Sentiment Analysis on Chinese Health Forums: A Preliminary Study of Different Language Models

Yan Zhang[1(✉)], Yong Zhang[1], Jennifer Xu[2], Chunxiao Xing[1], and Hsinchun Chen[1,3]

[1] Research Institute of Information Technology, Tsinghua National Laboratory for Information Science and Technology, Department of Computer Science and Technology, Tsinghua University, Beijing, China
zhang-yanl4@mails.tsinghua.edu.cn, hchen@eller.arizona.edu
[2] Computer Information Systems, Bentley University, Waltham, USA
jxu@bentley.edu
[3] MIS Department, University of Arizona, Tucson, USA

Abstract. Sentiment analysis on Chinese health forums is challenging because of the language, platform, and domain characteristics. Our research investigates the impact of three factors on sentiment analysis: sentiment polarity distribution, language models, and model settings. We manually labeled a large sample of Chinese health forum posts, which showed an extremely unbalanced distribution with a very small percentage of negative posts, and found that the balanced training set could produce higher accuracy than the unbalanced one. We also found that the hybrid approaches combining multiple language model based approaches for sentiment analysis performed better than individual approaches. Finally we evaluated the effects of different model settings and improved the overall accuracy using the hybrid approaches in their optimal settings. Findings from this preliminary study provide deeper insights into the problem of sentiment analysis on Chinese health forums and will inform future sentiment analysis studies.

Keywords: Sentiment analysis · Chinese health forum · Language model

1 Introduction

Chronic diseases have become more common in many countries. In China, the number and percentage of patients diagnosed with chronic diseases continue to increase rapidly in recent years. For example, a recent study estimates that there are 113.9 million Chinese people with diabetes and 493 million with pre-diabetes [17]. To better manage their health, more and more people use the Internet to seek health-related information such as the symptoms, causes, and treatments of chronic diseases. Among many types of online services, discussion forums have become increasingly popular. According to a study by the Pew Research Center [3], 80 % of American Internet users have looked for health-related information online, among whom 34 % have read about health-related

X. Zheng et al. (Eds.): ICSH 2015, LNCS 9545, pp. 68–81, 2016.
DOI: 10.1007/978-3-319-29175-8_7

personal stories from other users; and 5 % have posted health-related comments, questions, or information in online forums. Many health forums have also emerged in China to offer a platform for patients to communicate with other users.

These health forums can help alleviate the demand pressure on healthcare resources. Meanwhile, forum users may form communities, and give and receive psychological and social support, which is crucial to their recovery. Hence, analyzing the sentiment polarity (i.e., positive or negative) that users express in their posts offers a great opportunity to understand the opinions, feelings, and emotions associated with their health conditions. It may also have a practical impact on online healthcare services by helping moderators of online health communities more efficiently prioritize their responses to users [5], identify influential members, provide better social support to their members, and get innovative ideas for designing new forums.

Although many new approaches have been proposed, sentiment analysis remains a challenging task because of the language, platform, and domain characteristics. First, automatically extracting sentiment from texts is difficult as people's expressions of their feelings may be obscure, ambiguous, and hard to understand for both humans and computers. Second, analyzing sentiment in user generated contents posted on forums is more difficult than mining regular, formal texts such as news reports. Forum posts usually are short and colloquial, and may contain typos, grammatical errors, forum-specific terms and symbols, and noise such as ads and irrelevant messages. Third, sentiment analysis must consider domain specific characteristics. In the health domain, for example, many sentimental words are used for descriptions of symptoms rather than for expressing personal feelings. A negative sentiment word may not necessarily indicate a negative sentiment. For example, the sentence "When (you're) not feeling well, sweating and shaking, a portable blood glucose meter will help" is actually a neutral, objective expression although negative words are used.

Sentiment analysis on Chinese health forums is even more challenging because the performance of existing natural language processing (NLP) tools is limited and standard Chinese sentiment lexicons do not yet exist. Few studies have been done in this area. The objective of this research is to propose an effective and efficient approach for sentiment analysis on Chinese health forums.

