Wang et al. EURASIP Journal on Wireless Communications and Networking (2016) 2016:253
DOI 10.1186/s13638-016-0745-7

EURASIP Journal on Wireless Communications and Networking

RESEARCH Open Access

A hybrid model of sentimental entity recognition on mobile social media

Zhibo Wang[1,2], Xiaohui Cui[1*], Lu Gao[1], Qi Yin[1], Lei Ke[1] and Shurong Zhang[1]

Abstract

With new forms of media such as Twitter becoming increasingly popular, the Internet is now the main conduit of individual and interpersonal messages. A considerable amount of people express their personal opinions about news-related subject through Twitter, a popular SNS platform based on human relationships. It provides us a data source that we can use to extract peoples' opinions which are important for product review and public opinion monitoring. In this paper, a hybrid sentimental entity recognition model (HSERM) has been designed. Utilizing 100 million collected messages from Twitter, the hashtag is regarded as the label for sentimental classification. In the meanwhile, features as emoji and N-grams have been extracted and classified the collected topic messages into four different sentiment categories based on the circumplex sentimental model. Finally, machine learning methods are used to classify the sentimental data set, and an 89 % precise result has been achieved. Further, entities that are behind emotions could be gotten with the help of SENNA deep learning model.

Keywords: Feature selection, Sentiment analysis, Sentiment classification, Entity recognition

1 Introduction

The social network is highly developed nowadays, and people can almost access it anywhere. In this aspect, there are many projects which are worth doing research on. Twitter allows users to login, through the webpage, mobile devices, PC, or other client. Users can send an about 140-word message to Twitter every time for sharing information, and the fans who follow these users can repost or share these messages. These messages which express the users' opinions, thoughts, and emotions in 140 words are called tweets. By the end of March 2012, there were more than 140 million active users, living in all over the world, in Twitter. Every day, Twitter's users posted about 340 million tweets which use over 10 TB storage and the number is rising continuously. Because of these, Twitter is one of the top 10 most visited websites.

For these reasons, Twitter attracts a large number of natural language processing experts to research on this field. The data mining and analysis of Twitter can be used in various fields, such as epidemic prediction, population migration, public opinion surveillance, and so on [1]. To be specific, comments about products in tweets are well worth mining. Sellers can get the buyers' comments in real time and then update their own products to be more competitive in the market place; buyers can get others' experience through these comments to help them decide whether to buy a product. Due to the characteristics of a tweet such as they can be produced in real time and can spread widely and quickly, they have large influence on network public opinion transmission [2]. It is necessary for the government to know and analyze the public opinion on some hot social issues preventing the views of the public from being deluded by criminals who may harm the country and society. So, recognizing emotions and entities of Twitter data has a very important reference value.

2 Related work

As a main form of media in the social network and the main microblog abroad, Twitter attracts people more and more. Tweets contain different tendencies and emotional characteristics; and mining these features is meaningful for public opinion monitoring, marketing, and rumor control. In general, most emotional analysis only divides the text emotion into three categories:

* Correspondence: xcui@whu.edu.cn
[1]International School of Software, Wuhan University, Wuhan 430079, China
Full list of author information is available at the end of the article

neutral, positive, and negative. It is limited to help people listen to the real voice and emotion of society.

The study of social media is relatively new. As early as 2005, Park et al. [3] began to analyze the emotion on Twitter. They labeled more than 20,000 tweets and created an emotion polarity dataset by labeling the neutral-positive-negative emotional tag manually. Next, they developed an emotion classifier by using machine learning method based on Naïve Bayes, support vector machine (SVM), and conditional random fields (CRF). Read et al. [4] put forward that they used twitter application program interface (API) to get a great number of emotion icon and demonstrated these icon's effect on emotion classification in detail. Go et al. [5] developed three machine learning classifiers based on Naïve Bayes, maximum entropy, and SVM by using a non-supervisor machine learning algorithm. They added the emotion icons into the selected features which caused the accuracy of the emotional tendency discrimination to be more than 80 %. This research has been applied to many business fields such as online shopping, online film review, and online 4S shop message. For instance, Fei HongChao analyzed the review text aimed at Yahoo English Sports. Through that, the attitude of the investors to the stock market could be discovered. Ghose et al. researchers started to apply the LingPipe for emotion classification. They tried to increase the accuracy of classifiers by labeling the training set manually and then recognized the emotion tendency of the original text. The amount of research about text mining based on emotion is growing, and the related research fields are extended at the same time. R. Pavitra [6] established an analysis model based on the weakly supervised joint sentiment-topic mode and created a sentiment thesaurus with positive and negative lexicons to find the sentiment polarity of the bigrams. Wang and Cui [7, 8] worked on group events and disease surveillance to research the sentiment analysis. They also extended the data source to multimedia for research on sentiment analysis [9].

Recently, with the development of computer technology on information searching and search engines, named-entity recognition has been a hot topic in the field of natural language processing. Asahara [10] performed the automatic identification on names and organizations by SVM and got some good results. Tan utilized a method based on transferring the wrong drive to get context contact rules of naming entity places. Next is using rules to implement automatic identification the names of places. According to the data test, the accuracy of this method can achieve 97 %. Huang et al. [11] got a large amount of statistics data from vast real text data, and they calculated every one's reliability of continually-words-construction and words construction. Finally, combining some rules, the names could be recognized automatically. Turkish scholars [12] did the named-entity recognition on their domestic twitter. In their article, a new named-entity-annotated tweet corpus was presented and the various tweet-specific linguistic phenomena were analyzed. After that, Derczynski L and his group [13] also worked on the similar field. Even Xu et al. [14] have a patent on named-entity recognition inquiry.

