Intensive Maximum Entropy Model for Sentiment Classification of Short Text

Article *in* Lecture Notes in Computer Science · April 2015

DOI: 10.1007/978-3-319-22324-7

Yanghui Rao
City University of Hong Kong

Haoran Xie
The Education University of Hong Kong

Intensive Maximum Entropy Model for Sentiment Classification of Short Text

Yanghui Rao[1], Jun Li[1(✉)], Xiyun Xiang[1], and Haoran Xie[2]

[1] School of Mobile Information Engineering, Sun Yat-sen University, Zhuhai, China
raoyangh@mail.sysu.edu.cn, {lijun223,xiangxy3}@mail2.sysu.edu.cn
[2] Caritas Institute of Higher Education, Hong Kong, China
hrxie2@gmail.com

Abstract. The rapid development of social media services has facilitated the communication of opinions through microblogs/tweets, instant-messages, online news, and so forth. This article concentrates on the mining of emotions evoked by short text materials. Compared to the classical sentiment analysis from long text, sentiment analysis of short text is sometimes more meaningful in social media. We propose an intensive maximum entropy model for sentiment classification, which generates the probability of sentiments conditioned to short text by employing intensive feature functions. Experimental evaluations using real-world data validate the effectiveness of the proposed model on sentiment classification of short text.

Keywords: Sentiment classification · Short text analysis · Intensive maximum entropy model

1 Introduction

Nowadays, the thriving of short text-based social media services has provided us with torrents of news on diverse topics and entities. Leveraging by the convenience of communication among online users, microblogs/tweets and instant-messages are constantly filled with opinions towards a large amount of topics such as politics, sports and other prevailing topics. Thus, it is important for us to identify and classify sentiments from short texts automatically.

Different from the normal documents, the number of words is few and most words only occur once in each short text. Thus, for the tasks of sentiment classification and annotation, it is usually impossible to classify or annotate emotions consistently using the limited information contained in short text [1]. Most existing approaches enriched the context of short text by retrieving relevant

Jun Li—The research work described in this article has been substantially supported by "the Fundamental Research Funds for the Central Universities" (Project Number: 46000-31610009).

A. Liu et al. (Eds.): DASFAA 2015 Workshops, LNCS 9052, pp. 42–51, 2015.
DOI: 10.1007/978-3-319-22324-7_4

long documents from the web, which may result in waste of both memory and computation time. In light of these considerations, we first propose to model sentiments and the limited words using intensive feature functions, Then, sentiments of unlabeled short text is classified according to the principle of maximum entropy [2]. Experimental results show that the proposed model can effectively integrate prior conditions from the sparse text.

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

2 Related Work

In this section, we review some previous works on sentiment classification and short text analysis, which shed light on sentiment classification of short text for our work.

Sentiment Classification. Sentiment classification mainly concentrate on extracting emotions from reviews, messages and news documents, 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 [3–5] 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 naïve Bayes, maximum entropy and support vector machines [6]. 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 [7]. However, those lexicon-based, supervised or unsupervised learning strategies were mainly designed for long text.

Short Text Analysis. The main feature of short text is the sparsity of words. Due to the fact that most words only occur once in each short text, it is difficult to accurately conduct classification, clustering, retrieval and other tasks.

In a preliminary work, Sahami and Heilman [8] defined a web-based kernel function to measure the semantic similarity of short text snippets. The general process was as follows: First, each short text was inputted to the web search engine as a query. Second, a large amount of retrieved long documents was combined to enrich the context of each short text. Third, the vector space model [9] was used to estimate the similarity values. Banerjee et al. [10] also proposed a method of enhancing the representation of short text by including the additional features from Wikipedia. However, due to the high-dimensional vectors, those methods may be memory or time-consuming. To address this issue, a collapsed Gibbs sampling algorithm for the Dirichlet multinomial mixture model was proposed [11]. Unfortunately, the model was designed for short text clustering rather than classification.

Our Work. Research into sentiment classification of short text began with the "affective text" in SemEval-2007 tasks [1], which aimed to annotate news headlines according to the predefined emotions. In the SWAT system [1], a word-emotion mapping dictionary was first constructed, in which, each word was scored according to multiple emotion labels. Then, the dictionary was used to classify the emotions of unlabeled news headlines. Recently, the emotion-term(ET) algorithm and the emotion-topic model(ETM) [12] were proposed to improve the performance of existing systems. ET is a variant of the naïve Bayes classifier and ETM is model associating emotions with topics jointly. Nevertheless, experimental results have shown that the performance of sentiment classification of short text is limited for SWAT, ET and EMT [3]. The reason may be that short documents lack enough context from which statistical conclusions can be drawn earily [13]. We here develop an intensive maximum entropy model to classify sentiments of short text, which has a concentrated representation for modeling emotions and the limited words.