The main contribution of our study is three-fold. First, we explored the characteristics of Chinese health forum data and examined the impact of the distribution of sentiment categories on classification performance. By manually labelling a large number of forum posts, we found that the distribution of different sentiment categories was extremely unbalanced with a very small percentage of negative posts. Our experimental results showed that such a distribution could undermine the classification performance and a balanced training set could produce a better classifier. Second, we examined a number of hybrid approaches that combine different language model based approaches for sentiment analysis on Chinese health forums. Third, we evaluated the effects of different model settings on classification performance and found the optimal settings for improving the overall accuracy of the hybrid approaches. Some of the settings in our study of Chinese health forum posts were different from what was reported in previous studies of movie reviews.

The rest of the paper is organized as follows. Section 2 reviews the related work and identifies research gaps. Section 3 presents our framework for sentiment analysis of

Chinese health forums. Section 4 reports on experimental results. Section 5 discusses the results. Section 6 concludes the paper and suggests future directions.

2 Related Work

The sentiment analysis in our study is to determine the polarity orientation, positive or negative, of a given text (i.e., a forum post). There are two main approaches for this task, namely, the *lexicon-based approach* and the *text classification approach*.

The *lexicon-based approach* uses a dictionary of sentiment words with orientations and strength (e.g., −4 for "hate" and +2 for "inspire") [15]. By aggregating scores of all words, the overall score is derived as positive (+) or negative (−). This is simple, but it cannot be applied in our study as few Chinese sentiment lexicons are available.

The *text classification approach,* a supervised learning approach, is to build classifiers from labeled instances [13], which is the focus of our study. Based on the language models used for feature learning, it can be further categorized into three types: N-gram model and its variants, structured language models, and neural net models.

N-gram Model and Variants. The N-gram model is one of the most widely used models for feature representation. It assumes that the probability of a given word is only conditional on its preceding n-1 word, where n could be 1 (the unigram model), 2 (the bigram model), 3 (the trigram model), or any whole number. This approach converts a collection of text documents into feature vectors by recording the n-gram frequency counts, and uses these vectors as input to classifiers. Wang and Manning proposed an n-gram based method called *NBSVM*, which combines Naive Bayes log-count ratio and Support Vector Machine via linear interpolation [16], and achieves outstanding performance across datasets. However, it remains an open question to determine the optimal n for sentiment analysis. Pang et al. showed that unigrams alone perform better than combining unigrams and bigrams on a movie review dataset [13]. Wang and Manning found that the inclusion of bigrams gives consistent performance gains compared with unigrams alone [16]. Mesnil et al. improved the performance on the Internet Movie Database (IMDB) dataset by adding trigrams [9].

Despite the simplicity, efficiency and accuracy of the n-gram models, their feature spaces grow linearly with the vocabulary size, which often leads to data sparsity. They also simply count the word frequency without considering word semantics.

Structured Language Model. The Structured Language Model (SLM) identifies syntactic structure of sentences by combining automatic parsing and language modeling [2]. Socher et al. developed a *Sentiment Treebank* approach to determining sentiment of sentences represented as a parser tree [14]. With rules for combining sentiment polarity of two semantic units, one can derive the polarity of words, phrases, and finally the whole sentence using a "bottom-up" method. It performs well on treebank sentences and is good at finding negations. However, its applicability is limited in the Chinese context since the performance of existing NLP tools for Chinese dependency parsing is not satisfactory, especially for long sentences.

Neural Net Language Model. The Neural Net Language model (NNLM) overcomes the curse of dimensionality in n-gram models by learning fixed-dimensional distributed representation of words [4]. The widely adopted NNLMs are RNNLM, Word2Vec, and Paragraph Vector.

