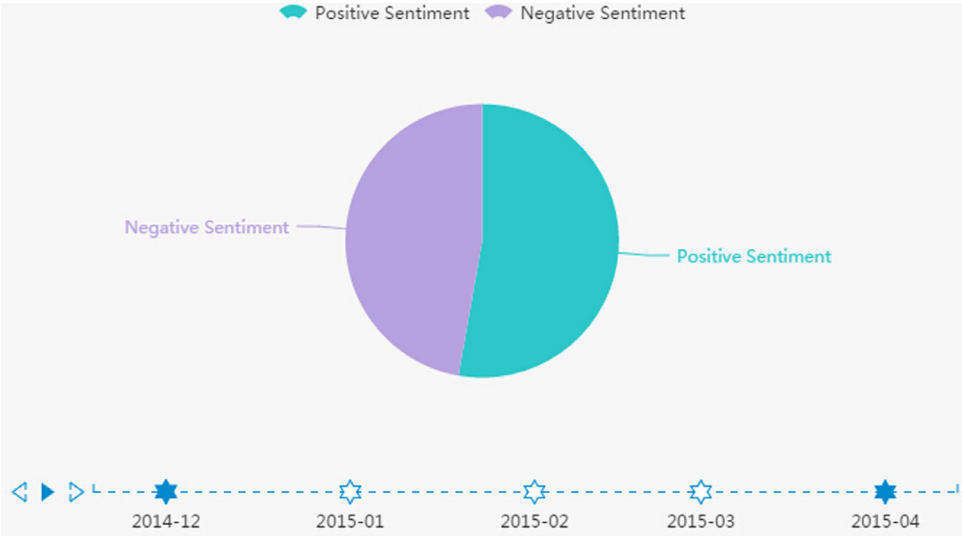

Fig. 3. Distribution of sentiment for emergency event.

sentiment comparison. The more accurate the prediction model is, the more realistic the distribution is.

Figs. 4 and 5 show the generated word clouds from positive and negative microblogs when a user clicks on the sentiment distribution. The size of the word represents the importance of the word, which provides valuable cues for emotional analysis.

Based on the word cloud, the visualization of the opinion summary is generated as shown in Figs. 6 and 7. The size of the sector represents the importance of the opinion. From these figures, we can observe that the reduplication of an emoticon plays an important role in expressing strong feelings, for instance, *[tears] [tears] [tears]* and *[sad] [sad] [sad]* are used to show that users are very upset and sad. *[candle][candle][candle]* and *[plane][plane][plane]* express the strong feeling of mourning for the victims and the hope that people come back soon. Besides the reduplication of emoticons, the model also finds some diverse and interesting patterns, such as significant word or sentence structure. For example, the mined sentence structure including "Let……," "Hope……" and "Wish……" is usually used for offering expectations and wishes. Tag question such as *can't be safe* are bitter complaints about the event. These found opinion summaries provide an efficient way for users to better understand the semantics of the public's opinions.

When the user clicks on an opinion sentence of interest, the related microblogs together with their publication time, authors,

Fig. 4. Positive word cloud.

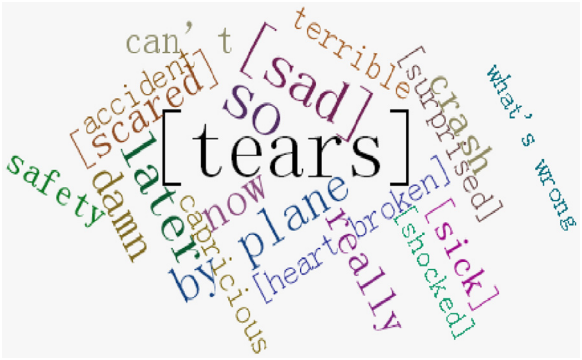

Fig. 5. Negative word cloud.