Some relevant techniques in data mining of tweets' fine-grained sentiment analysis will be researched, including the methods of tweets collection, the tweets pre-processing, and the construction of knowledge. Based on a tweets' emotional dictionary, sentiment analysis based on weighted emotional words meaning and sentiment analysis based on multi-feature fusion. Tweets text has the features of a large amount of data, covering a wide range and rapidity, so it impossible to monitor hot events and analyze guidance of public opinion manually. For processing the huge amount of unstructured text data, machine learning and deep learning have had some certain breakthrough in the field of text processing. In the part of sentiment analysis, we will build a circumplex sentiment model by using hashtags as the classification tags, catching N-gram, and emoji features. Then, the emotion will be classified through the processing of a SENNA model. It is possible to classify four kinds of emotions which we described in advance.

3 Definition of question

We aim to deduce the user's emotion by analyzing their tweet text. To give formalized definition of this question, we predefine some like below:

Definition 1 (Tweet words w): Since each word in a tweet is possible related to users' emotion, so we add up all the words in blog text and use a two-tuples to represent it, $w = \{t, a\}$. t is the text form of w, a is the frequency of w in a tweet.

Definition 2 (sentiment dictionary D): for each sentiment, we can design a dictionary which can represent it sharply, called sentiment dictionary. The dictionaries of different sentiment can include same words, since dictionary exacts influence on the sentiment analysis as entirety. We use a two-tuples to represent each sentiment dictionary: $\mathrm{Di} = \{d, v\}$. d is each word in dictionary, v is central vector of this sentiment. The closer user's vector model is to central vector, the more likely the user is to be this sentiment. The words in dictionary also can be represented as two-tuples: $d = \{t, c\}$. t is the text form of d, c is the relevancy of the word d and the sentiment.

4 Methods

4.1 Data collection

Twitter has played a vital role in the spread of social hot spots; therefore, a comprehensive, timely, and accurate access to twitter data is the basis and premise of our data analysis. At present, there are three main acquisition methods for Twitter: API acquisition, webpage analysis acquisition, and APP capture acquisition.

Our training set data uses dictionary tables as keywords. Besides, a tweet acquisition system based on a search API has been designed. Aims to categorize four different emotions, it contains 50 thousand tweets as training data.

The system contains the following modules:

A, the keyword dictionary module: used to construct data dictionaries and send requests to the API according to the words in the dictionary
B, the pre-process module: used to preprocess the data in real time, such as preprocessing time and extraction of useful data
C, the database storage module: based on the returned data format, to use MongoDB, NoSql database, and set up a collection of four different emotional words dictionaries

4.2 Text feature extraction

At present, the mainstream research of text feature extraction aims in the feature selection algorithm and text representation model.

The basic unit of the feature is called text features. Text feature must have the following characteristics:

(1) The ability to distinguish features of the target and the non-target texts
(2) The ability to clearly express and identify text content
(3) Easiness to implement the classification between features
(4) The number of features should not be too many

On the condition that avoids damaging the core content of texts, the main purpose of feature extraction is to reduce the dimension of feature vectors as far as possible, reducing the number of words needed to process. Thus, it should simplify the computational complexity and ultimately improve the efficiency and speed of text processing.

There are several factors that could be used to implement text feature extraction, such as word frequency method, principal component analysis method, TF-IDF method, N-gram method, and emoji method

Actually, word frequency method may filter some words which have useful information but small frequency, TF-IDF method cannot differently calculate the weight based on the different specific locations of the word, but N-gram method can filter the text information according to the given threshold value and there are lots of emoji to express their emotion in tweets nowadays. Considering the format and content of Twitter, features will be extracted from the text by N-gram and emoji methods in this project.

N-gram model will be a continuous identification of large quantities of data words. In general, these data vocabulary can be letters, words, syllables, and so on, and we use N-gram model to implement the automatic classification of emotion.

(1) Revised N-gram model

In statistical concepts, N-gram model can be understood as any word in a text such as algorithm w_i, and its emergence probability only has some connection with the times it appeared. This determines the w_i's independence. We can apply the probabilistic conditional probability can be used to predict the probability of the w_i appeared.

If using W represents a string of sequence which have N elements, then $W = \ldots$. And the appearance probability of the sequence is $P(W)$:

$$P(W) = P(\,)P(\ |\ )P(\ |\,) \ldots P(\ |\ \ldots\ ) \tag{1}$$

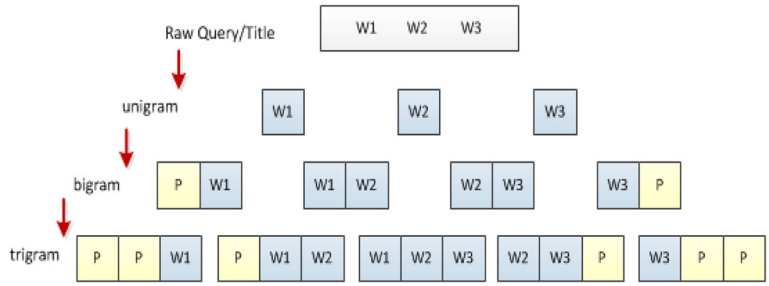

Fig. 1 Revised N-Gram model

Table 1 Emoji classification table of Twitter

Emotion Type	Typical Emoticon
Happy Active	
Unhappy Active	
Happy Inactive	
Unhappy Inactive	

