MoodLens: An Emoticon-Based Sentiment Analysis System for Chinese Tweets

Jichang Zhao* zhaojichang@nlsde.buaa.edu.cn
Li Dong* donglixp@gmail.com
Junjie Wu† wujj@buaa.edu.cn
Ke Xu*‡ kexu@nlsde.buaa.edu.cn

*State Key Lab of Software Development Environment, Beihang University
†Beijing Key Laboratory of Emergency Support Simulation Technologies for City Operations, School of Economics and Management, Beihang University
‡Corresponding author

ABSTRACT

Recent years have witnessed the explosive growth of online social media. Weibo, a Twitter-like online social network in China, has attracted more than 300 million users in less than three years, with more than 1000 tweets generated in every second. These tweets not only convey the factual information, but also reflect the emotional states of the authors, which are very important for understanding user behaviors. However, a tweet in Weibo is extremely short and the words it contains evolve extraordinarily fast. Moreover, the Chinese corpus of sentiments is still very small, which prevents the conventional keyword-based methods from being used. In light of this, we build a system called *MoodLens*, which to our best knowledge is the first system for sentiment analysis of Chinese tweets in Weibo. In *MoodLens*, 95 emoticons are mapped into four categories of sentiments, i.e. *angry*, *disgusting*, *joyful*, and *sad*, which serve as the class labels of tweets. We then collect over 3.5 million labeled tweets as the corpus and train a fast Naïve Bayes classifier, with an empirical precision of 64.3%. *MoodLens* also implements an incremental learning method to tackle the problem of the sentiment shift and the generation of new words. Using *MoodLens* for real-time tweets obtained from Weibo, several interesting temporal and spatial patterns are observed. Also, sentiment variations are well-captured by *MoodLens* to effectively detect abnormal events in China. Finally, by using the highly efficient Naïve Bayes classifier, *MoodLens* is capable of online real-time sentiment monitoring. The demo of *MoodLens* can be found at *http://goo.gl/8DQ65*.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications—*Data Mining*; H.3.3 [**Information Search and Retrieval**]: [Text Mining]; J.4 [**Social and Behavioral Sciences**]: [Miscellaneous]

KDD'12, August 12–16, 2012, Beijing, China.

General Terms

Measurement, Experimentation

Keywords

Sentiment Analysis; Chinese Short Text; Online Social Media; Weibo

1. INTRODUCTION

The development of online social networks has attracted enormous Internet users in this decade. They are becoming the mainstream online social media for information sharing. Twitter (www.twitter.com), a microblog site launched in 2006, has over 300 million registered users, with over 140 million microblog posts, known as *tweets*, being published every day. In China, Weibo (www.weibo.com), a Twitter-like service launched in 2009, has accumulated more than 300 millions users in less than three years. Every second, more than 1000 Chinese tweets are posted in Weibo.

Each user in the network can be viewed as a social sensor, which publishes and propagates the information through the tweets. Therefore, the huge amount of tweets convey complicated signals of the authors and the real-world events, in which the sentiment is an essential part. In [3], the authors argued that the events in the social, political and cultural fields did have a significant effect on the users' mood, which could be detected by their tweets. It was also claimed in [2] that the stock market even could be predicted by the sentiment analysis of the Twitter users.

Disclosing the emotions in tweets therefore plays a key role in understanding the user behaviors in social media. However, both Twitter and Weibo only allow users to post messages up to 140 characters, which makes the tweets extremely short, and the sentiment analysis therefore becomes a very challenging task. In particular, few works have been done to reveal how to perform sentiment analysis for Chinese tweets in Weibo.

