A Depression Detection Model Based on Sentiment Analysis in Micro-blog Social Network

Xinyu Wang[1], Chunhong Zhang[1], Yang Ji[1], Li Sun[1], Leijia Wu[2], and Zhana Bao[3]

[1] Beijing University of Posts and Telecommunications (BUPT), Beijing, China
{wxinyu906,zhangch.bupt.001,ji.yang.0001,buptsunli}@gmail.com
[2] University of Technology, Sydney, Australia
leijia.wu@alumni.uts.edu.au
[3] Graduate School of Global Information and Telecommunication Studies, Waseda University, Japan
znabao@ruri.waseda.jp

Abstract. Datasets originating from social networks are valuable to many fields such as sociology and psychology. But the supports from technical perspective are far from enough, and specific approaches are urgently in need. This paper applies data mining to psychology area for detecting depressed users in social network services. Firstly, a sentiment analysis method is proposed utilizing vocabulary and man-made rules to calculate the depression inclination of each micro-blog. Secondly, a depression detection model is constructed based on the proposed method and 10 features of depressed users derived from psychological research. Then 180 users and 3 kinds of classifiers are used to verify the model, whose precisions are all around 80%. Also, the significance of each feature is analyzed. Lastly, an application is developed within the proposed model for mental health monitoring online. This study is supported by some psychologists, and facilitates them in data-centric aspect in turn.

Keywords: data mining, Chinese sentiment analysis, depression.

1 Introduction

The rise of online social network provides unprecedented opportunities for solving problems in a wide variety of fields with information techniques [1]. For example, traditional psychology research is based on questionnaires and academic interviews, but many psychologists are now turning their sights to web media. They try to analyze the data of social networks from the view of psychology. Undoubtedly, this discipline integration injects vigor into psychology, however, the supports from technical perspective are far from enough. Only some simple statistic tools are applied in such kind of research, and little attention is paid to design specific data mining methods, especially in Chinese text processing.

J. Li et al. (Eds.): PAKDD 2013 Workshops, LNAI 7867, pp. 201–213, 2013.

This paper applies data mining techniques to psychology, specifically the field of depression, to detect depressed users in social network services (SNS). The expansion of data mining to psychology is of great technical and social significance. It is proved that the proposed model in this paper could effectively help for detecting depressed ones and preventing suicide in online social networks.

"I suffer from depression, thus trial for death. There's no special reason. Please do not care about my departure. Bye, everyone." On March 18^{th}, 2012, a micro-blog posted by Zoufan dropped a bomb in Sina Micro-blog. Then on March 19^{th}, JiangNing Police confirmed that Zoufan had committed suicide. Zoufan, whose real name is MaJie, was a talented university student in Nanjing. Her suicide for depression left us with endless sorrow. Zoufan Tragedy resulted from depression has aroused extensive concern of the whole society in China. The micro-blog contents of Zoufan are like death signals, revealing total despair and depression characteristics. But unfortunately, they did not attract attentions in time, until she resolutely said goodbye to the world with her last micro-blog.

This research is accomplished with the help of psychologists. They detect some depressed users within psychological diagnostic criteria, and observe the online behaviors of them. This paper constructs a depression detection model based on the features of depressed users derived from psychological observations. The result is verified to be efficient, so this model could be used for large-scale mental health monitoring, and avoid the tragedy of Zoufan occurring again.

Definitely, there are some challenges in our study. The complexity of Chinese text processing is one of them. The ambiguity definition relies on almost all the levels of syntactical unit in Chinese linguistics, which brings difficulties for word segmentation and linguistic rules construction [2]. Coupled with the particularity of the micro-blog content of depressed users, it takes great effort to adjust our model to improve its performance. Another challenge is how to distinguish training dataset and select features as criteria in the model. Due to the particularity of depression, this work is not easy for data mining experts, so the help of psychologists is asked as instructions.

To sum up, the contributions of this paper are: (1) From the aspect of methodology, data mining techniques is expanded to depression area. (2) Sentiment analysis algorithm, specifically for Chinese micro-blog is proposed for calculating the depression inclination. (3) An association model is established between features abstracted from Micro-blog system and depression inclination. The model also determines the principle features which affecting depression detection significantly. (4) An application in Sina Micro-blog is developed for monitoring the users' mental health in SNS. The basic idea of this paper could be explicitly extended to other language scenarios.

