Maximum Entropy Model based on Feature Extraction for Sentiment Detection of Text

Jun Li[1, a], Wei Jin[2, b], Zihao Zhang[3]

[1] School of Data and Computer Science, Sun Yat-Sen University, Guangzhou, 510006, China

[2] School of Data and Computer Science, Sun Yat-Sen University, Guangzhou, 510006, China

[3] School of Data and Computer Science, Sun Yat-Sen University, Guangzhou, 510006, China

[a]email: zsdlijun@163.com, [b]email:sky__jinwei@163.com

Keywords: Sentiment Detection; Feature Extraction; Maximum Entropy Model

Abstract. The rapid development of social media services has facilitated the communication of opinions through online news, blogs, post bar, microblogs/tweets, and so forth. This article concentrates on the mining of emotions evoked by newmaterials. Compared to the classical sentiment analysis by using the word-emotion lexicon in the text, we combine the word with emotion via the intensive feature functions. We propose a maximum entropy model based on the feature extraction for sentiment classification, which generates the probability of sentiments conditioned to news text. In addition, one effective feature extraction strategies are proposed to refine the original miscellaneous news text. Experimental evaluations using real-world data validate the effectiveness of the proposed model on sentiment classification of news text.

Introduction

Since the world go into the internet century, thousands of billions of texts are published on the internet every day. The texts contains lots of valuable information which can be used in different ways. However, it is not realistic to deal with such huge amount of texts with manpower. Therefore, applying machine learning method to this work can give great help.

Among the information from the texts on the internet, sentiment of the text is the great valuable one. News on the internet are published in time in contrast to other media, and it can show what happen in the world right now. What is more, sentiment in the news can be used to show the public's attitude and predict the trend.

In this paper, we focus on the sentiment classification on the news texts published on the internet. Therefore, this work can help to keep in touch with public's sentiment. We propose a framework to deal with the text to get the sentiment information. In our work, we apply singular value decomposition to the data sets to reduce dimension and restrain noise, and besides, we propose to model sentiments and words using intensive feature functions. Thus, sentiments of unlabeled news text is classified according to the principle of maximum entropy [1]. The result of the experiments on the real data sets shows that the framework we propose can gain a good result on the prediction of sentiment from text.

The remainder of this paper is organized as follows. In Section 2, we summarize the related work in sentiment classification on the text. Then, we propose our model and conduct experimental analysis in Section 3 and Section 4, respectively. Finally, we present conclusions in Section 5.

Related work

In this section, previous works on sentiment classification, news text analysis and our work will be shown.

Sentiment classification mainly concentrate on extracting emotions from reviews, messages and news text, which convey the opinion of writers or readers. The existing methods of sentiment classification can be divided into three categories primarily: lexicon-based, supervised and unsupervised learning strategies. The lexicon-based method [2-4] classified sentiments by constructing word- or topic-level emotional dictionaries. The supervised learning strategy used existing classification algorithms to split the emotional orientation of words or phrases into positive and negative, which included naive Bayes, maximum entropy and support vector machines [5]. An unsupervised learning technique was also utilized to classify the emotional orientation of users' reviews (e.g., reviews of movies, travel destinations, automobiles and banks), which computed the overall polarity of the review by counting the occurrence of positive and negative terms [6]. Recently, the Emotion-Term(ET) algorithm and the Emotion-Topic Model(ETM) [7] were proposed to improve the performance of existing systems. ET is a variant of the naive Bayes classifier and ETM is model associating emotions with topics jointly.

We analysed the features of text which has many redundant information and noise. So we reduced the dimension and remove some noise data of the original data. And we extracted the main features to represent one document, then we develop an intensive maximum entropy model to classify sentiments of these features, which has a concentrated representation for modeling emotions.

Maximum Entropy Model via Intensive Feature Functions

In this section, we propose Singular Value Decomposition(SVD) and the Maximum Entropy Model based on Feature Extraction(MEFE), a maximum entropy model via intensive feature functions for text sentiment classification. The problem is first defined, including the relevant terms and notations, and then the MEFE is presented in detail. Finally, we describe the estimation of parameters.