- *RNNLM.* There are two architectures for NNLM, feedforward and recurrent (See Fig. 1). The feedforward neural network model (FNNM) is limited in the same way as in N-gram model where only preceding $n - 1$ words in the history are taken as context. Mikolov et al. proposed RNNLM, a recurrent neural network based language model, where the history is represented by neurons with recurrent connections and hence the context length is unlimited [11]. Plenty of empirical evidence suggests that RNNLM outperforms n-grams significantly.
- *Word2Vec.* Word2Vec is a well-known RNN based implementation of distributed word representation with two novel architectures: CBOW (Continuous Bag-of-Words Model) and Skip-gram (Continuous Skip-gram Model) (Fig. 2) [10]. CBOW sums up or averages the context on the projection layer and ignores word order as BoW model does. To improve the efficiency of the two models, binary Huffman tree based *hierarchical softmax* is used as an approximation of the full softmax. In [12], Mikolov et al. further speeded it up with the *negative sampling* approach, which subsamples the frequent words instead of building a binary tree. Word2Vec can grasp semantic relationship between words, e.g., vector("king") − vector ("man") + vector("woman") ≈ vector("queen").
- *Paragraph Vector.* Paragraph Vector (PV) is an improved version of Word2Vec and maps *sentences, paragraphs, and documents* rather than only words to continuous vector representations. Hence, this approach is more suitable for sentiment analysis than Word2Vec [6]. PV adopts similar architectures and efficiency tricks as Word2Vec, except that it adds a paragraph vector as the context during training. Le and Mikolov reported that for the PV approach, Skip-gram is better than CBOW using hierarchical softmax, and it is the new state of the art in sentiment analysis on the above mentioned two movie review datasets.

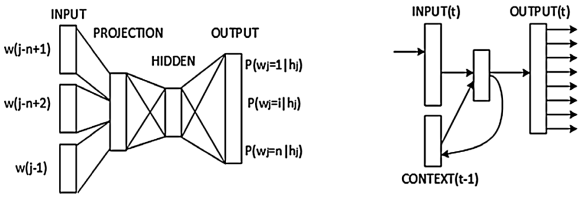

Fig. 1. Feedforward neural network (left) vs. Recurrent neural network (right) [11].

Based on our literature review we found that most prior language model based sentiment analysis studies have been done in the business and entertainment domains such as product reviews and movie reviews, little has been done in the medical and health domain. Among the state-of-the-art approaches, NBSVM, RNNLM, and PV can be applied for feature learning using limited Chinese resources. However, each

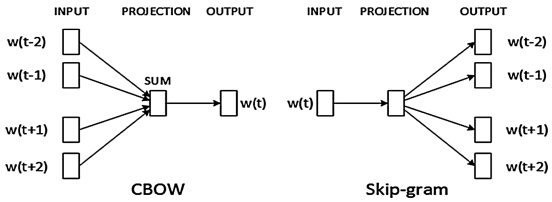

Fig. 2. Two model architectures for Word2vec [10].

approach needs to adjust several parameters for optimal performance including different history lengths in n-grams (e.g., unigrams, bigrams, trigrams, etc.), two architecture options (CBOW and Skip n-gram), as well as two efficiency tune-ups for PV. Different types of approaches can also be combined to produce better performance. For example, Mesnil and Mikolov reported that combining all three approaches yields the best performance on an English movie review dataset [9]. However, it is still not clear under which conditions and with what kind of combination these advanced techniques could achieve the best performance for sentiment analysis on Chinese health forums. Hence, we propose the following research questions:

- How can we apply the state-of-the-art language model based approaches for sentiment analysis on Chinese health forums?
- Which hybrid approach performs the best?
- How do different model settings affect sentiment classification performance?

3 Research Design

In this section, we present the process for sentiment analysis on Chinese health forums, including Data Collection and Preprocessing, Design of Data Sampling Strategy, Evaluation of Individual Approaches, and Combination of Different Models.

3.1 Data Collection and Preprocessing

Data for sentiment analysis can be collected from a health related forum platform using a web crawler. Usually each forum has a specific focus (e.g., diabetes, heart diseases, and arthritis). A text parser is then used to extract specific attributes from posts, including the forum ID, URL, post ID, post title, post time, and text content.

Data preprocessing consists of two tasks: data cleansing and word segmentation. Data cleansing includes removing duplicate posts and irrelevant information added by the platform (e.g., date of the last edit), and integrating each post into one paragraph. Word segmentation is an essential step for feature extraction in applications of mining Chinese text. Unlike English that uses space to naturally separate words, there are no delimiter characters in Chinese. In our study, we used NLPIR (http://www.nlpir.org/), a widely adopted NLP toolkit, for Chinese word segmentation.