2 Related Work

2.1 Research of Depression in Psychology

Depression is the world's fourth largest disease and will be in the second place in 2020 according to World Health Organization statistics [3]. The main clinical symptom of depressed patients is lasting depressed state of mood and lack of positive emotions. They prefer to be alone rather than together with others. What's more, most of depressed patients suffer from chronic insomnia.

The research of depression in social network in psychology comes in two types: one is to discover disciplines of a crowd of depressed users [4-5]; the other is to look into a specific case elaborately [6]. Literature [4] observes linguistic markers of depression through collecting posts by depressed and non-depressed individuals from Internet forum. It analyzes the text with LIWC, a computerized word counting tool, and shows that the online depressed writers use more first person singular pronouns but less first person plural pronouns, more negative emotion words but less positive emotion words. Literature [6] discusses the relations between SNS behaviors and depression levels based on Zoufan Event. It is established by questionnaire and statistic tools, and reveals that frequencies of the original posts could indicate micro-bloggers' depressed levels. Also, the period of time users post micro-blogs is a consideration as most depression patients suffer from chronic insomnia.

These researches of depression features in the perspective of psychology provide reliable background knowledge for our study. However, when comes to data analysis problems, only some simple statistic tools are designed for them, which undoubtedly limit their researches. Therefore a specific data mining technique to detect depressed users is designed in this study based on their results.

2.2 Sentiment Analysis Techniques

As the cardinal symptom of depression is severe negative emotions and lack of positive emotions, sentiment analysis is the most important step in depression detection. Sentiment analysis aims at mining users' opinions and sentiment polarity from the texts they posted [7]. Recently many progresses have been made in sentiment analysis on Twitter data. These researches include two aspects:

- Subject-independent analysis, namely judging the polarity of the tweets without considering if it is relevant to a subject [8-10]. The main approaches are based on hashtags, smileys and some abstract features.
- Subject-dependent analysis, namely judging the polarity of the tweets based on the given subject [11-12]. The sentiments of the tweets as positive, negative or neutral in [11], according to not only the abstract features but also the target-dependent features, which refers to the comments on the target itself and the related things, which are defined as extended targets.

Sentiment analysis research on Chinese text is still in its starting stage [13]. The highest accuracy of polarity discrimination on Chinese text is only 59.27% in the latest NTCIR evaluation [14]. Little study has been made for solving problems in a specific field, although analysis strategy differs a lot for different fields. For example, depression sufferers tend to think the topic about "death", so this kind of words should be paid special attention to when constructing the vocabulary. Micro-blogs are often written in a colloquial style, which also bring new challenges when instituting the linguistic rules in the proposed method.

The problem addressed in this paper is subject-dependent sentiment analysis of micro-blogs. Inspired by the work in literature [11], abstract features and target-dependent features are taken into account. This study stresses the particularity of depression and micro-blog content, and the whole model is specifically designed based on them. As shown in Fig.1, a sentiment analysis method is firstly proposed utilizing vocabulary and man-made rules to calculate the depressive inclination of each micro-blog in Fig.1 (A). The vocabulary and man-made rules in sentiment analysis method are constructed based on the Chinese syntax rules, the particularity of depression and micro-blogs (section 3). Then as shown in Fig.1 (B), a depression detection model is constructed based on the proposed method and 10 features of depressed users derived from psychological research (section 4). Lastly, the significance of each feature is analyzed and a simplified model is proposed for the application in Sina Micro-blog.