If there is a large amount of data, according to the Markov assumption, the probability of a word appears only associated with the probability of the word in front of it, and then problem becomes simple. Therefore, uni-gram model changed to the binary model bi-gram.

$$P(W) \text{ material } P(\,)\,P(\,|\,)\,P(\,|\,)\ldots P(\,|\,) \tag{2}$$

In the same way, we can get tri-gram, the probability of words appearance only related to the probability of two words in front of it.

$$P(W) \text{ material } P(\,)\,P(\,|\,)\,P(\,|\,)P(\,|\,)\ldots P(\,|\,) \tag{3}$$

In the same way, we can also get the concept of N-gram model.

In our research, we used the improved version of the N-gram model, namely adding padding characters (general spaces or whitespace) at the beginning of each bi-gram and tri-gram to increase the number of grams, improving the prediction accuracy of the models, as shown in Fig. 1. Sometimes, a tweet contains only a few words, using the tri-gram model can only static few characteristics, but the feature quantity is improved significantly after adding the padding characteristic.

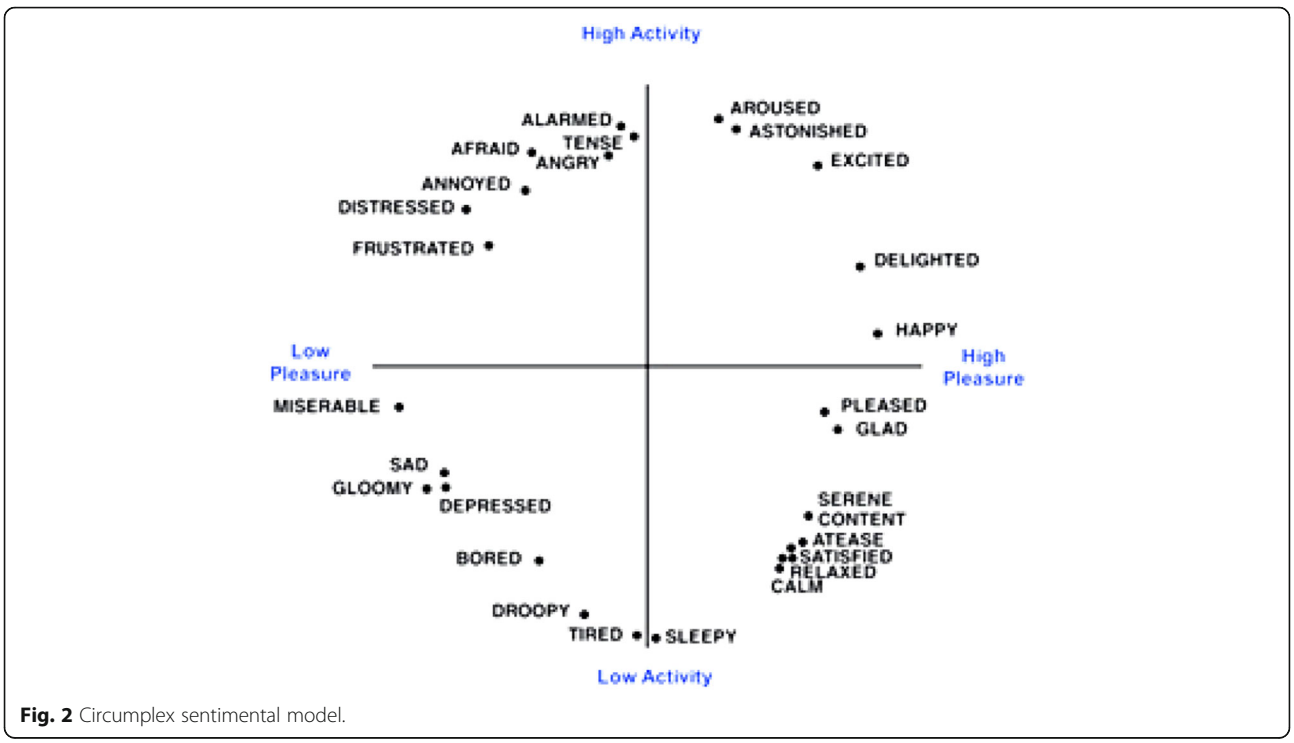

Fig. 2 Circumplex sentimental model.

(2) Emoji

The emoji classification table of Twitter found in Table 1.

4.3 Sentimental entity analysis

This section contains the application of sentimental models in the analysis of emotion and emotion detection of Twitter text. At present, most sentimental model are based on psychological theories and social network sentimental hypothesis.

According to the theme needed and the variety of sentimental, this paper adopts the circumplex and four coordinate pole edge as sentimental model and keywords. Figure 2 is an example of circumplex sentimental model. In order to maintain the independence of the four kinds of emotion and reduce affection overlaps, it removes sentimental words between axes.

Finally, it acquires four types of emotional keywords in the dictionary. Keywords in this dictionary are the training data set that used to train the data. The dictionary table is shown in Table 2.

4.4 Text classifier training

Scikit-learn is a Python module for machine learning, built on the basis of SciPy and 3-ClauseBSD open source license. Its main features are (1) the advantages of simple operation, efficient data mining, and data analysis, (2) unrestricted access that can be re used in any cases, (3) basis on NumPy, SciPy, and Matplotlib, and (4) commercial open source BSD license certificate.