Problem Definition

For convenience of defining the issue of sentiment classification in news, we here defined the following terms and notations:

A text collection consists of D documents $\{t_1, t_2, ..., t_D\}$ with word tokens and multiple emotion labels. We represent the set of all word tokens by $W = \{w_1, w_2, ..., w_V\}$ and emotion labels by $E = \{e_1, e_2, e_3, ..., e_M\}$ where V is the number of unique word tokens, and M is the amount of pre-defined emotion labels such as "joy", "anger", "fear", "surprise", "touching", "empathy", "sadness", "boredom" and "warmness". The co-occurrence of word token w and emotion label e in the training set is denoted by (w, e), and we denote the set of all word-emotion pairs by Φ. For the multiple emotion labels and reader ratings, a matrix Θ is used to represent the reader votes over each emotion label for all documents in the training set. In the $D \times M$ matrix Θ, the element Θ_{nk} denotes the ratings over the $k-th$ emotion among all users who have read the $n-th$ text. The reader ratings were normalized and summed to one for each document. Table 1 summarizes the notations of frequently-used variables.

Table 1: Notations of frequently-used variables

Symbol	*Description*
D	Number of documents
V	Number of unique word tokens
M	Number of pre-defined emotion labels
F	Number of unique features
W	Set of word tokens
E	Set of social emotions
Φ	Set of all word-emotion pairs
Θ	$[\theta_{nk}]$: $D \times M$ matrix of normalized ratings for each emotion

Singular Value Decomposition

Singular Value Decomposition (SVD) is an important method to reducing dimension and extract the main features of high dimension data. Then we will introduce the SVD method for our Sina news text data.

First, we should construct the document-word matrix while row represent a document and the array represent a word appeared in documents. The matrix is represented in C. If C is a square matrix whose number of rows and arrays are same, we could decompose this matrix easily. There exists a formula: $C = U\Lambda U^{-1}$, and the arrays of U is the feature vector of C. Λ is a diagonal matrix, the value in diagonal line is the feature value of matrix C, and they arrange in descending order.

$$\Lambda = \begin{pmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \cdots & \\ & & & \lambda_M \end{pmatrix} \text{,where } \lambda_i \geq \lambda_{i+1}$$

In the most time, the document-word matrix is not a square matrix so we couldn't decompose it directly. If the matrix C is a $M \times N$ matrix, U is a $M \times M$ matrix, the array of U is the orthogonality feature vector of CC^T, V is a $N \times N$ matrix, the array of V is the orthogonality feature vector of C^TC, r is the rank of matrix C, there exist the singular value decomposition as follows: $C = U\Sigma V^T$, the feature values of CC^T and C^TC are same which are $\lambda_1, \lambda_2, ..., \lambda_r$ respectively. For $1 \leq i \leq r$, let $\sigma_i = \sqrt{\lambda_i}$, with $\lambda_i \geq \lambda_{i+1}$, then the $M \times N$ matrix Σ is composed by setting $\Sigma_{ii} = \sigma_i$ for $1 \leq i \leq r$, and zero otherwise. σ_i is singular value of matrix C.

Then we can select the largest k singular value in matrix Σ to construct an approximate matrix Σ_k, while this matrix represent the main features of the original data. Calculate the new document-word matrix $C_k : C_k = U_k \Sigma_k V_k^T$.

Through the conversion we have described above we can get the new matrix C_k to represent the original document, we can select the main words to represent a document in this matrix that they are the key features of this document. So we have reduced the dimension and remove some noise of the original data by the SVD method. The specific algorithm is shown as Algorithm 1.