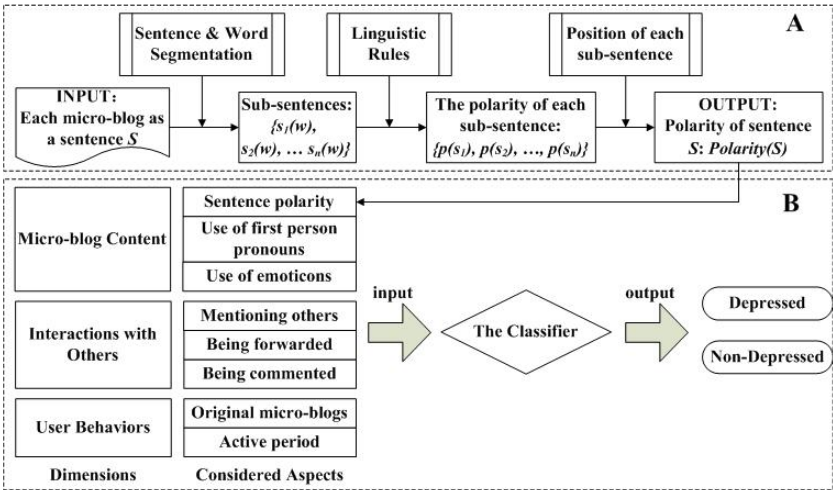
A
Sentence & Word Segmentation
Linguistic Rules
Position of each sub-sentence
INPUT: Each micro-blog as a sentence S
Sub-sentences: $\{s_1(w), s_2(w), \ldots s_n(w)\}$
The polarity of each sub-sentence: $\{p(s_1), p(s_2), \ldots, p(s_n)\}$
OUTPUT: Polarity of sentence S: $Polarity(S)$
B
Micro-blog Content
Sentence polarity
Use of first person pronouns
Use of emoticons
Interactions with Others
Mentioning others
Being forwarded
Being commented
User Behaviors
Original micro-blogs
Active period
Dimensions
Considered Aspects
input
The Classifier
output
Depressed
Non-Depressed

Fig. 1. Framework of the proposed model in this paper

3 Sentiment Analysis of Micro-blog Content

The most direct expression of depressed mood is the users' micro-blog content, so the sentiment analysis method in this section helps to figure out the polarity of each piece of micro-blog, which emphasizes the depression inclination reflected from the content. A vocabulary is constructed based on HowNet [15], and the sentence structure patterns and calculation rules are derived according to Chinese syntax rules. As described above, the particularity of depression and micro-blogs are paid special attention to the whole process.

3.1 Vocabulary Construction

The most essential particularity of depression and micro-blogs is the use of words. A vocabulary fitting for depression detection is constructed based on How-Net, a comprehensive vocabulary of Chinese words, as shown in Table 1.

Table 1. Words in HowNet vocabulary [15]

2 Item		Num.	Example
Emotion Words	Positive	4566	pretty, love, like, happy, good
	Negative	4370	ugly, sad, depressed, unhappy, bad
Degree Modifiers		219	most(2), over(1.75), very(1.5), more(1), -ish(0.75),insufficient(0.5)

HowNet contains most of the popular emotion words and degree modifiers. The weights of degree modifiers are quantified into six levels according to their intensities. HowNet is designed for general sentiment analysis. In order to make it fit for depressed inclination calculation, several adjustments are made as follows:

1. Emotion words, cyberspeaks, modal particles and negative words are added:
   1) Depressed users tend to use more emotion words, especially negative emotion words [4], some of which are even only for them. For example, “bye” is a neutral word for normal people, but it’s a typical negative one for depressed users. So these typical emotion words for depression are added.
   2) Considering that cyberspeaks are in prevalent in Internet, they are playing an important role in micro-blogs. Therefore these words are also added, for example, “smilence”, which means “smile silently”, into the vocabulary.
   3) As micro-blogs are often written in a colloquial style, modal particles often occur in micro-blogs to express feelings directly, such as “ha-ha” and “a-ha”, so these modal particles are added into the vocabulary too.
   4) Besides degree modifiers, negative words could also modify the expressions, which do not exist in HowNet. Negative words totally reverse the meaning of the expression, such as “not”, “never”. The negative words selected from a dictionary are imported as a new item into the vocabulary.
2. The part of speech of each word are recognized:
   The proposed calculation rules is derived from Chinese syntax rules, which are defined by parts of speech. So the part of speech of each word are recognized and also imported as an attribute into the vocabulary.

Finally, the vocabulary is constructed with three items as shown in Table 2. 1210 emotion words and 36 negative words are added. Each word holds its own part of speech and weight. Degree modifiers are inherited from HowNet.