3.2 Design of Data Sampling Strategy

Most prior sentiment analysis studies employing language model based approaches have used movie review data, where instances are evenly distributed between positive and negative sentiment classes [8, 9, 12]. However, many datasets have an unbalanced distribution among classes. A classifier trained with unbalanced data may have bias towards the majority class [1]. To address this issue, a balanced sample can be generated out of the data by designing a proper sampling strategy [7]. However, it remains unknown whether, when tested with a real, unbalanced set, the classifier trained with a balanced training set would produce better predictive accuracy than with the real, unbalanced training set, as the performance also depends on the domain, feature learning approaches, and the classification approach [1].

Hence, the sentiment distribution of the health forum data must be analyzed first by manually labelling a large sample of posts. An appropriate data sampling strategy can then be designed based on the analysis result. Specifically, if sentiment polarity in the data is evenly distributed, a simple random sampling approach will produce a balanced training set. Otherwise, an undersampling strategy [7], which randomly draws a matching number of instances out of the majority class with the minority class, can be used to create a balanced training set. Then we compare the performance of a balanced training set with the unbalanced one, and the one shows an higher accuracy on the real unbalanced test dataset is selected as the final approach.

3.3 Evaluation of Individual Feature Learning Approaches

Feature learning is the prerequisite for classification. In our study, features are learned using the state-of-the-art language model based approaches: NBSVM (n-gram), RNNLM (neural net), and PV (neural net). As mentioned above, the performance of NBSVM and PV can be affected by specific settings. Hence, we test the performance of NBSVM under three most widely adopted settings: unigrams, unigrams + bigrams, and unigrams + bigrams + trigrams. We evaluate the performance of PV under four (2X2) possible settings with two architectures (CBOW and Skip-gram), and two alternative efficiency tune-ups (hierarchical softmax and negative sampling).

3.4 Combination of Different Approaches

In statistics and machine learning studies, combining different learning methods often produces better performance than using a single approach. In this study, we examine hybrid approaches that combine the three language models (NBSVM, RNNLM, and PV), each of which is under their best settings derived from the previous individual approach evaluation step. The final output is generated by a weighted linear combination of each approach; and the weight of each model is set according to its accuracy on the validation set. The performance of all possible combinations is then evaluated.

The learned features are represented as vectors and used as input to a classifier, which assigns a label for each instance. In both the individual approach evaluation and the model combination steps, we adopt a logistic regression classifier as it shows

similar performance with linear SVM but requires less training time. The accuracy rate is calculated based on manually labeled data to assess the performance.

4 Experiments

4.1 Datasets

To evaluate our proposed hybrid approach for sentiment analysis on Chinese health forums, we collected posts from an online diabetes forum (http://bbs.tnbz.com/), which is popular among diabetes patients in China. We collected 184,708 posts in total. The date range was between September 2005 and December 2013.

We manually labeled each instance (i.e., forum post) to prepare for the training and test datasets. Our aim is to identify the negative posts to help prioritize patients in need, which is especially beneficial to the health domain. Rather than simply classifying the instances into either negative or non-negative like in many prior studies [8, 9, 12–14], we labeled it as negative, positive, or neutral. To ensure coding reliability, we first sampled 1,000 posts and hired 3 coders to label the sample. Fleiss' kappa was calculated as a measure of inter-coder agreement. The kappa values were 0.76 for the coding of three categories (negative, positive, neutral) and 0.70 for two categories (negative, non-negative), both of which indicate high inter-coder reliability.

The coding result of a much larger sample of 50,000 posts shows that the data are distributed quite unbalanced, i.e., there are only a small percentage of negative posts (see Table 1). This is interesting because we expected that there were many negative posts in health related forums, where people share their feelings (e.g., worries, frustration, and anxiety) about their health conditions. Such an unbalanced distribution makes it extremely difficult to find enough negative posts for training the classifier and may significantly lower the performance.

Table 1. Sentiment polarity distribution in the health forum dataset

		Non-negative	
	Negative	Positive	Neutral
# posts	2,806 (5.6 %)	998 (2 %)	46,196 (92.4 %)

As our data was distributed unbalanced, the undersampling strategy was selected over the simple random sampling strategy. We evaluated the performance of each language model based approach using both balanced and unbalanced training sets.