A pipeline refers to the data processes for what a pipeline always contain multiple feature conversion phases. The feature transformation could be regarded as a new column that is added to an existing column. An example of the pipeline is text segmentation that divides texts into a large number of words; TF-IDF transforms it into a feature vector. In this process, the tag will be processed for model fitting when the ID, text, and words are transferred into the conversion process.

Table 2 Four types of sentimental dictionaries

HappyActive	UnhappyActive
#elated,#overjoyed,#enjoy,#excited,	#nervous,#anxious,#tension,#afraid,
#proud,#joyful,#feelhappy,#sohappy	#fearful,#angry,#annoyed,#annoying
#veryhappy,#happy,#superhappy,	#stress,#distressed,#distress,
#happytweet,#feelblessed,#blessed,	#stressful#stressed,#worried,#tense,
#amazing,#wonderful,#excelent,	#bothered,#disturbed,#irritated,
#delighted,#enthusiastic	#mad,#furious
HappyInactive	UnhappyInactive
#calm,#calming,#peaceful,#quiet,	#sad,#ifeelsad,#feelsad,#sosad,
#silent,#serene,#convinced,#consent	#verysad,#sorrow,#disappointed,
#contented,#contentment,#satisfied,	#supersad,#miserable,#hopeless,
#relax,#relaxed,#relaxing,#sleepy,	#depress,#depressed,#depression,
#sleepyhead,#asleep,#resting,	#fatigued,#gloomy,#nothappy,
#restful,#placid	#unhappy,#suicidal,#downhearted,
	#hapless,#dispirited

Table 3 Precision between the classifiers and characteristics

Characteristics	Naïve Bayes	Logistic regression	SVM	KNN
Uni-gram	86.3	85.2	89	88.1
Uni-gram, emoji	86.4	84.2	89.1	88.5
Uni-gram, punctuation	86.6	85.6	88.9	88.8
All features	86.9	85.9	89.8	89.1

Classification is an important task in data mining. Using a machine learning classifier to train a set is for the purpose to let the machine trained classifier automatic classify the content of new data and thus liberating human resources to undertake artificial classification. Therefore, it is very important to clarify which kind of classifier to use to classify data in the field of data mining.

There are four different classifiers to choose to classify data: Naïve Bayesian, logical regression, support vector machines, and K-Nearest Neighbor algorithm. Naïve Bayesian is a generative model, by calculating the probability to classify, can be used to handle multiple classifications. Logical regression can be implemented by simple and need low storage resources. SVM is the best existing classifier, and existing means can be used directly without modification. You can get a lower error rate, and SVM can do well in classification decisions on the data outside the training set. K neighbor algorithm can not only be used for classification but also can be used for regression analysis.

In order to achieve data classification, Naïve Bayesian, logistic regression, SVM, and K-Nearest Neighbor algorithm have been implemented in our classifiers, and a five-fold cross validation also contributes to the evaluation of the final results. Table 3 shows the precision of the classifiers, and Table 4 shows the recall rate of the results.

In text classification, support vector machine (SVM) is one of the best classifier algorithms. It is because the use of a support vector machine can transform linearly non-separable text to high-dimensional space that can make it classifiable. The kernel function is used to transform from low dimensions to high dimensions and obtain the

Table 4 Recall rate between the classifiers and characteristics

Characteristics	Naïve Bayes	Logistic regression	SVM	KNN
Uni-gram	86.3	85.5	89.3	88.3
Uni-gram, emoji	86.4	84.9	89.7	88.5
Unigram, punctuation	86.5	85.3	88.6	88.9
All features	87	86	90	89

result of the high dimension. Compared with the Naïve Bayes and KNN algorithms, SVM has the advantages of high classification speed, high accuracy of the training data size, and universal property.

We choose to use SVM classifier in our experiment and the following is the principle:

SVM method is to make the sample space mapped into a high-dimensional and even infinite dimensional feature space (Hilbert space) through a non-linear mapping p, so that the nonlinear separable problem in original sample space is transformed into linearly separable problem in feature space.

In our paper, we regard precise rate and recall rate as the evaluation index.

Assume

$\mathrm{M}_{\underset{i}{\text{right}}}$ represents the number of tweets which is correctly divided into class Ci

$\mathrm{M}_{\underset{i}{\text{wrong}}}$ represents the number of tweets which is mistakenly divided into class C_i

$\mathrm{M}_{\underset{i}{\text{all}}}$ represents the number of tweets which is included in the class C_i actually

Then

$$\text{Precise rate}(i) = \frac{\mathrm{M}_{\underset{i}{\text{right}}}}{\mathrm{M}_{\underset{i}{\text{right}}} + \mathrm{M}_{\underset{i}{\text{wrong}}}} \times 100\% \tag{4}$$

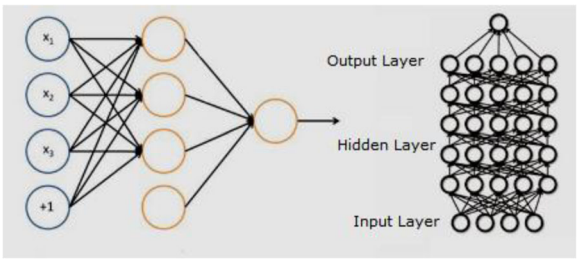

Fig. 3 Deep learning training model

$$\text{Recall rate}(i) = \frac{\mathrm{M}_{\underset{i}{\text{right}}}}{\mathrm{M}_{\underset{i}{\text{all}}}} \times 100\% \tag{5}$$

And the average precise rate and average recall rate can be calculated according to the formulas:

$$\text{Average precise} = \frac{\sum_{i=1}^{m} \text{Precision rate}_i}{m} \tag{6}$$

$$\text{Average recall} = \frac{\sum_{i=1}^{m} \text{Recall rate}_i}{m} \tag{7}$$

Algorithm SVM: The SVM based on the analysis of Twitter user's sentiment.