3.2 Linguistic Rules Construction

The meaning of a sentence could not be decided only by the words it uses, but also by the order of words, named the structure of the sentence. For example, “我

Table 2. Words in the proposed vocabulary

3 Items		Num.	Part of speech	Weight	Notation
Emotion Words	Positive	5127	verb, adjective, adverb, modal particle	+1	W_EW
	Negative	5019		-1	
Degree Modifiers		219	adjective, adverb	Inherited from HowNet	W_DM
Negative Words		36	adverb	-1	W_NW

不很高兴(slightly unhappy)", " 我很不高兴(very unhappy)", the two sentences share the same 5 characters, but have obvious different extent of how happy it is, so the structure of sentence should be taken into account in the process of polarity calculation. The structure of sentences could be described as the linguistic rules, which reflects the complexity of Chinese language in one aspect. In this section, linguistic rules based on the proposed vocabulary is constructed by taking the colloquial style of micro-blog into account.

According to Chinese syntax rules, the proposed linguistic rules are derived as shown in Table 3. In the rules, Sentence structure patterns are recognized with different items of words in the vocabulary, and each pattern has its own calculation rule based on the weight of each word. If the sentence is recognized as "Partial Negative Structure", a coefficient should be multiplied for precise result, which is set as -0.5 in Table 3. How to calculate the polarity of a given sentence according to the rules will be introduced in the next section.

Table 3. Linguisitc rules based on the proposed vocabulary

Chinese syntax rules	Linguistic rules in our method	
	Structure pattern	Calculation rules of each pattern
Single Word Structure: v, adj, modal particles	W_EW	玩/play = +1 = 1 哈哈/ha-ha = +1 = 1
Adverbial-Modifier Structure	W_DM+W_EM	很高兴/very happy = 1.5×(+1) = 1.5
Verb-Complement Structure	W_EM+W_DM	好得很/awfully well = (+1)×1.5 = 1.5
Negative Structure	W_NW+W_EW	不高兴/not happy = (-1)×1 = -1
Complete Negative Structure	W_DM+W_NW +W_EW	很不高兴/very unhappy = 1.5×(-1)×1 = -1.5
Partial Negative Structure	W_NW+W_DM +W_EW	不很高兴/slightly unhappy = (-0.5)×[(-1)×1.5×1] = 0.75 *: -0.5 is the coefficient in our rule.
Coordinate Structure	W_EW+W_EW	幸福开心/glad and happy = (+1)+(+1) = 2

3.3 Procedure of the Proposed Method

Within the preparation of vocabulary and linguistic rules construction, the proposed method contains 3 main steps as shown in Fig.1 (A).

Sentence Segmentation and Word Segmentation. A piece of micro-blog allows 140 characters at most, so it may contain several sub-sentences. Punctuations are taken as symbols to segment sentences. ICTCLAS Chinese word segmentation systems [16], the most popular one throughout the world, is applied for segmenting word and labeling part of speech of each word.

- Each micro-blog is regarded as a sentence S, and each S is a sequence of N sub-sentences denoted by S= $\{s_1$,s_2,...$s_n\}$, where s_n is the n_{th} sub-sentence.
- A sub-sentence s_i is a collection of M words denoted by s_i=$\{w_1(\mathrm{s}p_1)$,$w_2(\mathrm{s}p_2)$,...,$w_M(\mathrm{s}p_M)$ $\}$, where $w_m(sp_m)$ is the m_{th} word in the collection, and sp_m refers to its part of speech.

Polarity Calculation of Each Sub-sentence. After being segmented, the polarity of each sub-sentence s_i could be calculated by structure pattern mining and the corresponding calculation rules. The process is implemented as follows:

Algorithm 1. Polarity calculation algorithm

```
Sub-sentence $s_i$: I am extremely happy and very glad today.
Word segmentation: I(noun), today(noun), extremely(adverb), happy(adjective), and(adverb), very(adverb), glad(adjective).
Keyword extraction in vocabulary: extremely(adverb)#W_DM, happy(adjective)#W_EW, very(adverb)#W_DM, glad(adjective)#W_EW
Structure pattern mining:
    W_EW+W_EW: happy(adjective)#W_EW+ glad(adjective)#W_EW;
    W_DM+W_EM: extremely(adverb)#W_DM+ happy(adjective)#W_EW;
    W_DM+W_EM: very(adverb)#W_DM + glad(adjective)#W_EW.
Polarity calculation of sub-sentence $s_i$: p($s_i$)= [weight(extremely)×weight(happy)] +[weight(very)×weight(glad)]= [2×(+1)]+[1.5×(+1)]= +3.5.
```