Algorithm 1 Algorithm for SVD

```
Input:
    Word token set $W$;
Output:
    Optimal number of features in the news text;
    1. Construct document-word matrix $C$;
    2. Decompose the matrix $C$ via the formula under:
       $C = U\Sigma V^T$
       where $C$ is the original word-document matrix, $U$, $\Sigma$ and $V^T$ are the matrices decomposed by SVD method;
    3. Calculate the new matrix $C_k$ approximate to matrix $C$: $C_k = U_k \Sigma_k V_k^T$
       where $\Sigma_k$ is an approximate matrix constructed by the largest $k$ singular value in matix $\Sigma$. $U_k$ is the first $k$ rows extracted by $U$ matrix and $V_k$ is the first $k$ columns extracted by $V$ matrix. Then $C_k$ is the approximate matrix of $C$ generated using the formula above;
    4. Set threshold value to 0.9; Extract the words whose weight value exceed threshold in $C_k$.
```

Maximum Entropy Model

In this section, we first briefly introduce the principle of entropy and maximum entropy model [1], and then introduce maximum entropy model for sentiment classification of text.

Entropy is the average amount of information contained in each message. The general idea is that the less likely an event is, the more information it provides when it occurs, i.e., the larger entropy it has. The probability distribution of the events, coupled with the information amount of each event, forms a random variable whose expected value is the average amount of information, generated by this distribution.

The principle of maximums entropy indicates that when predicting the probability distribution of a random event, the distribution should satisfy all our prior conditions and knowledges (e.g. the training set that expressed testable information), and make none subjective assumptions about the unknown case. Under this condition, the probability distribution has the largest value of entropy, and the error of prediction could be minimized.

Table 2: Samples of the training set

News text	Word tokens	Emotions and ratings
t_1	$\{w_3, w_4\}$	$e_1:\theta_{11}, e_2:\theta_{12}, e_3:\theta_{13}$
t_2	$\{w_1, w_2\}$	$e_1:\theta_{21}, e_2:\theta_{22}, e_3:\theta_{23}$
t_3	$\{w_2, w_3, w_4\}$	$e_1:\theta_{31}, e_2:\theta_{32}, e_3:\theta_{33}$

In our task of social emotion classification, the prior conditions are the co-occurrences of word tokens and emotion labels. Table 2 shows the samples of a training set, where $D=3$, $V=4$, $M=3$, and $F=12$. To make a concentrated representation of modeling emotion labels and the word tokens in text, we propose the intensive feature function as follows:

$$f_i(w,e)=\begin{cases}1 & w\in \mathrm{W}, e\in \mathrm{E} \text{ and } (w,e)\in\Phi \\ 0 & otherwise.\end{cases} \tag{1}$$

We then define the empirical probability distribution of the training set $\bar{p}(w,e)$, as follows:

$$\bar{p}(w,e)=\frac{1}{|\Phi|}\times\sum_{t_n\in \mathrm{T}_j}\Theta_{nk} \tag{2}$$

where T_j is the collection of documents that contain w_j.

The expected value of feature functions $f_i(w,e)$ with respect to the empirical distribution can be estimated by:

$$\bar{E}(f_i)=\sum_{w,e}\bar{p}(w,e)f_i(w,e) \tag{3}$$

The expected value of $f_i(w,e)$ with respect to the probability of emotion label e conditioned to word token w, i.e., $p(e\mid w)$ is derived as follows:

$$E(f_i)=\sum_{w,e}\bar{p}(w)p(e\mid w)f_i(w,e) \tag{4}$$

where $\bar{p}(w)$ is the empirical distribution of w in the training set. Thus, the first constraint condition of the MEFE is as follows:

$$E(f_i)=\bar{E}(f_i) \tag{5}$$

According to Eq.(3) and Eq.(4), we get

$$\sum_{w,e}\bar{p}(w)p(e\mid w)f_i(w,e)=\sum_{w,e}\bar{p}(w,e)f_i(w,e) \tag{6}$$