Input

- D: sentimental classification dictionary
- T: all target tweet

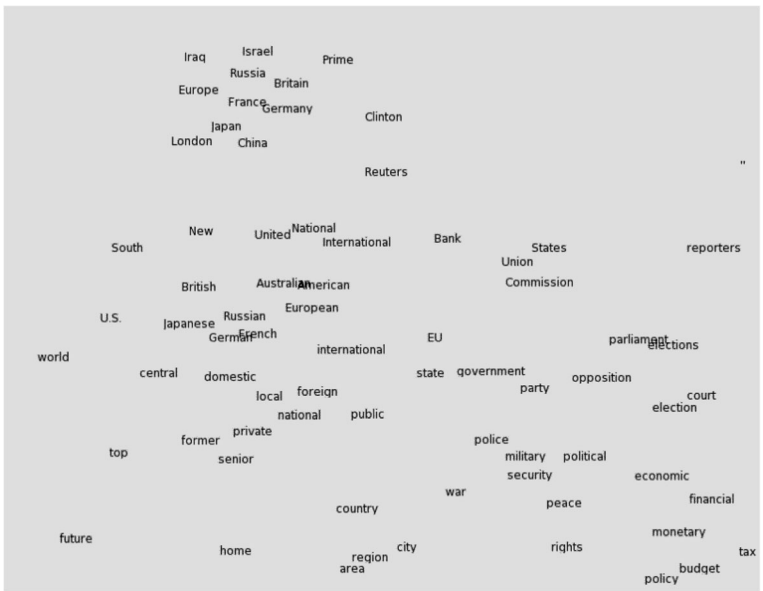

Fig. 4 Words in high-dimensions space

- S: sentimental category table included in system

Output: the sentiment E of target tweet t

1. For each T's every word w do
2. For each α1 not satisfied KKT condition
3. If min = distance (α1, α2)
4. Check $\sum_{i\text{-}1}^{m} \alpha i \cdot \text{label} = 0$
5. Calculate α1, α2
6. Check label * $(W^{T}X + b) \geq 1.0$
7. Calculate b1,b2
8. Returnα matrix and b matrix

4.5 Entity detection

Named-entity recognition, also called entity recognition, means the entities with specific meaning in a series of text and it mainly refers to the names of people, places, organizations, and proper noun. There are four main techniques for named-entity recognition:

(1) The statistical-based recognition method. The main statistical models for named-entity recognition include hidden Markov model, decision tree model, support vector machine (SVM) model, maximum entropy model, and conditional random fields model.
(2) The rule-based recognition method. It mainly uses two pieces of information: restrictive clauses and named-entity words.
(3) The recognition method of combining rules and statistics. Some mainstream. Named-entity recognition systems combine the rules and the statistics. First, they use the statistical methods to the image to recognize the entities, and then, correct and filter them by the rules.
(4) The recognition method based on machine learning. This technology in English is developed to some extent. Classifying English words by SVM machine methods can achieve an accuracy more than 99 % when the places or names of people are recognized.

Deep learning is a new branch in the field of machine learning and a kind of algorithm that stimulates functions of the brain. Deep learning originated from the deep belief nets originated from the Boltzmann machine presented in the Hinton paper. The basic idea of deep learning is that, for a system *S* with *N* level (S1, S2… SN). If input is I and output is *O*, a formula could be used to express this system as $I => S1 => S2 => S3 => Sn => O$. This system should automatically learn some features to help people make decisions. By adjusting parameters in each layer of the system, the output of the lower level could be the input for the higher level. And by piling from various layers, there could be a hierarchical expression of input information.

Deep learning training model of the system is in Fig. 3.

Natural language is the human-use communication and direct language, but in the order for computers to make computational identification, it needs to convert natural language into computer-use symbols, usually called the digital of natural language. In deep learning, words are embedded to represent words. The word embedding method was proposed by Bengio more than a dozen years ago. The words in the language are mapped into the high-dimensional vector with 200 to 500 dimensions. By training word vector with deep learning, each word has their corresponding spatial coordinates in high-dimensional space. The sample of space coordinate map is in Fig. 4:

At the beginning of the training process of the word vector, each word will be given a random vector. For example, deep learning is used to predict if a quintet phrase is true, such as "Benjamin likes play the basketball". If taking the replacement for any one of the words in this sentence, such as replacing "the" with "theory", "Benjamin likes play theory basketball" is obviously not true in the grammar. Using models trained by deep learning. It is possible to predict whether changed quintet phrases are true or not.

