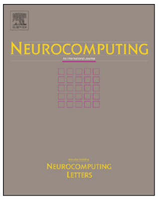
Contents lists available at ScienceDirect

Neurocomputing

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



Sentiment analysis for Chinese microblog based on deep neural networks with convolutional extension features

Xiao Sun [a,*], Chengcheng Li [a], Fuji Ren [a,b]

[a] School of Computer and Information, Hefei University of Technology, TunXi Road No. 193, 230009 Hefei, Anhui, China
[b] Faculty of Engineering, The University of Tokushima, 770-8506 Tokushima, Japan

ARTICLE INFO

Article history:
Received 15 April 2015
Received in revised form
24 September 2015
Accepted 16 February 2016

Keywords:
DNN
Sentiment analysis
Microblog conversation
RBM
ClassRBM

ABSTRACT

Related research for sentiment analysis on Chinese microblog is aiming at the analysis procedure of posts. The length of short microblog text limits feature extraction of microblog. Tweeting is the process of communication with friends, so that microblog comments are important reference information for related post. A contents extension framework is proposed in this paper combining posts and related comments into a microblog conversation for features extraction. A novel convolutional auto encoder is adopted which can extract contextual information from microblog conversation as features for the post. A customized DNN (Deep Neural Network) model, which is stacked with several layers of RBM (Restricted Boltzmann Machine), is implemented to initialize the structure of neural network. The RBM layers can take probability distribution samples of input data to learn hidden structures for better high level features representation. A ClassRBM (Classification RBM) layer, which is stacked on top of RBM layers, is adopted to achieve the final sentiment classification label for the post. Experimental results show that, with proper structure and parameters, the performance of proposed DNN on sentiment classification is better than state-of-the-art surface learning models such as SVM or NB, which proves that the proposed DNN model is suitable for short-length document classification with the proposed feature dimensionality extension method.

1. Introduction

Microblog has been widely popularized in recent years. It acts not only as a way for interaction and communication among people, but also as a way to express individual emotions at work or in daily life. Some related studies measure the preferences and political orientations of microblogger through sentiment analysis of microblog text. The emotion polarity of bloggers might reflect his or her hobbies and interests [1–4]. Microblog sentiment classification emerges as a challenging task [5–8]. The general realization of emotion polarity detection commonly includes extraction of features [9–11] and selection of machine learning methods [12–15].

Extraction of features from text is a process that produces some vector representations [16–19] for essential features and characteristics of original observation contents (or texts). At present, there are commonly two types of feature extraction methods: literal contents based method [20–23] and external knowledge information based method [24–28]. The methods based on literal contents are to calculate the probabilities of character or word sequences in text under the precondition of target classification using statistics methods. For example, a N-gram system is designed to solve the web categorizing problem [16], which takes Chinese character as minimal unit of textual feature representation for Chinese, as Chinese word segmentation is considered to cause loss of semantic or syntactic information and some Chinese word segmentation could not handle out-of-vocabulary words which might lead to the problem of word meaning ambiguities. In contrast, most researchers adopt word level frequencies and features because word is the least unit of semantic representation [22,23]. Related algorithms are adopted to calculate the feature weight of a word, such as CHI, IG, and TF-IDF. These textual feature extraction methods are widely adopted. These methods mostly consider morphological connections between words but the meaning embedded in texts is ignored or seldom considered. Sentiment is semantic information embedded in text, so some deep learning method should be adopted to detect such deep semantic knowledge. The method based on external knowledge information is applied to analyze word sense, semantics and grammatical structure and so on [27,28]. The knowledge information mainly includes system sentiment word lexicons, expression rules, syntactic model, etc. The sentiment polarities of words are based

[*] Corresponding author.
E-mail address: sunx@hfut.edu.cn (X. Sun).

http://dx.doi.org/10.1016/j.neucom.2016.02.077
0925-2312/© 2016 Elsevier B.V. All rights reserved.

on subjective opinion under certain circumstances or in specific field, which means that a certain sentiment dictionary is usually domain related and not universal. Choi [17] proposed a novel method to transform existing sentiment lexicon for specific domain into a new one to reflect the characteristics of new domain more directly, so that it can be used in cross-domain sentiment analysis. Pang [18] and Saif [19] use sentiment words as additional features combined with domain characteristics, and they find that semantic word features produce better recall and F-score than word frequency features. Ye [25] presents an improved semantic oriented approach for sentiment classification of Chinese movie reviews. This semantic approach introduces two-word phrases patterns based on POS (parts-of-speech), which is primarily applied in English movie review classification. Because of different language expression structure of English and Chinese, the method does not have great effects on classification results. Hiroshi [22], Wilson [23] and Subrahmanian [24] find the dependency relationship among words through building syntactic parsing tree for a sentence. The parsing tree includes semantic structure for the whole sentence and grammatical role of words. The word features are adjusted by considering sentence modifiers and syntactic structure information, which are treated as classification features. By adopting semantic features, the approach is proved to have better performances than literal content-based feature extraction methods.