In light of this, we propose a system called *MoodLens* to perform the sentiment analysis for Chinese Weibo. The main contributions lie in the following aspects: (a) *Moodlens* employs an emoticon-based method for sentiment classification, which helps to tackle the longstanding sparsity problem of short texts; (b) *MoodLens* can detect four types of sentiments: *angry*, *disgusting*, *joyful*, and *sad*, which goes beyond the traditional binary sentiment (positive vs. negative) analysis studies, and is crucial for unveiling the abundant

sentiments contained in tweets; (c) *MoodLens* implements an incremental learning scheme to deal with the problems of the sentiment shift of words and the generation of new words; (d) *MoodLens* is capable of real-time tweet processing and classification, and therefore can serve as a real-time abnormal event monitoring system. The demo of *MoodLens* is now available at *http://goo.gl/8DQ65*.

2. EMOTICON-BASED METHOD

We have noticed that the graphical emoticons are popular in Weibo. In recent work [1], it has been found that the graphical emoticons can convey strong sentiment. They help the users to express their mood when post the tweet. Hence, we could treat these emoticons as sentiment labels of the tweets. In fact, it is a kind of crowdsourcing, i.e., the users label the tweet with emoticons to express their emotion themselves. Because of this, categorizing the emoticons into different sentiments would make the tweets divided into different emotion classes. Among over 1000 emoticons, we manually select 95 ones as the sentiment labels(denoted as E) and divide them into four different sentiment categories, including *angry*, *disgusting*, *joyful* and *sadness*. As show in Figure 1, there are 9 emoticons in *angry*, 14 emoticons in *disgusting*, 50 emoticons in *joyful* and 22 emoticons in *sad*, respectively.

Sentiment	#Emoticons	Typical emoticons
Angry	*9*	
Disgusting	*14*	
Joyful	*50*	
Sad	*22*	

Figure 1: Sentiment categories and the typical emoticons in each class.

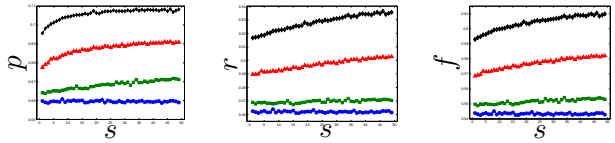
From Dec. 2010 to Feb. 2011, *MoodLens* has collected more than 70 million tweets from Weibo. We extract over 3.5 million tweets that contains emoticons in E as the labeled tweets set, denoted as T. It indicates that in Weibo, there is nearly 5% of the tweets labeled by the sentiment emoticons. Finally we obtain 569,229 *angry* tweets, 290,444 *disgusting* tweets, 2,218,779 *joyful* tweets and 607,715 *sad* tweets. These tweets could be used to as a initial sentiment corpus for Weibo. For each tweet t in T, *MoodLens* converts it into a sequence of words $\{w_i\}$, where w_i is a word and i is its position in t.

In *MoodLens*, we employ the simple method of Naïve Bayes (NB) to build the classifier, which consumes little training time and predicts the category fast. From the labeled tweets, we could obtain the word w_i's prior probability of belonging to the sentiment category c_j is $P(w_i \parallel c_j) = \frac{n^{c_j}(w_i)+1}{\sum_q (n^{c_j}(w_q)+1)}$, where $j = 1, 2, 3$ *or* 4, $n^{c_j}(w_i)$ is the times that w_i appears in all the tweets in the category c_j and *Laplace smoothing* is used to avoid the problem of zero probability. Then we could establish the Naïve Bayes classifier as follows, for a unlabeled tweet t with word sequence $\{w_i\}$, its category could be obtained as $c^*(t) = arg\max_j P(c_j)\Pi_i P(w_i \parallel c_j)$, where $P(c_j)$ is the prior probability of c_j.

In order to validate the performance of the classifier, the set of labeled tweets is divided into two sets randomly, including training set, denoted as T_{train} and testing set, denoted as T_{test}. The fraction of T_{train}, i.e., the fraction of the training data, is denoted as $f_t = \frac{|T_{train}|}{|T|}$. In T_{train}, the set of tweets labeled as c_j is denoted as $T^{c_j}_{train}$, similarly, the tweets in T_{test} of c_j is denoted as $T^{c_j}_{test}$. In the testing set, the correctly predicted tweets of c_j is denoted as P^{c_j}. From these definitions, we mainly employ three metrics in this paper to describe the effectiveness of the classifier, which are listed as follows. **Precision** is defined as $p = \frac{\sum_{j=1}^{4} |P^{c_j}|}{|T_{test}|}$. **Recall** is defined as $r = \frac{1}{4}\sum_{j=1}^{4} \frac{|P^{c_j}|}{|T^{c_j}_{test}|}$. **F-measure** is defined as $f = 2pr/(p+r)$.