- Balanced training set: 2,520 (90 % of 2,806) negative and non-negative instances;
- Balanced testing set: 280 (10 % of 2,806) negative and non-negative instances;
- Unbalanced training set: 2,520 negative and 42,340 non-negative instance.

For the unsupervised PV approach, we used all 50,000 instances without class labels. For approaches with several possible settings, we randomly assigned a specific one (i.e., *skip n-gram* architecture with *negative sampling* efficiency tune-up for PV

approach, and using *unigrams, bigrams, and trigrams* as history length for NBSVM). Table 2 reports the accuracy of different models trained using balanced and unbalanced sets and tested using a balanced set. Tests using unbalanced sets produced similar results, which suggests that a balanced training set can deliver better performance and generalizes well to the real unbalanced test set.

Table 2. Accuracy with different training sets

	RNNLM	PV	NBSVM
Balanced training set	75.02 %	82.46 %	82.71 %
Unbalanced training set	53.46 %	58.27 %	56.24 %

4.2 Evaluation

Performance of Individual Approaches. We evaluated the performance (accuracy) of individual approaches under all possible settings as shown in Table 3. Statistical analyses were performed using a one-tailed paired sample t test. Among the three approaches, NBSVM shows the highest accuracy, followed by PV, and then RNNLM.

We also explore the effects of different settings on these approaches. For the two architectures of PV, Skip n-gram significantly outperforms CBOW, which is consistent with previous findings using movie review data [6]. As for the two speed tune-ups, the accuracy of negative sampling (NS) is significantly higher than that of hierarchical softmax (HS). Moreover, NS requires much less training time (30 min) than HS (90 min) in our experiments using one CPU only. Thus NS + Skip n-gram is the best setting for PV in terms of both effectiveness and efficiency.

Table 3 also reflects the impact of different lengths of n-grams on NBSVM, where unigrams (U) and unigrams + bigrams (UB) outperform unigrams + bigrams + trigrams (UBT) significantly, and there is no significant difference between U and UB while UB produces a slightly higher accuracy rate. In terms of efficiency, U takes slightly less training time than UB as it has smaller size of feature space. Hence, both U and UB could be the best settings for NBSVM.

Table 3. Accuracy of individual approaches with ten-fold cross validation

Approach			Accuracy	p-value
RNNLM			75.02 %	–
PV	Hierarchical Softmax (HS)	CBOW	78.34 %	0.008
		Skip n-gram	**79.98 %****	
	Negative Sampling (NS)	CBOW	81.32 %	0.003
		Skip n-gram	**82.46 %****	
NBSVM	Unigrams (U)		83.64 %	–
	Unigrams + Bigrams (UB)		**83.73 %**	–
	Unigrams + Bigrams + Trigrams (UBT)		82.71 %	–

$*p < 0.05; **p < 0.01$.

Performance of Hybrid Approaches. We evaluated the performance of the hybrid models to find whether combining different models produces higher accuracy and to assess the contribution of each individual approach (see Table 4).

We first evaluated the performance of the combination of two models. Among such combinations, PV + NBSVM produced the highest accuracy, followed by RNNLM + NBSVM, and RNNLM + PV, indicating that NBSVM contributes most to the overall performance. Moreover, the significant difference between the performance of PV + NBSVM and RNNLM + NBSVM suggests the importance of PV.

The performance of three-model combination was worse than PV + NBSVM, indicating that the RNNLM reduces the overall accuracy. Still, no significant difference was found between the accuracy under unigram and unigram + bigram settings of NBSVM, while the latter shows a slightly higher average accuracy.

Overall, the PV + NBSVM (UB) approach performed the best with a 2.47 % increase in accuracy from NBSVM (UB), the best state-of-the-art individual approach.