A mathematical measure of the uniformity of the conditional distribution $p(e\mid w)$ is provided by the conditional entropy:

$$H(P)=-\sum_{w,e}\bar{p}(w)p(e\mid w)\log p(e\mid w) \tag{7}$$

Then, the MEFE is formulated as the following optimization problem:

$$\begin{aligned} &\text{maximize } H(P)=\sum_{w,e}\bar{p}(w)p(e\,|\,w)\log\frac{1}{p(e\,|\,w)} \\ &\text{subject to } \quad E(f_i)-\bar{E}(f_i)=0 \qquad 1\le i\le F \\ &\qquad\qquad\quad \sum_{e}p(e\,|\,w)-1=0 \quad \text{for all } w \end{aligned} \tag{8}$$

To estimate the value of $p(e\,|\,w)$ that maximizes $H(P)$, we resolve the above primal optimization problem to an unconstrained dual optimization problem by introducing the Lagrange parameters λ, as follows:

$$p_{\lambda}(e\,|\,w)=\frac{1}{Z_{\lambda}(w)}\exp(\sum_{i=1}^{F}\lambda_i f_i(w,e)) \tag{9}$$

$$Z_{\lambda}(w)=\sum_{e}\exp(\sum_{i=1}^{F}\lambda_i f_i(w,e)) \tag{10}$$

Parameter Estimation

In terms of a mountain of information irrelative to social emotions in original news text, the motivation of MEFE is to improve the robustness of our model in the extremely miscellaneous training set, where a part of noisy features may largely decrease the accuracy of social emotion classification (as will be illustrated in Section 4). To estimate the parameters of MEFE, i.e., λ, we use an iterative method as shown in Algorithm 2 after reducing the redundant features by SVD .

```
Algorithm 2 Iterative algorithm for MEFE
Input:
    Feature functions $f_1, f_2, \dots, f_n$
    Empirical distribution $\bar{p}(w,e)$;
Output:
    Optimal values of each parameter $\lambda_i$;
    Set $\lambda_i^{(0)}$ to some initial values, e.g.: $\lambda_i^{(0)}=0$.
    Repeat
        $\lambda_i^{(r+1)}=\lambda_i^{(r)}+\frac{1}{C}\log\frac{\bar{E}(f_i)}{E^{(r)}(f_i)}$
    until convergence
    where $r$ is the iteration index and the constant $C$ is defined
    as follows:
        $C=\max_{w,e}\sum_{i=1}^{n}f_i(w,e)$
```

After estimating the optimal values of each parameter λ_i, predicting the emotion label of unlabeled text is straightforward.

Table 3 presents an example of the testing set. Given and unlabeled text t with three words tokens $\{w_1,w_2,w_3\}$, and two predefined emotion $\{e_1,e_2\}$, we get six intensive feature functions in total.

Table 3: Samples of the testing set.

predifined emotion label	w_1	w_2	w_3
e_1	f_1	f_5	f_6
e_2	f_2	f_3	f_4

According to Eq.(9) and Eq.(10), we have:

$$p_\lambda(e_1 \mid t) = \exp(\lambda_1 f_1 + \lambda_5 f_5 + \lambda_6 f_6) / Z_\lambda(w)$$
$$p_\lambda(e_2 \mid t) = \exp(\lambda_2 f_2 + \lambda_3 f_3 + \lambda_4 f_4) / Z_\lambda(w) \quad (11)$$
where
$$Z_\lambda(w) = \exp(\lambda_1 f_1 + \lambda_5 f_5 + \lambda_6 f_6) + \exp(\lambda_2 f_2 + \lambda_3 f_3 + \lambda_4 f_4)$$

Experiments

In this section, we evaluate the performance of the MEFE for sentiment classification of text. we designed the experiments to achieve the following goals: (i) to analyze the influence of number of iterations on the accuracy of sentiment classification in the MEFE via comparing with the Maximum Entropy Model (MEM), and (ii) to conduct comparative analysis with various baselines.