Polarity Calculation of Sentence S. The polarity of sentence S is determined by the polarities and positions of its sub-sentences, as the position of a sub-sentence s_i in S can indicate its importance [17]. This is especially noticeable in micro-blog, because it enables people to record their immediate feelings at any time. If a micro-blog content is long, the beginning and ending sub-sentences often reflect the writer's feelings more directly, thus more important than those in the middle. Therefore, the higher weights are assigned to sub-sentences at the two ends of the micro-blog as (1). With the polarity and position of each sub-sentence, the polarity of S is calculated as (2).

$$\lambda(s_i) = \frac{1}{\min(i, N-i+1)}, 1 \leq i \leq N \tag{1}$$

$$Polarity(S) = \sum\nolimits_{i=1}^{N} [\lambda(s_i) \times p(s_i)] \tag{2}$$

N is the number of sub-sentences in S, and i is the position of s_i in S. A positive *polarity(S)* means the sentence expresses a positive sentiment of the user, and a negative one means opposite. If the polarity equals to zero, then it is objective. The absolute value $|polarity(S)|$ shows the intensity of the sentiment.

Research in psychology shows that depressed individuals focus more on negative aspects of their lives [4]. So the polarity of users' micro-blog contents is an important feature in depression detection in section 4, which is normalized as (3), where $|S|$ is the total number of micro-blogs during a given period of time.

$$NormalizedSentencePolarity = \sum_{|S|} {}^{Polarity(S)}\!/_{|S|} \tag{3}$$

4 Depression Detection Model

When it comes to depression detection, many other features need to be considered. In this section, the work of psychologists is firstly introduced, and then the classifier based on their work is designed for depression detection.

4.1 Psychologists' Work

Psychologists observe the online behaviors of depressed users, and discover potential features that could be used to distinguish depressed and non-depressed individuals. All these features mainly come from three dimensions: micro-blog content, interactions and behaviors [4-6]. Table 4 lists the statistical data of ten features of two depressed and two normal samples in two weeks, in which A and B are the anonymized users. It reveals that most features show significant differences. For example, depressed users tend to use more first person singular pronouns but less emoticons. However, some features show little influence on these four users, such as times of being forwarded and commented. The proposed model obtained by training data will further illustrate the significance of each feature for depression detection in section 5.

4.2 Model Construction

Taking the achievement of psychologists as background knowledge, their observations need to be converted to parameters that are easily imported into the model as shown in Fig.1 (B). As the calculation of sentence polarity has been discussed in section 3, how to obtain, process and normalize other features will be introduced in this part.

The Use of First Person Singular and Plural Pronouns. As the result in [4] shows, depressed users tend to focus on themselves and detach from others. They use more first person singular pronouns ("I") but less first person plural pronouns ("We"). So the use of first person pronouns is considered in two aspects:

- The quantity of first person pronouns, reflecting their focuses on themselves.
- The ratio of first person singular pronouns to first person plural pronouns as (4), where Q_{fs} and Q_{fp} represent the quantities of first person singular and plural pronouns respectively.

Table 4. Features of typical depressed and non-depressed samples

		Depressed samples		Non-depressed samples	
		Zoufan	A	Li Kaifu	B
Number of micro-blogs($\|S\|$)		215	124	1156	621
Micro-blog content	1^{st} person singular	264 (123%)	142 (115%)	404 (34.9%)	97 (15.6%)
	1^{st} person plural	3 (1.4%)	7 (5.6%)	94 (8.1%)	8 (1.3%)
	Positive emoticons	8 (3.7%)	23 (18.5%)	234 (20.2%)	76 (12.2%)
	Negative emoticons	9 (4.2%)	19 (15.3%)	99 (8.6%)	24 (3.9%)
Interactions	Mentioning	22 (10.2%)	10 (8.1%)	574 (49.7%)	621 (100%)
	Being Forwarded	161592 (752%)	7 (5.60%)	3229210 (2793%)	35 (5.6%)
	Being Commented	210902 (981%)	50 (40.30%)	1458263 (1262%)	129 (20.80%)
Behaviors	Original blogs	183 (85.1%)	91 (73.4%)	758 (65.6%)	75 (12.1%)
	Blogs posted between $0:00-6:00$ o'clock	57 (26.5%)	67 (54.0%)	30 (2.6%)	12 (1.9%)