Table 4. Accuracy of hybrid approaches using ten-fold cross validation

Model combinations	Accuracy	p-value
RNNLM + PV	82.50 %	–
RNNLM + NBSVM (Unigrams)	83.68 %	0.1277
RNNLM + NBSVM (Unigrams + Bigrams)	84.09 %	
PV + NBSVM (Unigrams)	85.59 %	0.0738
PV + NBSVM (Unigrams + Bigrams)	**86.20 %**	
RNNLM + PV + NBSVM (Unigrams)	85.30 %	0.2216
RNNLM + PV + NBSVM (Unigrams + Bigrams)	82.50 %	

5 Discussion

Two findings distinguish our sentiment analysis on Chinese health forum data from prior studies of movie reviews in English:

First, the effect of the length of n-grams is found to be different from previously reported. Our experiments on the NBSVM approach show that both unigrams (U) and unigrams + bigrams (UB) are better than unigrams + bigrams + trigrams (UBT). However, a previous study using the IMDB dataset shows that UBT is the best among the three, followed by UB, and then U [9].

Second, we found that the effects of the two efficiency tune-ups used in the PV approach are also different. Our experiments show that when using the PV approach for sentiment analysis on Chinese health forum data, negative sampling (NS) produces higher accuracy than hierarchical softmax (HS) does. Yet in a previous study, HS is selected as a better choice for learning paragraph vector [6].

To find out the causes for the accuracy differences, we compared two types of errors under different settings, i.e., the ratio of positive posts misclassified as negative

(i.e., false negative) and the ratio of negative posts misclassified as positive (i.e., false positive). Table 5 shows that the differences are mainly caused by a higher false positive error in UBT, and a higher false negative error in HS.

We further analyzed the impact of post length on accuracy. In our training set, the average lengths of positive and negative posts are 579 and 364, respectively. We first examined the posts misclassified by UBT/HS but correctly classified by UB/NS. As shown in Table 6, posts misclassified by UBT generally are shorter than those in the training set, indicating that UBT may not work well for short posts. This is because a Chinese medical term usually consists of multiple characters, each of which is a word by itself, and is longer than its English counterpart. For example, "糖尿病," the Chinese term corresponding to the single-word term "diabetes" in English, has three characters. As a result, a short post in Chinese may not be long enough to generate as many trigram features as in English.

We then analyzed posts misclassified by HS but correctly classified by NS, both of which use the Skip n-gram architecture. Because the skip n-gram architecture is essentially an n-gram model with a history length of at most 5 in our experiment, which is a stricter matching criterion than trigrams, these misclassified posts should be shorter than the average posts in the training set. However, as shown in Table 6, the false negative posts in HS, on average, are longer than those in the training set. Looking into these posts, we found that the longer average length is caused by a small percentage of rather long posts (i.e., 20 % posts longer than 1100), indicating that HS does not work well for long posts.

Table 5. Types of errors in different settings

Approach	Settings	False positive	False negative
NBSVM	Unigrams + Bigrams (UB)	18.93 %	12.14 %
	Unigrams + Bigrams + Trigrams (UBT)	**22.86 %**	10.71 %
PV	NS	15.35 %	19.64 %
	HS	16.43 %	**23.57 %**

Table 6. Average length of posts misclassified in UBT/HS only

	False positive	False negative
UBT	311	67
HS	319	567

Examples of posts misclassified by UBT/HS but correctly classified by UB/NS are shown in Table 7. Note that we use "/" to show word segmentation returned by the NLPIR tool in the original Chinese posts, and use "()" to indicate the originally omitted words in the translated version.

Table 7. Examples of posts misclassified by UBT/HS but correctly classified by UB/NS