In this demo, we use a standard bag of words as the feature, set $f_t = 0.9$, $P(c_j) = 0.25$ and get a Naïve Bayes classifier, its precision is 64.3%, recall is 53.3% and F-measure is 58.3%.

We also present a simple incremental learning approach to complement the original Naïve Bayes classifier. Here, we could assume the tweets in Weibo is a stream, in which there is a fraction of tweets (denoted u) are sentimentally labeled, then these labeled tweets could be used to update the prior probability of words. To verify the effectiveness of the method, the following experiment is performed. we randomly shuffle T and divide it into 50 pieces of same size. Then we use the first piece as the training set and obtain an initial classifier. For the other 49 pieces, we treat them as the tweet stream, which means they enter into the classifier one by one, and in each of them, there is a fraction(u) of tweets are randomly selected as labeled tweets and could be used to update the classifier. As shown in Figure 2, as the index of pieces, denoted as s, grows, the p, r and f of the classifier indeed grows. Particularly, higher u means a larger fraction of labeled tweets are used to update the classifier, and then the more updates produce more improvements.



(a) Precision (b) Recall (c) F-measure

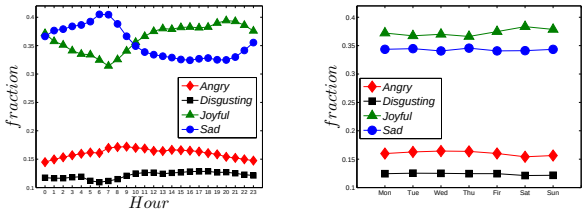
Figure 2: Experiments of incremental learning, $u = 0, 0.01, 0.05, 0.1$ from bottom to top, respectively.

In summary, *MoodLens* employs Naïve Bayes classifier with incremental learning to predict the detailed sentiment of the tweets. For other solutions, like Liblinear [4], which consumes much more training time while gain less than 5% improvement of precision. Moreover, it is also hard to incorporate incremental learning approaches into it.

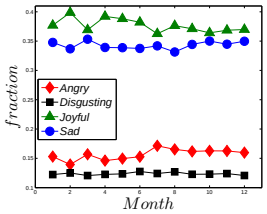
3. APPLICATIONS

Data Collection Weibo has published its APIs since 2010 and through these APIs, it is easy to obtain the public tweets and some basic demographic attributes of the users. We build a Weibo application named *"Are you happy?!"* and

Weibo grant us the opportunity to access its APIs with the application-level. For the limitation of requesting the API, it is necessary to select some probes from Weibo and then collect data from them. From a large-scale user-pool we collected before 2011, which contains more than 2.2 million users, *MoodLens* randomly selects 6,800 active users, 200 users for each province or region in China. Here the "active users" means these users should be true users but not spam. A simple filtering rule is used to filter them out, which is that *MoodLens* only chooses the user with more than 200 but less than 3,000 followers and has published less than 3,000 tweets.



(a) Hourly pattern. (b) Weekly pattern.



(c) Monthly pattern.

Figure 3: Temporal sentiment patterns.