$$Ratio\_StoP = \frac{\sum_{|S|} Q_{fs}}{\sum_{|S|} Q_{fp} + 1} \tag{4}$$

For this purpose, all the first person singular and plural pronouns in users' micro-blogs need to be detected. It requires that the first person pronouns must be the subject of the sentence. So besides detecting all the first person words, whether they are the subject of the sentence or not should also be checked. To solve this problem, we choose to check if the word following them could be used as predicate. If they could, this sentence is considered as the first person pronoun. Obviously, this method would meet problems when the sentence structure is too complex, for example, some adverbials are following the subject. But as discussed above, micro-blogs are often written in a colloquial style, and complex structures are not frequent, so the proposed method is effective after being tested.

The Use of Emoticons. As shown in Table 4, depressed individuals use less emoticons, and the result in [4] shows they focus more on negative aspects of their lives. So similar with the consideration of first person pronouns, the use of emoticons is also considered in two aspects: the absolute quantity of emoticons, and the relative ratio of positive emoticons to negative ones. For this purpose, all the emoticons should be detected and distinguished as positive or negative ones. So firstly the common 76 emoticons are divided into 3 categories: 34 positive ones, 32 negative ones and 10 neutral ones. Then all the emoticons in users' micro-blogs detected are matched with the 3 emoticon categories. In this way both the numbers of positive and negative emoticons are obtained.

User Interactions with Others. Micro-blog provides three ways for users to interact with each other: one is to mention others in order to attract them via *@username*; another is to forward, aims to pass along those information to their own followers; and the last one is to make comments, which only appear under the original micro-blog to express their own attitudes. In our model, user interactions are measured with these three common parameters: times of mentioning others, times of being forwarded and times of being commented, which could be collected through Sina Micro-blog open platform API [18]. The features input into the classifier are normalized by $Times/|S|$.

User Behaviors in Micro-blog. As the discovery in [6], frequencies of the posting original blogs could indicate depression levels of the user. So the percentage of original micro-blogs is taken as one of the features in user behaviors, calculated as *(number of original posts)/*$|S|$.

It is also found in [6] that the period users post micro-blogs is another indicator of depression level. Table 4 reveals depressive ones tending more active between 0:00-6:00a.m. Therefore, the percentage of micro-blogs posted in this period is calculated as *(number of micro-blogs posted during 0:00-6:00a.m.)/*$|S|$, which is used as another feature in user behaviors.

5 Experiment

5.1 Data Acquisition and Experiment Result

The proposed model is applied to detect depressed users in Sina Micro-blog, a social network service like Twitter. It is one of the most influential SNS in China.

During August 1^{th} -15^{th}, 2012, a group of psychologists made diagnosis on hundreds of volunteers with the means of questionnaire and interviews. They identified 122 depressed sufferers and 346 normal ones. Among them 90 depressed and 90 non-depressed users who use Sina Micro-blog are picked as training dataset. Their information during August 1^{th} -15^{th} are collected through Sina Micro-blog Open Platform API [18]. A total of 6,013 micro-blogs are collected, of which the user who owns the most micro-blogs owns 173, and the least one owns 3 pieces. Over 50,000 sub-sentences are obtained after sentence segmentation. Our experiment is based on these data.

After data processing with methods in section 3&4, ten features are obtained for depression detection. Waikato Environment for Knowledge Analysis (Weka), one of the most useful tools for classification [19], is applied to help classify the users into normal or depressed category. To ensure the result being more reliable, three different kinds of classification approaches are employed: Bayes, Trees and Rules [19-20]. The result is obtained with 10-fold cross validation in Fig.2 (A). ROC Area refers to the area under ROC curve, measuring the quality of a classifier. F-measure is the harmonic mean of precision and recall, denoting the accuracy of a classifier comprehensively.