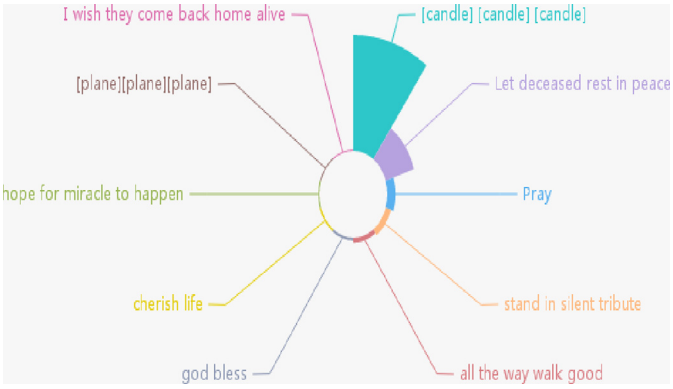

Fig. 6. Positive opinion summary.

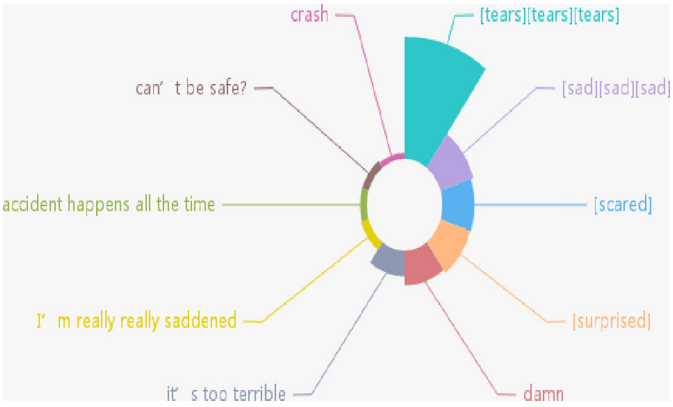

Fig. 7. Negative opinion summary.

reply number, retweet number and sentiment label will appear in a list. The user can interact with the system. For example, if he/she is not satisfied with the prediction result of the microblog, he/she can provide timely feedback by adjusting the sentiment label in the system; then the revised label data will be collected and used for training the model in the future.

From the above analysis, we can see that visual analytics of sentiment can help users understand the public's opinions comprehensively, thus, it is useful for management departments to quickly respond to a hot event.

6. Conclusions and future work

Chinese microblogging systems play an increasingly important role in our daily life. The large amount of microblogs makes it extremely difficult for users to keep track of public opinions. Therefore, it is crucial to mine and summarize opinions from microblogs automatically.

This paper aims to address the following two unique challenges posed by Chinese opinion summarization: automatic learning of important features and selection of representative sentences. We propose a CNN-based opinion summarization method for Chinese microblogging systems. The model first uses a convolution operation to produce feature maps; then the resolution of the feature maps is reduced by a pooling operation; finally, the obtained useful local features are fed to a fully connected softmax layer to predict the sentiment label of the microblog. Based on the prediction results, we build positive and negative datasets by selecting the posts and features with assigned sentiment labels. A two-step opinion generation method is proposed to generate the opinion summary: a semantic relationship mining model is used to identify important features and Maximal Marginal Relevance is applied to extract representative sentences to generate the opinion summary. The extracted representative sentences meet "relevant novelty" requirements. Experimental results on COAE and emergency/disaster real-world datasets from Chinese microblogging systems verify the efficacy of the proposed model.

Our work can be extended as follow. First, we would like to provide an efficient procedure for considering the user's feedback during the learning process. Second, we will further study the characteristics of emergency and natural disaster events on Twitter by applying the proposed model to analyze the semantic of opinions for these kinds of events and evaluating the efficacy. Third, it will be interesting to study whether the proposed model can be used for other kinds of events. Finally, how to incorporate other information such as temporal signals, geographical information and user influence information into summarization is an interesting and challenging research direction.

Acknowledgment

This research is supported in part by National Natural Science Foundation of China under Grant No. 91224008, 61172106, 61402123, 71402177. The Important National Science & Technology Specific Project under Grant No. 2013ZX10004218.