App	Err	Original post	Translation
UBT	False Pos.	糖/妈妈/们/请/帮帮/我/最/好/是/生/过/的/。/。/我/是/个/糖/妈妈/现在/怀孕/8/个/月/了/宝宝/马上/要/降生/了/在/怀孕/期间/发生/过/6–7/次/低血糖/低血糖/发生/的/时候/大多/都/是/3点/多/请问/对/孩子/有/影响/吗/?/?/我/好/怕/我/低血糖/宝宝/生/下来/会/脑瘫/啊/有/过来/的/妈妈/告诉/下/会/发生/这样/的/情况/吗/?/?/?/?	Could any diabetic mothers help me? It would be better if you already have a baby…I was diagnosed with gestational diabetes in my 8th month of pregnancy and would give birth soon. Hypoglycemia has occurred 6-7 times during my pregnancy and it happens mostly at 3:00 PM. Would it have any impact on the baby?? I'm afraid that my baby will be born with cerebral palsy since I have a low blood glucose level (.) Could any experienced mom tell me whether it will happen????
	False Neg.	'肾功能不全/:/上/星期六/和/老伴/同/测/了/个/尿/四/样/,/今天/见/报告单/上/写/着/"/肾功能不全/"/,/但/指标/都/正常/,/虚惊/一/场	Renal insufficiency: My wife and I did a urine test last Saturday. The report showed that it was renal insufficiency. However, all indicators were normal. It turned out to be a false alarm.
HS	False Pos.	俞老/,/请/教/:/我/是/11月/4日/看/了/DM/的/宣传/海报/上/说/的/症状/和/偶/一样/,/口渴/,/喝/大量/的/水/,/消瘦/,/马上/到/了/医院/检查/:/先是/查/了/尿/尿糖/:/3 +/胆红素/:/阴性/酮/体/:/微量/蛋白质/:/1 +/亚硝酸盐/:/阴性/血红蛋白/:/微量/白细胞/:/阴性/PH/:/5.5/尿/比重/:/1.030/红细胞计数/:/0/白细胞计数/:/0……/后面/的/就/不/写/了/,/看/完/尿/,/医生/让/偶/做/了/生化/血/葡萄糖/:/14.91/总胆固醇/:/4.99/甘油三酯/:/1.84/高密度脂蛋白胆固醇/:/1.19/低密度脂蛋白胆固醇/:/3.40C/肽/:/1.724/胰岛素/:/67.20/医生/说/我/是/糖尿病/。/俞老/,/有/几/个/问题/请教/,/感谢/!/1/,/我/不/要/做/葡萄糖/耐量/测试/就/确定/是/DM/了/吗/?/2/,/我/尿/里/有/蛋白/,/为什么/医生/不/让/偶/做/微量/检查/?/是/不/是/我/	Dr. Yu, I have some questions to ask you. I saw the DM poster on November 4. I had the symptoms mentioned in the poster: thirsty, drinking plenty of water, and weight loss. I went to the hospital immediately, and did a urine test: glycosuria: 3+ (,) bilirubin: negative(,) ketones: trace(,) protein: 1 + (,) nitrite: negative(,) hemoglobin: trace(,) WBC: trace(,) urine specific gravity: 1.030(,) RBC count: 0(,) WBC count: 0…… The rest is omitted here. After the urine test, the doctor asked me to do a biochemical test: blood glucose: 14.91(,) total cholesterol: 4.99(,) triglycerides: 1.84(,) high-density lipoprotein cholesterol: 1.19(,) low-density lipoprotein cholesterol: 3.40C(,) peptide: 1.724(,) insulin: 67.20(.)

(*Continued*)

Table 7. (*Continued*)

App	Err	Original post	Translation
		肾/有/病/了/,/这个/是/我/最/担心/的/,/家族/里/有/人/得/过/,/怕/怕/。	The doctor said that I have diabetes. Thus, Dr. Yu, I have some questions to ask you and thank you in advance: 1. Is it certain that I have DM without doing a glucose tolerance test? 2. Why did the doctor not ask me to do a micro inspection since I have protein in my urine? Is there something wrong with my kidney? This is my greatest worry since I have a family history of kidney diseases. It is scary.
	False Neg.	建议/版/主/详细/介绍/一下/活力/试纸/行货/、/水货/和/假货/的/识//学会/识别/行/、/水/、/假货/试纸/的/本领/,/对/想/省/银子/的/DMer/十分/重要/。	(I) suggest that moderators explain in detail about the recognition of properly licensed, parallel and counterfeit active test strips. Learning to recognize properly licensed, parallel and counterfeit test strips is important to DMers who want to save money.