The result in Fig.2 (A) reveals the precisions of the proposed model with different classifiers are all around 80%, which is considered acceptable by psychologists to detect depressed users in SNS. Among the incorrect cases, the number of individuals incorrectly classified into normal category and incorrectly into depressed category are approximately equal. Most of these individuals own

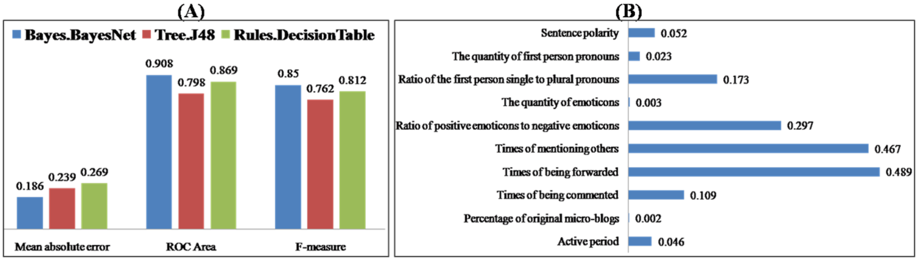

Fig. 2. Classification results of training dataset & Sig. of 10 features in the model

less than 10 pieces of micro-blogs, indicating that lack of data information would bring error to the proposed model.

According to the psychological research, only 3 out of 1000 users online are depressed, so it's very difficult to conduct experiments on large data. However, more than 100 non-depressed familiar friends in Micro-blog are tested with the application in section 6, and more than 85% of them are correctly classified.

5.2 Model Simplification

Besides verifying the effectiveness of the 10 features in the proposed model, the significance of each feature is also studied. Binary logistic regression analysis with Statistical Product and Service Solution (SPSS) is applied to evaluate the significance of each feature in the model [19]. The result is shown in Fig.2 (B). A lower Sig. represents that it is more important. The threshold for Sig. is set as 0.1 and five features are selected for simplifying the model. Experiments show that the precision of the simplified model is declined by less than 5%, however, the bytes of data needed to collect and computing time are significantly reduced. So the application in section 6 is developed with the simplified model.

Furthermore, as Fig.2 (B) shows, the total number of emoticons and original micro-blogs are the most important features, and times of mentioning others and being forwarded are the least important ones. This is a little different from the observation of psychologists in Table 4, which shows times of mentioning others could easily distinguish depressed and non-depressed individuals. It may enlighten psychologists about some further research.

6 Application

Since mental health problem has become the most serious one for modern urban people, the proposed method of sentiment analysis and depression detection model in this paper can be applied in social network services for user mental health state assessment and monitoring. For this purpose, an application in Sina Micro-blog is developed named "Mental Health Testing".

The application provides two functions. One is to calculate the polarity of each piece of micro-blog with the sentiment analysis method, which reflects whether the user is optimistic or not. A user is considered to be “very optimistic”, “a little optimistic” or “pessimistic” according to the total popularity of his latest micro-blogs in a week. The other function is to analyze whether the given user is inclined to be depressed or not with the simplified depression detection model. If a user is tested to be depressed, the application also provides diagnostic messages including some suggestions from the psychologists on active self-regulating strategies.

7 Conclusions and Future Work

A model for detecting depressed users in social network based on sentiment analysis is proposed in this paper. The sentiment analysis method pays special attention to the characteristics of depression and Chinese micro-blog content, and ten features are applied in the depression detection model. The precisions obtained from training dataset are all around 80%, and the significance of each feature in the model is also analyzed for model simplification. An application in Sina Micro-blog is developed to test the polarity of micro-blog and the mental state of users, which has helped psychologists detect several potential depressed users in Sina Micro-blog. Although the depression detection model is proposal based on Chinese vocabulary, the basic idea of the frame especially the sentence structure pattern mining and principle micro-blog features related to depression could be explicitly extended to other language scenarios.

In this work, the training data detected by psychologists are widely scattered in social network. It is hard to analyze the relationship between them, so user interaction is paid little attention and simply three parameters are considered. However, homophily is manifest in the group of depression users, which means, the friends of the depression are more likely to be depressive, and different kinds of interactions indicate different results. Thus the influence of ties between users is contemplated to be studied in the future and a deeper understanding about depression in SNS will be provided.