SENNA not only proposed the method for building word embedding but also solved the natural language

Table 6 Label table of sentiment classification in Twitter

Emotion type	Label
Happy-active	1
Happy-inactive	2
Unhappy-active	3
Unhappy-inactive	4

Table 5 A sample table of sentiment classification in Twitter

Tweets	Type
I love this apple! 😝	Happy-active
You are really a bad 😒 guy.	Unhappy-active
I feel surprise by these roses. 😿	Happy-active
Pray for MH17! 😞	Unhappy-inactive

Table 7 Sample labeling of Twitter

Tweets	Label
I love this apple! 😝	1
You are really a bad 😒 guy.	3
I feel surprise by these roses. 😿	1
Pray for MH17! 😞	4

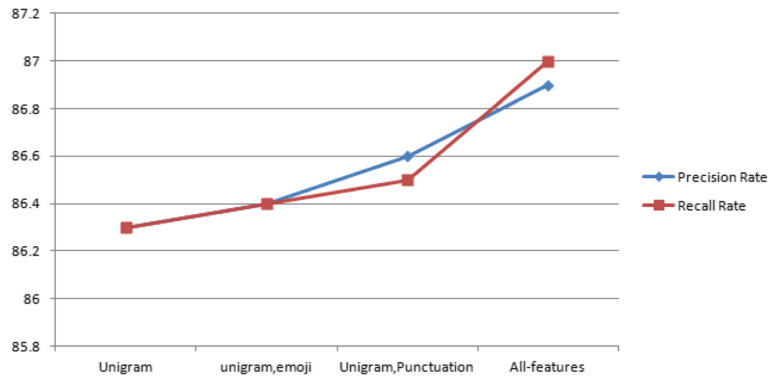

Fig. 5 Precision rate and recall rate of Naïve Bayes

processing tasks (POS, Chunking, NER, SRL) from the perspective of neural network language model system. In SENNA, each word can be directly found from the lookup table.

The word vectors of HLBL in SENNA which are different from each other are used to depict different semantics and grammar usage. The word vectors of a word are combined by various forms of the vector word eventually. SENNA directly pieced the vectors together and represented the words.

Then, the emotion will be classified through the processing of a SENNA model. It is possible to classify four kinds of emotions which we described in advance.

5 Experiment and analysis

5.1 Data acquisition

The data set used in the paper is collected from Twitter between Mar. 1st and May 1st of 2015 by means of the search API in the open platform of Twitter. The dictionary of keywords is included in four kinds of sentiment whichused to get the training data. It takes a great number of resources to classify the tweets due to its enormous quantity. Therefore, the paper selects hashtag in the tweets as the tag which are automatically classified by using a machine method. A hashtag, as a level of tag in the tweets, is used to record a certain topic. The paper believes that the tweets with the label of a certain sentiment category belong to the category, so as to implement the automatic classification of the machine.

5.2 Data preprocessing

Not like the traditional news or media data, tweets are a kind of data as daily expression, so they have a lot of error and "noise". Before classification, the "noisy" data should be deleted as following:

1) Deleting username and URL
   A part of usernames contains sentiment words and are even contained by our sentiment dictionary. However, the username has little use in the sentiment analysis. In order to exclude the interference on the experimental results by the username, the paper regards all the labels with the "@" symbol as USERID. Meanwhile, the tweets contain a lot of URL which are useless for text processing, so they are replaced by the word URL.

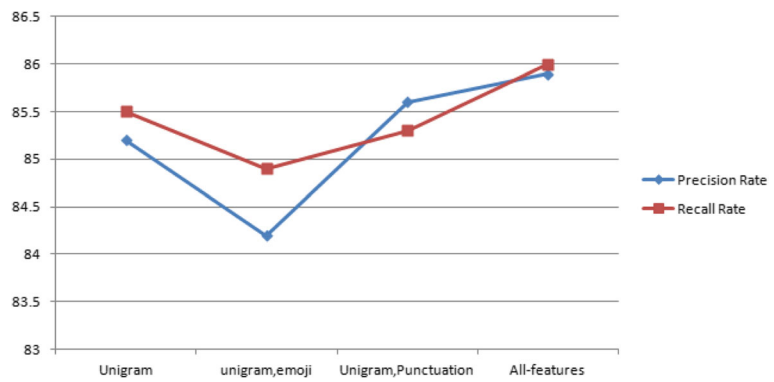

Fig. 6 Precision rate and recall rate of logistic regression

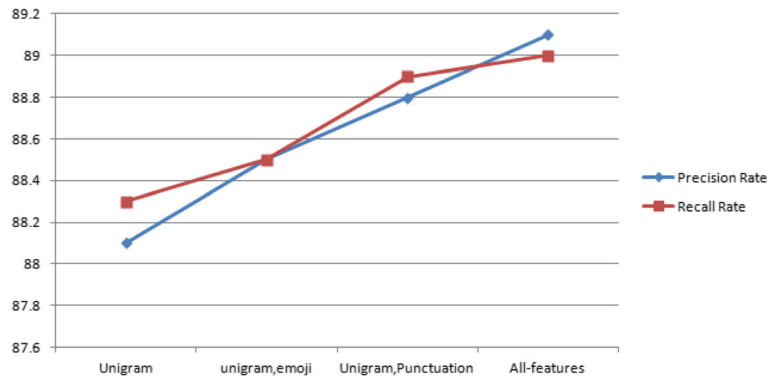

Fig. 7 Precision rate and recall rate of KNN

2) Spelling mistakes
   Tweets have a large number of words with spelling mistakes, such as "asd" (sad), "Happyyyyyy!!", "Cooool~~", some of which are unintentional misspellings and some for emphasis. In order to reduce the interference, the module corrects the misspelled words by using the wrong words dictionary of tweets.
3) Abbreviations
   There are many abbreviations of words in tweets such as "good4you" which means "good for you". A particularly large number of words are expressed by the short form of letters and numbers; therefore, the correcting of abbreviation is also a significant part of preprocessing.
4) Confliction between sentiment categories
   There are four sentiment categories, and some large tweets contains hashtag in two different sentiment categories. These tweets are named conflicting tweets. In order not to affect tweets classification accuracy, the module deletes these tweets!