References

[1] J. Bai, Weibo User Development Report in 2014, 2014 Available: http://data.weibo.com/report/reportDetail?id=215 .
[2] Y. LeCun, Y. Bengio, G. Hinton, Deep Learn. Nature 521 (7553) (2015) 436–444.
[3] 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.
[4] R. Socher, A. Perelygin, J.Y. Wu, J. Chuang, C.D. Manning, A.Y. Ng, C. Potts, Recursive deep models for semantic compositionality over a sentiment treebank, in: Proceedings of the conference on empirical methods in natural language processing (EMNLP), 2013, pp. 1631–1642.

[5] C.N. dos Santos, M. Gatti, Deep convolutional neural networks for sentiment analysis of short texts, in: Proceedings of the 25th International Conference on Computational Linguistics (COLING), Dublin, Ireland, 2014, pp. 23–29.
[6] Y. Kim, Convolutional neural networks for sentence classification, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2014, pp. 1746–1751.
[7] A.E. Cano Basave, Y. He, R. Xu, Automatic labelling of topic models learned from Twitter by summarisation, in: Proceedings of the Association for Computational Linguistics (ACL), 2014, pp. 618–624.
[8] J. Carbonell, J. Goldstein, The use of MMR, diversity-based reranking for reordering documents and producing summaries, in: Proceedings of the 21st annual international ACM SIGIR Conference on Research and Development in Information Retrieval, 1998, pp. 335–336.
[9] R. Mihalcea, P. Tarau, TextRank: bringing order into texts, in: Proceedings of the Association for Computational Linguistics, 2004, pp. 404–411.
[10] L. Barbosa, J. Feng, Robust sentiment detection on twitter from biased and noisy data, in: Proceedings of the 23rd International Conference on Computational Linguistics, 2010, pp. 36–44.
[11] D. Davidov, O. Tsur, A. Rappoport, Enhanced sentiment learning using twitter hashtags and smileys, in: Proceedings of the 23rd International Conference on Computational Linguistics, 2010, pp. 241–249.
[12] L. Jiang, M. Yu, M. Zhou, X. Liu, T. Zhao, Target-dependent twitter sentiment classification, in: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, 2011, pp. 151–160.
[13] B. O'Connor, R. Balasubramanyan, B.R. Routledge, N.A. Smith, From tweets to polls: linking text sentiment to public opinion time series, in: Fourth International AAAI Conference on Weblogs and Social Media, 2010, pp. 122–129.
[14] B. Pang, L. Lee, S. Vaithyanathan, Thumbs up?: sentiment classification using machine learning techniques, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2002, pp. 79–86.
[15] X. Hu, J. Tang, H. Gao, H. Liu, Unsupervised sentiment analysis with emotional signals, in: Proceedings of the 22nd International Conference on World Wide Web, 2013, pp. 607–618.
[16] L. Zhu, A. Galstyan, J. Cheng, K. Lerman, Tripartite graph clustering for dynamic sentiment analysis on social media, in: Proceedings of the 2014 ACM SIGMOD international conference on Management of data, 2014, pp. 1531–1542.
[17] S. Brody, N. Diakopoulos, Cooooooooooooooolllllllllllllll!!!!!!!!!!!!!!!: using word lengthening to detect sentiment in microblogs, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2011, pp. 562–570.
[18] T. Li, V. Sindhwani, C.H. Ding, Y. Zhang, Bridging domains with words: opinion analysis with matrix tri-factorizations, in: SDM, 2010, pp. 293–302.
[19] Q. Miao, Q. Li, D. Zeng, Fine-grained opinion mining by integrating multiple review sources, J. Am. Soc. Inf. Sci. Technol. 61 (11) (2010) 2288–2299.
[20] C. Zhang, D. Zeng, J. Li, F.Y. Wang, W. Zuo, Sentiment analysis of Chinese documents: from sentence to document level, J. Am. Soc. Inf. Sci. Technol. 60 (12) (2009) 2474–2487.