Data Set

To test the effectiveness of the proposed model, we collected 500 news from the society channel of Sina (news.sina.com.cn/society/). The news, and user ratings across eight emotions (i.e.,touching, empathy, boredom, anger, amusement, sadness, surprise, and warmness) were gathered. The publishing dates of the news range from January to April of 2012. To ensure that the stability of user ratings, the data set was crawled from half a year after the publishing date. Table 4 summarizes the statistics for each emotion label of the data set. The number of titles for each emotion label represents the amount of the news that had the highest ratings for that emotion. For example, there are 65 news that had the highest user ratings for "Touching", with a total number of ratings of 4,298 for that emotion.

Table 4: Statistics of the data set.

Emotion label	Number of titles	Number of ratings
Touching	65	4,298
Empathy	47	1,557
Boredom	53	2,159
Anger	145	10,840
Amusement	80	5,579
Sadness	55	2,395
Surprise	31	1,587
Warmness	24	911

In the preprocessing step, a Chinese lexical analysis system (ICTCLAS) is used to perform the Chinese word segmentation. ICTCLAS is an integrated Chinese lexical analysis system based on multi-layer HMM. We random select 80 percent of news as the training set and the rest as the testing set. The MEM, Naive Bayes (NB) and Emotion-Topic Model (ETM) [7] were implemented for comparison. All hyper parameters were set at default. We also included a dummy algorithms, MaxC as the baselines. MaxC always picks the emotion label with the largest total user ratings in the training set. To make an appropriate comparison with other models, the accuracy was employed as the indicator of performance [7]. The accuracy is essentially the micro-averaged F1 measure, which equally weights precision and recall.

Experiment Design

Our experiment consists of two main parts: (1)Preprocess the original data which is the Sina news text data. We divide the text data into some separate words and remove some stop words without any practical significance. Then we reduce the dimension of this text data using the SVD method and remove some noise and abnormal data in this process. (2)Train these data have been processed using maximum entropy model, then test the testing data by this model when getting the appropriate parameters. Calculate the accuracy of this model in sentiment classification.

The main purpose of this experiment design is that text information is complex and redundancy, we can remove noise and outliers by SVD reducing dimension method. So we can extract the main feature of this text, then we can train the maximum model with the main feature, which will improve the efficiency and accuracy of this algorithm effective.

We also have done an experiment to compare the effect between text data without SVD reducing dimension and with reducing dimension. So we can explore the advantage of classification after reducing dimension more.

Influence of the Number of Iterations

To evaluate the influence of iteration number, we varied the number of iterations from 1 to 20. Fig.1 presents the performance of MEFE, MEM when using different numbers of iterations.

We found that MEFE outperformed MEM with the improvement ratio of 4.5% on average. The result indicate the proposed algorithm can capture the relationship between social emotions and word tokens more accurately. The reason is that for the Sina News data set, a multitude of miscellaneous features which is irrelative to social emotions will influence the prediction. According the parameter estimation section discussed above, the sum of the weight corresponding to these features (i.e., λ) can be very large for a great amount of features, despite of small weight of each feature. Hence, the irrelative data can influence the social emotions prediction in the MEM. We also found that the accuracy of MEFE was more steady than that in MEM as the increasing of the number of iterations from the curves shown in the Figure 1, thus indicate the robustness of parameter estimation in MEFE outperformed the MEM.