5.3 Text feature extracting

During the building of pipeline, the module adopts the text feather of N-gram and adds English punctuation, emotional symbols, and bag of words as feathers. Taking the emoji as an instance, the tweets containing the same category of emoji are classified as the same category (Table 5).

Emoji is stored as unicode, so it is the first to classify the unicode character and signed by number (Table 6).

Scanning every tweet, and add a key-value pair in list when meeting an emoji (Table 7).

Wordnet builds a network including all synonyms and can replace all words in the same synonym network as the same word. Therefore, in the sentiment classification, the module reduces the feature dimensionality and increases the classification accuracy by using the replacement.

This module adopts a sentiment dictionary provided by Harvard University that has fewer words so as to reduce the dimension of the feature vector space and the amount of calculation.

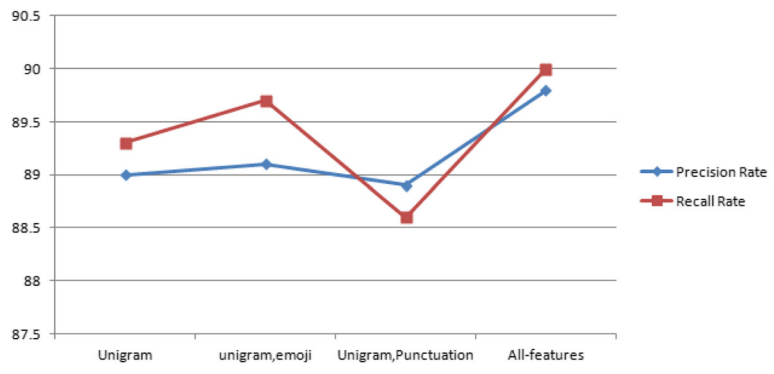

Fig. 8 Precision rate and recall rate of SVM

Table 8 Four kind of emotion and the entity extracted from them

Happyactive	HappyInactive	UnhappyActive	UnhappyInactive
Ganada	Netflix	Mad Men	Pocket Full Of Gold
NDSU	youtube	LiveYourTruth	AmnestyOnline
Sensation	Levis	HolyBible	AhmedahliCom
Game	Newspaper	KingLikeAQueen	EdgarAllanPoe
Filmphotography	VallartaGV	NTUWFC	Elena
KimFCoates	yoga	backstreetboys	JonnyValleyBoy
Ft. Beyonce	CandyCrushSaga	BLOOMparenting	GinyTonicBlog
Drake	ICandlelighters	STFU Louise	Havasupai
StuartPWright	HillCountry	YL train	Bethany
Longley_Farm	SLU	Rebecca De Mulder	David Letterman

5.4 Classifier training and result analysis

From the graph, we can see that by using different way of feature extraction, the precision rate and recall rate of Naïve Bayes is different. The value of precision rate is around 86.5 % (86.3~86.9 %), and the recall rate is around 86.5 % (86.3~87 %), too. It proved that with uni-gram, emoji, and punctuation of these all features, the precision rate can up to the maximum 86.9 % (Fig. 5).

From the graph we can see that by using different way of feature extraction, the precision rate and recall rate of Logistic regression is different. The value of precision rate is around 85 % (84.2~85.9 %), and the recall rate is around 85 % (84.9~86 %), too. It proved that with uni-gram, emoji, and punctuation of these all features, the precision rate can be up to the maximum 85.9 % (Fig. 6).

From the graph, we can see that by using different way of feature extraction, the precision rate and recall rate of KNN is different. The value of precision rate is around 88.5 % (88.1~89.1 %), and the recall rate is around 88.5 % (88.3~89 %), too. It proved that with uni-gram, emoji, and punctuation of these all features, the precision rate can be up to the maximum 89.1 % (Fig. 7).

From the graph, we can see that by using different way of feature extraction, the precision rate and recall rate of SVM is different. The value of precision rate is around 89 % (88.9~89.8 %), and the recall rate is around 89 % (88.6~90 %), too. It proved that with uni-gram, emoji, and punctuation of these all features, the precision rate can be up to the maximum 89.8 % (Fig. 8).

Comparing the data in this project, it is obvious that by using uni-gram, emoji, and punctuation as characteristics and SVM as emotional classifiers, the classification accuracy could reach 89.8 %. SVM is the best sentimental classification method for our experiment.

5.5 Results and analysis of named-entity recognition

After the training of emotional classifiers, an automatic classifier has been implemented to deal with 5000 new data values. Also, the named-entity recognition has been processed for each type of data. In this section, the SENNA deep learning toolkit is adopted for entity extraction for each type of data and at the same time all of these words are sorted. Table 8 shows the top 10 results.

For each type of emotional entities, visual graphic displays as follows (Figs. 9, 10, 11, and 12).