[21] L. Zhou, P. Chaovalit, Ontology-supported polarity mining, J. Am. Soc. Inf. Sci. Technol. 59 (1) (2008) 98–110.
[22] T. Narock, L. Zhou, V. Yoon, Semantic similarity of ontology instances using polarity mining, J. Am. Soc. Inf. Sci. Technol. 64 (2) (2013) 416–427.
[23] J. Zhao, L. Dong, J. Wu, K. Xu, Moodlens: an emoticon-based sentiment analysis system for chinese tweets, in: Proceedings of the 18th ACM SIGKDD international Conference on Knowledge Discovery and Data Mining, 2012, pp. 1528–1531.
[24] K. Ravi, V. Ravi, A survey on opinion mining and sentiment analysis: tasks, approaches and applications, Knowl. Based Syst. 89 (2015) 14–46.
[25] J. Serrano-Guerrero, J.A. Olivas, F.P. Romero, E. Herrera-Viedma, Sentiment analysis: a review and comparative analysis of web services, Inf. Sci. 311 (2015) 18–38.
[26] J. Liang, C. Yumei, Y. Huibin, Z. Hongying, L. Ming, Deep learning for Chinese micro-blog sentiment analysis, J. Chinese Inf. Process. 28 (5) (2014) 155–161.
[27] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, arXiv:1301.3781, 2013.
[28] N. Kalchbrenner, E. Grefenstette, P. Blunsom, A convolutional neural network for modelling sentences, in: Proceedings of the Association for Computational Linguistics, 2014, pp. 655–665.
[29] 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.
[30] M. Hu, B. Liu, Mining opinion features in customer reviews, in: AAAI, 2004, pp. 755–760.
[31] L. Zhuang, F. Jing, X.-Y. Zhu, Movie review mining and summarization, in: Proceedings of the 15th ACM International Conference on Information and knowledge management, 2006, pp. 43–50.
[32] I. Titov, R. McDonald, Modeling online reviews with multi-grain topic models, in: Proceedings of the 17th International Conference on World Wide Web, 2008, pp. 111–120.
[33] X. Meng, F. Wei, X. Liu, M. Zhou, S. Li, H. Wang, Entity-centric topic-oriented opinion summarization in twitter, in: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2012, pp. 379–387.
[34] N. Aletras, M. Stevenson, Labelling Topics using Unsupervised Graph-based Methods, in: Proceedings of the Association for Computational Linguistics, 2014, pp. 631–636.
[35] Y. Chang, X. Wang, Q. Mei, Y. Liu, Towards Twitter context summarization with user influence models, in: Proceedings of the Sixth ACM International Conference on Web Search and Data Mining, 2013, pp. 527–536.
[36] S. Song, Q. Li, X. Zheng, Detecting popular topics in micro-blogging based on a user interest-based model, in: Proceedings of the 2012 International Joint Conference on Neural Networks (IJCNN), 2012, pp. 1–8.
[37] S. Song, Q. Li, H. Bao, Detecting dynamic association among Twitter topics, in: Proceedings of the 21st International Conference Companion on World Wide Web, 2012, pp. 605–606.
[38] Z. Zhang, Q. Li, D. Zeng, H. Gao, User community discovery from multi-relational networks, Decis. Support Syst. 54 (2) (2013) 870–879.
[39] Y. Lu, X. Hu, F. Wang, S. Kumar, H. Liu, R. Maciejewski, Visualizing social media sentiment in disaster scenarios, in: Proceedings of the 24th International Conference on World Wide Web Companion, 2015, pp. 1211–1215.
[40] Z. Jin, Q. Li, D. Zeng, L. Wang, Filtering spam in Weibo using ensemble imbalanced classification and knowledge expansion, in: Proceedings of 2015 IEEE International Conference on Intelligence and Security Informatics, 2015, pp. 132–134.
[41] T. Fawcett, An introduction to ROC analysis, Pattern Recognit. Lett. 27 (8) (2006) 861–874.
[42] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I.H. Witten, The WEKA data mining software: an update, ACM SIGKDD Explor. Newslett. 11 (1) (2009) 10–18.