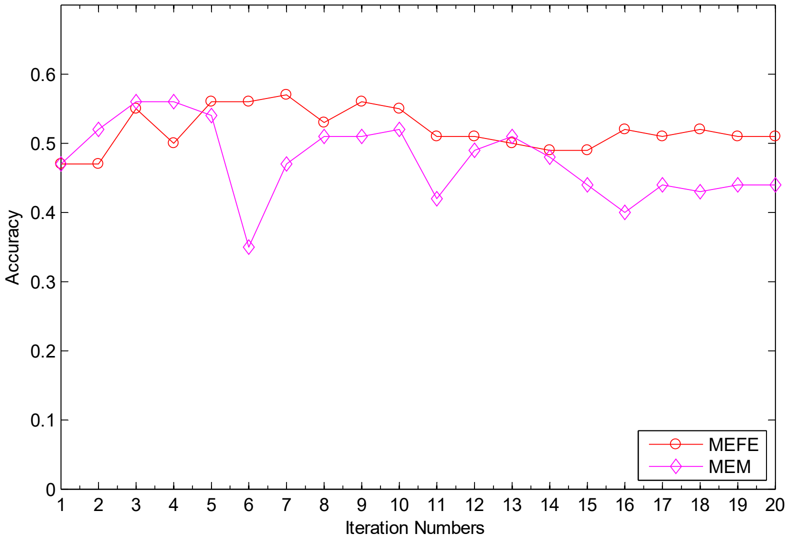

Fig.1. Performance with different iterative times

Comparison with Baselines

In this section, we measure and compare the performance of different models on sentiment classification of news text comprehensively. The accuracy of MEFE and four baselines is present in Table 5. Compared to the baselines of MEM, ETM, NB, and MaxC, the accuracy of MEFE improved 15.95%, 10.67%, 53.67%, and 96.23%, respectively.

Table 5: Statistics of different models

Models	Accuracy(%)	Improvement(%)
MEFE	51.02	—
MEM	44.00	15.95
ETM	46.10	10.67
NB	33.20	53.67
MaxC	26.00	96.23

Conclusion

Sentiment classification is helpful for understanding the preferences and perspectives of online users, and therefore can facilitate the provision of more relevant and personalized services, including hybrid search in social media [8], construction of user profiles [9], financials analysis [10], emotionbased text retrieval [11], and social emotion detection [12]. In this paper, we have proposed maximum entropy model based on feature extraction for sentiment classification of news text. We evaluate our model on real-world data and compare it to three existing models. The result show that our approach outperforms those baselines.

References

[1] A. Ratnaparkhi. Maximum entropy models for natural language ambiguity resolution [J]. Proc of the IEEE, 1998.

[2] Rao Y, Lei J, Liu W, et al. Building emotional dictionary for sentiment analysis of online news[J]. World Wide Web-internet & Web Information Systems, 2014 17(4) 723-742.

[3] Rao Y, Li Q, Mao X, et al. Sentiment topic models for social emotion mining [J]. Information Sciences, 2014 266(5) 90-100.

[4] Rao Y, Li Q, Liu W, et al. Affective topic model for social emotion detection [J]. Neural Networks the Official Journal of the International Neural Network Society, 2014 58(5) 29–37.

[5] Pang B, Lee L, Vaithyanathan S. Thumbs up? Sentiment Classification using Machine Learning Techniques [J]. Proceedings of Emnlp, 2002 79-86.

[6] Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews[A]. Meeting on Association for Computational Linguistics[C]. 2002.

[7] Bao S, Xu S, Zhang L, et al. Mining Social Emotions from Affective Text [J]. IEEE Transactions on Knowledge & Data Engineering, 2012 24(99) 1-1.

[8] Xie H, Li Q, Mao X, et al. Mining Latent User Community for Tag-Based and Content-Based Search in Social Media [J]. Computer Journal, 2014 57(9) 1415-1430.

[9] Xie H, Li Q, Mao X, et al. Community-aware user profile enrichment in folksonomy [J]. Neural Networks the Official Journal of the International Neural Network Society, 2014 58(5) 111-121.

[10] Li X, Xie H, Chen L, et al. News impact on stock price return via sentiment analysis [J]. Knowledge-Based Systems, 2014 69(1) 14-23.

[11] Ranking social emotions by learning listwise preference [A]. Proceedings 2011 First Asian Conference on Pattern Recognition[C]. 2011.

[12] Lei J, Rao Y, Li Q, et al. Towards building a social emotion detection system for online news[J]. Future Generation Computer Systems, 2014 37(7) 438-448.