By using SENNA, we extracted the emotional entities from 5000 new data values. Actually, these entities are the reasons why users show these types of emotions. The result shows that Netflix could let the users feel "HappyInactive". At the same time, when we read the

Fig. 9 HappyActive word clouds

Fig. 10 HappyInactive word clouds

Fig. 11 UnhappyActive word clouds

original essay, we noticed that foreign users often spend their leisure time on Netflix. Most people feel excited when they arrived in Canada, which is an extraordinary beautiful place. And students in NDSU always express their activity during the exam period or graduation.

6 Conclusions

This project aims at the current Tweets sentiment analysis technology such as traditional attitude classification for positive-negative and median-type, multivariate emotion model classification. At the same time, we alter the machine learning method in the emotion classification and deep learning method in entity recognition.

Innovation points of this project are the following:

1. In the current emotional analysis, most research just takes into account the classification of the positive and negative attitude of emotion, and no one worldwide found the composition causes behind the emotional analysis. But in our daily life, the emotional composition analysis of the hot issues and public opinion monitoring has a very wide application.
2. In order to use deep learning methods in named-entity recognition, previous researchers used the pos tags. But for the data with a lot of noise, using this method, it is difficult to achieve a considerable level. However, identification accuracy of entity recognition theory in deep learning can reach more than 95 %.

Fig. 12 UnhappyInactive word clouds

The technology provided in this paper can be applied to the public opinion monitoring of social media. The training of the classifier is can be also used for Sina Weibo in Chinese public opinion analysis. In this research, we find that the scale of the data used in the experiment has a huge effect on the precision of the final results. So, in order to improve the accuracy of the classifier, the extensibility of the classifier for all tweet data is of great significance. Besides, our work in the future will transfer the emotional mining technology to Chinese texts, so as to mine the feelings of Chinese Sina Weibo users.

Acknowledgements

This research is supported in part by the National Nature Science Foundation of China No. 61440054, Fundamental Research Funds for the Central Universities of China No. 216274213, and Nature Science Foundation of Hubei, China No. 2014CFA048. Outstanding Academic Talents Startup Funds of Wuhan University, No. 216-410100003. National Natural Science Foundation of China (No. 61462004). Natural Science Foundation of Jiangxi Province (No. 20151BAB207042). Youth Funds of Science and Technology in Jiangxi Province Department of Education (No. GJJ150572).

Competing interests

The authors declare that they have no competing interests.

Author details

[1]International School of Software, Wuhan University, Wuhan 430079, China. [2]School of Software, East China University of Technology, Nanchang 330013, China.

Received: 26 May 2016 Accepted: 2 October 2016
Published online: 24 October 2016

References

1. L Xiangwen, X Hongbo, S Le, Y Tianfang, Construction and analysis of the third Chinese Opinion Analysis Evaluation (COAE2011) Corpus. J. Chin. Inform. Process. **01**, 56–63 (2013)
2. A Apoorv, X Boyi, V Ilia, R Owen, P Rebecca, Sentiment analysis of Twitter data, in *Proceedings of the Workshop on Languages in Social Media [D], LSM '11* (Association for Computational Linguistics, Stroudsburg, PA, USA, 2011), pp. 30–38
3. Pak A, Paroubek P. Twitter as a corpus for sentiment analysis and opinion mining[C] Seventh Conference on International Language Resources & Evaluation. 2010
4. Read J. Using emoticons to reduce dependency in machine learning techniques for sentiment classification. Proceedings of the ACL Student Research Workshop. Association for Computational Linguistics, 2005:43–48
5. Go A, Bhayani R, Huang L. Twitter sentiment classification using distant supervision. Cs224n Project Report, 2009:1–12
6. R Pavitra, PCD Kalaivaani, Weakly supervised sentiment analysis using joint sentiment topic detection with bigrams. Electronics and Communication Systems (ICECS), 2015 2nd International Conference on. IEEE. **2015**, 889–893 (2015)
7. Zhibo W, Kuai H, Luyao C, Xiaohui C. Exploiting Feature Selection Algorithm on Group Events Data Based on News Data. IIKI2015(2015 International Conference on Identification, Information, and Knowledge in the Internet of Things.), 2015:62–65
8. C Xiaohui, Y Nanhai, W Zhibo, H Cheng, Z Weiping, L Hanjie, J Yujie, L Cheng, Chinese social media analysis for disease surveillance. Pers. Ubiquit. Comput. **19**, 1125–1132 (2015)

9. W Zhibo, Z Chongyi, S Jiawen, Y Ying, Z Weiping, C Xiaohui, *Key technology research on user identity resolution across multi-social media*. CCBD2015, 2015
10. CL Coh, M Asahara, Y Matsumoto, Chinese unknown word identification using character-based tagging and chunking, in *Proceedings of the 41st Annual Meeting of the association of Computational Linguistics, Sapporo*, 2003, pp. 197–200
11. H Degen, Y Yuansheng, W Xing, Z Yanli, Z Wanxie, Identification of Chinese names based on statistics. J. Chin. Inform. Process. **02**, 31–37+44 (2001)
12. Küçük D, Jacquet G, Steinberger R. Named entity recognition on Turkish tweets. Language Resources and Evaluation Conference. 2014
13. L Derczynski, D Maynard, G Rizzo et al., Analysis of named entity recognition and linking for tweets. Inform. Process. Manage. **51**(2), 32–49 (2015)
14. G Xu, H Li, J Guo, *Named Entity Recognition in Query: US, US9009134*, 2015