Table 1. Notations of frequently-used variables.

Symbol	Description
V	Number of unique word tokens
N	Number of short text
M	Number of emotion labels
Φ	Set of all word-emotion pairs
Θ	Set of distinct word-emotion pairs

3 Maximum Entropy Model via Intensive Feature Functions

In this section, we propose the intensive maximum entropy model(IMEM), a maximum entropy model via intensive feature functions for short text sentiment classification. The problem is first defined, including the relevant terms and notations, and then the IMEM is presented in detail. Finally, we describe the estimation of parameters.

3.1 Problem Definition

For convenience of defining the issue of short text sentiment classification, and describing our intensive maximum entropy model, we here defined the following terms and notations:

A document collection consists of N short text $\{t_1, t_2, \ldots, t_N\}$ with word tokens and an emotion label. We represent the list of word tokens and emotion labels by $\{w_1, w_2, \ldots, w_V\}$ and $e = \{e_1, e_2, e_3, \ldots, e_M\}$, respectively, where V is the numble of unique terms, and M is the amount of emotion labels. The common instances of emotion labels are "joy", "anger", "fear", "surprise", "touching", "empathy", "boredom", "sadness", "warmness", etc. Table 1 summarizes the notations of these frequently used variables.

3.2 Intensive Maximum Entropy Model

In this section, we first briefly introduce the principle of entropy and maximum entropy model [2], and then present out intensive maximum entropy model for sentiment classification of short 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 ors entropy, and the error of prediction could be minimized.

Table 2. Samples of the training set.

Short text	Word tokens	Emotion Label
t_1	$\{w_3, w_4\}$	e_1
t_2	$\{w_1, w_2\}$	e_2
t_3	$\{w_1, w_3\}$	e_3
t_4	$\{w_2, w_3, w_4\}$	e_1

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

$$f(w,e) = \begin{cases} 1 & w \in \{w_1, w_2, \ldots, w_V\}, e \in \{e_1, e_2, \ldots, e_M\} \ and \ (w,e) \in \Theta \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

where Θ is the set of distinct word-emotion pairs that occurred in the training set, i.e., $\Theta = \{(w_1, e_2), (w_1, e_3), (w_2, e_1), (w_2, e_2), (w_3, e_1), (w_3, e_3), (w_4, e_1)\}$. We then define the empirical probability distribution of the training set $\bar{p}(w, e)$, as follows:

$$\bar{p}(w,e) = \frac{1}{|\Phi|} \times count(w,e) \tag{2}$$

where Φ is the set of all word-emotion pairs in the training set, $|\Phi| = 9$, and $count(w, e)$ is the number of times that (w, e) occur in Φ.

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

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

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

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

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

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

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

$$\sum_{w,e} \bar{p}(w) p(e|w) f(w, e) = \sum_{w,e} \bar{p}(w, e) f(w, e) \tag{6}$$

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

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

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

$$\text{maximize } H(P) = \text{argmax} \sum_{w,e} \bar{p}(w) p(e|w) \log \frac{1}{p(e|w)} \tag{8}$$

subject to

$$E(f) - \bar{E}(f) = 0 \tag{9}$$

$$\sum_{e} p(e|w) - 1 = 0 \text{ for all } w \tag{10}$$

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

$$p_\lambda(e|w) = \frac{1}{Z_\lambda(w)} exp(\sum_{i=1}^{|\Phi|} \lambda_i f_i(w, e)) \tag{11}$$

$$Z_\lambda(w) = \sum_{e} exp(\sum_{i=1}^{|\Phi|} \lambda_i f_i(w, e)) \tag{12}$$

3.3 Parameter Estimation

To estimate the parameters of IMEM, i.e., λ, we use an iterative method as shown in Algorithm 1.

Algorithm 1. Iterative algorithm for IMEM

Input:
Feature functions $f_1, f_2, \ldots, f_n$;
Empirical distribution $\bar{p}(w, e)$s
Output:
Optimal values of each parameterλ_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 t 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 short text is straightforward.