The selection of machine learning methods is the process of selecting a proper classifier with tuned parameters for specific task. From the structure aspect, there are mainly two kinds of machine learning methods: surface learning models and deep learning models. Surface learning models can be regarded as a model with single hidden layer. For example, SVM, Boosting and Logistic Regression etc. belong to surface machine leaning model. Ye [25] compares three surface machine learning algorithms (SVM, Naïve Bayes and N-gram model) for review sentiment classification and proves that all three approaches reached accuracies of at least 80% on specific dataset. In Abbasi's work [26], the utility of semantic and syntactic features is integrated according to the characteristics of target text and learned by EWGA (Entropy Weighted Genetic Algorithm) and SVM model. The results indicate high performance in their datasets. These surface models require large number of labeled experiment data, and the common characteristic of such models is the limited ability of complicated object function or data representation. Deep leaning model could learn complex object function through building a deep nonlinear multi-layer network structure. The deep learning model has brought lots of attentions recently as a hot research topic in many fields such as image and speech processing. Xavier Glorot [10] proposes a deep learning approach which shows linear classifiers trained with higher-level learned feature representation of reviews outperforming traditional surface learning methods. Deselaers [11] and Chen [5] testify the validity of deep learning on NLP tasks. The deep learning model is showing its abilities of learning deep structure knowledge such as semantic information embedded in text, so deep learning models could solve classification problem with relatively better performance than surface learning models.

In tasks of micriblog sentiment analysis, the brevity of a post limits the feature expression and the feature vector extracted is excessively sparse as the average length of Chinese microblog post is 13 Chinese Hanzi in average. Short length of a post caused that traditional feature extraction method is hard to build reasonable representation of a sentence for machine learning. This paper presents a convolutional content extension feature extraction method for Chinese microblog sentiment classification. Compared with traditional feature extraction approaches discussed above, the comments of a post are adopted to expand feature dimensionality for sentiment analysis of post in microblog. The proposed method combines post and comments to form a new microblog conversation while extracting microblog textual feature to extend the feature dimensionality and solve the feature sparseness problem. This paper proposes a feature auto-encoder, named Conversation to Sentence Convolutional Auto Encoder (ConCAE), to extract the context information of a post from microblog conversation. Furthermore, a specific DNN model is constructed by stacking a Classification RBM [3,13,14] layer and several RBM layers together. At last, we design some experiments to choose proper feature set and optimal structure (including parameters) for the proposed DNN model. In the experiments, some comparisons are also performed on some public corpus such as Sina weibo to prove the effectiveness of proposed methods for microblog sentiment analysis.

2. Extended feature extraction for short text in microblog

Chinese microblogging posts are short texts, from which features extracted are limited because of its brevity [30]. The commonly used method for long text classification such as bag of word might cause the problem of feature sparseness [31]. In order to solve such problem, this paper proposes a content extension method for feature extraction in Chinese microblog sentiment analysis problem. For a post, it might be followed by several comments. These comments are responses or references to the emotion of microblogger. The content of a post is the key feature for sentiment analysis and its comments are used as assistance features. We combine a post with its comments into a microblog conversation and extract sentiment related information from the microblog conversation by ConCAE. The content extension method employs filtering approach to obtain proper sized microblog conversation which is composed of the post and its fixed-sized comments. The first step is to capture the emotion and semantic information of words. It is supposed that every post and its comments is consisted by m word $\{w_1, w_2, \ldots, w_m\}$, the word information w_i is the integration of emotion information w_{ei} and corresponding semantic information w_{si}. We compose size-fixed word bag and expression information to represent the feature of a post and its comments. Then the post feature V_{post} and comment feature sets $\{V_{com}^1, V_{com}^2, \ldots, V_{com}^L\}$ can be obtained, where L is the number of comments. After word bag and expression information are obtained, the auto-encoding network is adopted to capture context information for the post with its comments.

2.1. Word-level feature extraction

The aim of word feature extraction is to acquire emotional, semantic and integration information of words [12]. The steps of word information extraction for a post are based on word segmentation and semantic analysis of the sentence. The emotional information of a word can be obtained from prior knowledge, such as a system lexicon or dictionary. Semantic information of each word in microblog is semantic role label, which can be obtained by analyzing the semantic structure of the microblog content.

2.1.1. Emotion information of words

The steps of emotion feature extraction are based on word segmentation and semantic analysis for the sentence. Word is the smallest meaning expression units in English and Chinese, while Chinese word expression is composed of Chinese Hanzi (single character) or words. Some Chinese Hanzi has no meanings. For Chinese microblog sentiment analysis, the first step is Chinese word segmentation and part-of-speech (POS) tagging. All