6 Conclusion and Future Directions

This paper presents our research of sentiment analysis on online Chinese forums that are related to health topics. Our research generates three major findings. First, based on the manual labeling process on a large number of posts we found that the distribution of sentiment categories (positive, negative) in the health-related forum posts is extremely unbalanced. Using different data sampling strategies, we found that the sentiment category distribution can dramatically affect the classification performance and a balanced training set could produce higher accuracy than the unbalanced one. Second, we found that hybrid approaches combining different language models outperform individual approaches for sentiment analysis on Chinese health forums. Finally we evaluated the effects of different model settings for each approach and applied the optimal settings to the hybrid approaches to improve the overall accuracy. Some of the settings in our study of Chinese health forum posts were different from what was reported in previous studies of movie reviews. In the future, we will extend our work by incorporating prior knowledge into these models to further improve the performance of sentiment analysis.

Acknowledgments. This work was supported by the National High-tech R&D Program of China (Grant No. SS2015AA020102), National Basic Research Program of China (Grant No. 2011CB302302), the 1000-Talent program, and the Tsinghua University Initiative Scientific Research Program. We appreciate the research assistance provided by Qingbo Cao, Yanshen Yin, and Xinhuan Chen at Tsinghua University.

References

1. Chawla, N.V.: Data mining for imbalanced datasets: an overview. In: Maimon, O., Rokach, L. (eds.) Data Mining and Knowledge Discovery Handbook, pp. 853–867. Springer, New York (2005)
2. Chelba, C., Jelinek, F.: Recognition performance of a structured language model. arXiv:cs/0001022 (2000)
3. Fox, S.: The social life of health information 2011. Pew Internet & American Life Project Washington, DC (2011)
4. Hinton, G.E.: Learning distributed representations of concepts. In: Proceedings of the Eighth Annual Conference of the Cognitive Science Society, vol. 1, p. 12, Amherst, MA (1986)
5. Huh, J., Yetisgen-Yildiz, M., Pratt, W.: Text classification for assisting moderators in online health communities. J. Biomed. Inform. **46**(6), 998–1005 (2013)
6. Le, Q.V., Mikolov, T.: Distributed representations of sentences and documents. In: Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21–26 June 2014. JMLR Proceedings, vol. 32, JMLR.org (2014)
7. Lee, C.Y., Lee, Z.J.: A novel algorithm applied to classify unbalanced data. Appl. Soft Comput. **12**(8), 2481–2485 (2012)
8. Maas, A.L., Daly, R.E., Pham, P.T., Huang, D., Ng, A.Y., Potts, C.: Learning word vectors for sentiment analysis. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, vol. 1, pp. 142–150. Association for Computational Linguistics (2011)
9. Mesnil, G., Ranzato, M., Mikolov, T., Bengio, Y.: Ensemble of generative and discriminative techniques for sentiment analysis of movie reviews. CoRR bs/1412.5335 (2014)
10. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. CoRR abs/1301.3781 (2013)
11. Mikolov, T., Karafiát, M., Burget, L., Cernocký, J., Khudanpur, S.: Recurrent neural network based language model. In: INTERSPEECH 2010, 11th Annual Conference of the International Speech Communication Association, Makuhari, Chiba, Japan, pp. 1045–1048, 26–30 September 2010
12. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in Neural Information Processing Systems, pp. 3111–3119 (2013)
13. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up?: sentiment classification using machine learning techniques. In: Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing, vol. 10, pp. 79–86. Association for Computational Linguistics (2002)
14. Socher, R., Perelygin, A., Wu, J.Y., Chuang, J., Manning, C.D., Ng, A.Y., Potts, C.: Recursive deep models for semantic compositionality over a sentiment treebank. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), vol. 1631, p. 1642. Citeseer (2013)

15. Taboada, M., Brooke, J., Tofiloski, M., Voll, K., Stede, M.: Lexicon-based methods for sentiment analysis. Comput. Linguist. **37**(2), 267–307 (2011)
16. Wang, S., Manning, C.D.: Baselines and bigrams: simple, good sentiment and topic classification. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers, vol. 2. pp. 90–94. Association for Computational Linguistics (2012)
17. Xu, Y., Wang, L., He, J., Bi, Y., Li, M., Wang, T., Wang, L., Jiang, Y., Dai, M., Lu, J., et al.: Prevalence and control of diabetes in Chinese adults. JAMA **310**(9), 948–959 (2013)