Table 3 presents an example of the testing set. Given and unlabeled short text t with three words tokens $\{w_1, w_2, w_3\}$, and two predefined emotion labels $\{e_1, e_2\}$, 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 Eqs. (9) and (10), we have:

$$p_\lambda(e_1|t) = \frac{1}{Z_\lambda(w)} exp(\lambda_1 f_1 + \lambda_5 f_5 + \lambda_6 f_6)$$

$$p_\lambda(e_2|t) = \frac{1}{Z_\lambda(w)} exp(\lambda_2 f_2 + \lambda_3 f_3 + \lambda_4 f_4)$$

where

$$Z_\lambda(w) = exp(\lambda_1 f_1 + \lambda_5 f_5 + \lambda_6 f_6) + exp(\lambda_2 f_2 + \lambda_3 f_3 + \lambda_4 f_4)$$

4 Experiments

In this section, we evaluate the performance of the IMEM for sentiment classification of short text. We designed the experiments to achieve the following two goals: (i) to analyze the influence of number of iterations on the accuracy of sentiment classification for the IMEM, and (ii) to conduct comparative analysis with various baselines.

4.1 Data Set

To test the effectiveness of the proposed model, we collected 4570 news headlines from the society channel of Sina (news.sina.com.cn/society/). The news headlines, 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 headlines 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 headlines that had the highest ratings for that emotion. For example, there are 749 news headlines that had the highest user ratings for "Touching", with a total number of ratings of 41,796 for that emotion. Figure 1 presents an example of emotion labels and user ratings, in which multiple emotion labels were voted by 3064 users for a particular news text.

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 headlines as the training set and the rest as the testing set. The existing SWAT [1], emotion-term (ET) and emotion-topic model (ETM) [12] 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 [12]. The accuracy is essentially the micro-averaged $F1$ measure, which equally weights precision and recall.

4.2 Influence of the Number of Iterations

To evaluate the influence of iterative times, we varied the number of iterations from 1 to 400 (the amount of different iterative times tested was 49 in total).

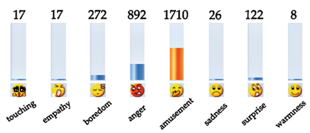

Fig. 1. Example of emotion labels and user ratings

Table 4. Statistics of the data set.

Emotion label	Number of titles	Number of ratings
Touching	749	41,796
Empathy	225	23,230
Boredom	273	21,995
Anger	2048	138,167
Amusement	715	43,712
Sadness	355	37,162
Surprise	167	11,386
Warmness	38	7,986

Figure 2 shows the performance of IMEM when using different numbers of iterations. The results indicate that the IMEM converges to its asymptote in less than 150 iterations. After its convergence, the variance of accuracy is noticeably less than that of initial values, Based on the above observation, we choose 300 iterations as the default setting unless otherwise specified.

4.3 Comparison with Baselines

In this section, we measure and compare the performance of different models on sentiment classification of short text comprehensively. The accuracy of IMEM and four baselines is present in Table 5. Compared to the baselines of ET, SWAT, ETM, and MaxC, the accuracy of IMEM improved 31.70 %, 15.64 %, 11.58 %, and 19.74 %, respectively.

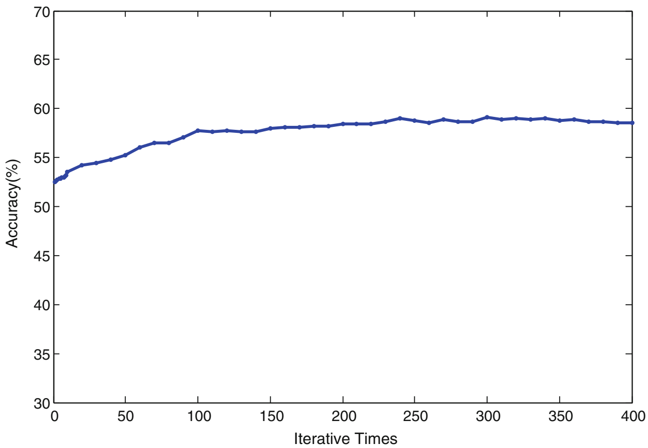

Fig. 2. Performance with different iterative times

Table 5. Statistics of different models.