Acknowledgement. This work is supported by Project on the Architecture, Key technology research and Demonstration of Web-based wireless ubiquitous business environment (No.2012ZX03005008-001) and China-Finland Cooperation Project on The Development and Demonstration of Intelligent Design Platform Driven by Living Lab Methodology (No.2010DFA12780). The authors would like to thank Xiaofeng Qiu, Cheng Cheng and Xinning Zhu for insightful discussions.

References

1. Aggarwal, C.C.: Social Network Data Analytics. Springer, New York (2011)
2. Hsieh, Y., Bolan, J.E.: Predicting Processing Difficulty in Chineses Syntactic Ambiguity Resolution: A Parallel Approach. Poster, The 84th Annual Meeting of the Linguistic Society of America, Baltimore, MD (2010)

3. World Health Organization, http://www.who.int/en/
4. Ramirez-Esparza, N., Chung, C.K., Kacewicz, E., Pennebaker, J.W.: The Psychology of Word Use in Depression Forums in English and in Spanish: Testing Two Text Analytic Approaches. In: Proceedings of the International Conference on Weblogs and Social Media, pp. 102–108. AAAI Press, Menlo Park (2008)
5. Moreno, M., Jelenchick, L., Egan, K., Cox, E., Young, H., Gannon, K., et al.: Feeling Bad on Facebook: Depression Disclosures by College Students on Social Networking Site. Depression and Anxiety 28, 447–455 (2011)
6. Ji, Y.: Social Displacement, Homophily and Depression Levels: The Case of Zoufan on a Chinese Social Network Site. Cyberpsychology, Behavior, and Social Networking. For Peer Review (2012)
7. Pang, B., Lee, L.: Opinion Mining and Sentiment Analysis. Now Publisher Inc. (2008)
8. Davidiv, D., Tsur, O., Rappoport, A.: Enhanced Sentiment Learning Using Twitter Hash-tags and Smileys. In: Proceedings of the 23rd International Conference on Computational in Linguistics, pp. 241–249. Coling 2010 Organizing Committee, Beijing (2010)
9. Barbosa, L., Feng, J.L.: Robust Sentiment Detection on Twitter from Biased and Noisy Data. In: Proceedings of the 23rd International Conference on Computational in Linguistics, pp. 36–44. Coling 2010 Organizing Committee, Beijing (2010)
10. Go, A., Huang, L., Bhayani, R.: Twitter Sentiment Classification using Distant Supervision. Project Report, CS224N (2009)
11. Jiang, L., Yu, M., Zhou, M., Liu, X., Zhao, T.: Target-dependent Twitter Sentiment Classification. In: Proceeding of 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, vol. 1, pp. 151–160 (2011)
12. Parikh, R., Movassate, M.: Sentiment Analysis of User-Generated Twitter Updates using Various Classification Techniques. Final Report, CS224N (2009)
13. Tan, S.B., Zhang, J.: An Empirical Study of Sentiment Analysis for Chinese Documents. Expert Systems with Applications 34, 2262–2269 (2008)
14. NII Test Collection for IR Systems, http://research.nii.ac.jp/ntcir/
15. Dong, Z., Dong, Q.: HowNet—A Hybrid Language and Knowledge Resource. In: Proceedings of International Conference on Natural Language Processing and Knowledge Engineering, pp. 820–824. IEEE Press, Los Alamitos (2003)
16. Institute of Computing Technology, Chinese Lexical Analysis System, http://ictclas.org/
17. Zhang, C.L., Zeng, D., Li, J.X., Wang, F.Y., Zuo, W.L.: Sentiment Analysis of Chinese Documents: From Sentence to Document Level. Journal of the American Society for Information Science and Technology 60, 2474–2487 (2009)
18. Sina Micro-blog Open Platform, http://open.weibo.com
19. Han, J.W., Kamber, M.: Data Mining Concepts and Techniques. Morgan Kaufmann, San Francisco (2006)
20. Witten, I.H., Frank, E.: Data Mining: Pratical Machine Learning Tools and Techniques, 2nd edn. Morgan Kaufmann, San Francisco (2005)