Models	Accuracy(%)	Improvement(%)
IMEM	59.08	—
ET	44.86	31.70
SWAT	51.09	15.64
ETM	52.95	11.58
MaxC	49.34	19.74

5 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 [14], construction of user profiles [15], financials analysis [16], emotion-based document retrieval [17], and social emotion detection [18]. In this paper, we have proposed an intensive maximum entropy model for sentiment classification of short text. We evaluate our model on real-world data and compare it to four existing models. The result show that our approach outperforms those baselines.

In our subsequent study, we plan to extend out method to multi-label conditions, and employ topic modeling to generate concept-level feature functions.

Acknowledgements. The authors are thankful to the anonymous reviewers for their constructive comments and suggestions on an earlier version of this paper. The research described in this paper has been supported by "the Fundamental Research Funds for the Central Universities" (46000-31121401), and a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (UGC/FDS11/E06/14).

References

1. Katz, P., Singleton, M. and Wicentowski, R.: Swat-mp: the semeval-2007 systems for task 5 and task 14, In: The 4th International Workshop on Semantic Evaluations. ACL, pp. 308–313 (2007)
2. Ratnaparkhi, A.: Maximum entropy models for natural language ambiguity resolution. Encyclopedia of Machine Learning, pp. 647–651(2010)
3. Rao, Y.H., Lei, J.S., Wenyin, L., Li, Q., Chen, M.L.: Building emotional dictionary for sentiment analysis of online news. World Wide Web Internet Web Inf. Syst. **17**, 723–742 (2014)
4. Rao, Y.H., Li, Q., Mao, X.D., Wenyin, L.: Sentiment topic models for social emotion mining. Inf. Sci. **266**, 90–100 (2014)
5. Rao, Y.H., Lei, J.S., Wenyin, L., Wu, Q.Y., Quan, X.J.: Affective topic model for social emotion detection. Neural Netw. **58**, 29–37 (2014)
6. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up? sentiment classification using machine learning techniques. In: Empirical Methods in Natural Language Processing. ACL, pp. 79–86 (2002)

7. Turney, P.D.: Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, pp. 417–424 (2002)
8. Sahami, M., Heilman, T.D.: A web-based kernel function for measuring the similarity of short text snippets. In: Proceedings of the 15th International Conference on World Wide Web, WWW 2006, Edinburgh, Scotland, UK, 23–26 May 2006
9. Salton, G., Wong, A., Yang, C.S.: A vector space model for automatic indexing. Commun. J. ACM **18**(11), 613–620 (1975)
10. Banerjee, S., Ramanathan, K., Gupta, A.: Clustering short texts using wikipedia, In: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 787–788 (2007)
11. Yin, J.H., Wang, J.Y.: A Dirichlet multinomial mixture model-based approach for short text clustering. In: The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 233–242 (2014)
12. Bao, S.H., Xu, S.L., Zhang, L., Yan, R., Su, Z., Han, D.Y., Yu, Y.: Mining social emotions from affective text. IEEE Trans. Knowl. Data Eng. **24**(9), 1658–1670 (2012)
13. Song, Y., Wang, H., Wang, Z., Li, H., Chen, W.: Short text conceptualization using a probabilistic knowledgebase. In: Proceedings of the 22nd International Joint Conference on Artificial Intelligence, pp. 2330–2336 (2011)
14. Xie, H.R., Li, Q., Mao, X.D., Li, X.D., Cai, Y., Zheng, Q.R.: Mining latent user community for tag-based and content-based search in social media. Comput. J. **57**(9), 1415–1430 (2014)
15. Xie, H.R., Li, Q., Mao, X.D., Li, X.D., Cai, Y., Rao, Y.H.: Community-aware user profile enrichment in folksonomy. Neural Netw. **58**, 11–121 (2014)
16. Li, X.D., Xie, H.R., Chen, L., Wang, J.P., Deng, X.T.: New impact on stock price return via sentiment analysis. Knowl.-Based Syst. **69**, 14–23 (2014)
17. Wang, Q.S., Wu, O., Hu, W.M., Yang, J.F., Li, W.Q.: Ranking social emotions by learning listwise preference. In: Proceedings of the 1st Asian Conference on Pattern Recognition, ACPR, pp. 164–168 (2011)
18. Lei, J.S., Rao, Y.H., Li, Q., Quan, X.J., Wenyin, L.: Towards building a social emotion detection system fro online news. Future Gener. Comput. Syst. **37**, 438–448 (2014)