Sentiment Patterns The hourly pattern of the sentiment is showed in Figure 3(a). It could be found that the time from 6:00 AM to 8:00 AM is the saddest moment, which is different from the recent study from Twitter [5]. In their data set, they found people are likely to be positive at early morning. While for Weibo, this period is also the angriest moment for most of the users during a day. Surprised by this difference, we carefully investigate the tweets published in Weibo from 6:00AM to 8:00AM and extract the commonly used words. And the results show that the "sad" mood is generally caused by the followed reasons. First, some people hate to get up early, but they have to. Second, some users do not want to work at this time. Third, some ones might have nightmare in the last night and have bad sleep. After 10:00 AM, users of Weibo seems to become more and more joyful gradually and the fraction of *joyful* tweets reaches the peak at 20:00 PM. The weekly pattern of the sentiment is showed in Figure 3(b). As can be seen, people seems to become happier since Friday, the *joyful* reaches peak at Saturday and then the mood of joy goes down as Sunday begins. As shown in Figure 3(c) is the monthly pattern of the sentiment. There are several outliers. For instance, in March of 2011, it shows the users of Weibo are sad and angry, it might be caused by the earthquake in Japan and the rumor that the iodized salt in China is also nuclear polluted. Another one is in July, the fraction of *angry* reaches the peak, it is mainly related with the accident of the bullet train. We also find the days of extreme sentiment in 2011. Namely, Jan. 1 is the most joyful day, Mar. 11 is the saddest day, Jul. 13 is the most disgusting day and Jul. 24 is the angriest day (refer to [6] for the link). *MoodLens* also draws the distribution graph of everyday sentiment for each region of China and show how the sentiment evolves day by day dynamically (refer to [7] for the link).

Abnormal Event Detection Intuitively, abnormal events in the real world would definitely affect the people's emotion, and then the mental change would be reflected by the tweets people publish. The basic idea of the detection method is first to find the turning point in the variation of the sentiment and then to extract information of event from the tweets. *MoodLens* defines a sequence of fraction for the sentiment c_j as $\{S_t^{c_j}\}$, where t is the observing time, its unit is likely to be a day or an hour. Assuming we observe the variation of the sentiment from $t = t_1$ to $t = t_2$, then the averaged fraction tweets in c_j could be defined as

$$\langle S_{t_1 \to t_2}^{c_j} \rangle = \frac{1}{t_2 - t_1} \sum_{t=t_1}^{t_2} S_t^{c_j}, \tag{1}$$

where $\Delta t = t_2 - t_1$ is the time window of observation. Hence, *MoodLens* could get the sequence of relative variation for c_j as

$$V_t^{c_j} = \frac{S_t^{c_j} - \langle S_{t_1 \to t_2}^{c_j} \rangle}{\langle S_{t_1 \to t_2}^{c_j} \rangle}. \tag{2}$$

Then *MoodLens* defines the sequence of sentiment variation as $\{\sum_{j=1}^{4} |V_t^{c_j}|\}$. This sequence could be sorted in descending order and the $top-k$ t is selected as the outlier time points, denoted as $\{t_1, t_2, .., t_k\}$. For each t_i, *MoodLens* could extract the tweets posted at that time and perform the information extraction for the event. Here *MoodLens* employs the simplest way, the top 5 bi-gram terms of high frequency would be extracted to depict the event happened. We perform this method on the data set of 2011 and find it could detect almost all the abnormal events happened during the whole year. As shown in Figure 4, we mark the $top-10$ events detected from A to J. In these detected events, A,D and E correspond to the event of bullet train crash. C and B correspond to the fact that Japan was hit by a magnitude 9.0 earthquake, while J corresponds to the news that people in China rushed to purchase the salt because of rumors. It should be noted that for J, the fraction of *angry* is larger than C and B. F corresponds to New Years' Eve of 2011. G and I correspond to Spring Festival of 2011. H corresponds to the death of Steven Jobs. It is also interesting that different from C, B, J, A, D and E, for H, although the fraction of *sad* is high but the fraction of *angry* is low. It indicates that detailed negative sentiments are useful for analyzing the essence of the abnormal event.

Real-time Sentiment Monitoring The NB classifier with incremental learning is speedy enough for the real-time sentiment analysis of the tweets in Weibo. Through the API provided by Weibo, *MoodLens* could obtain the most recent 400 public tweets every minute and these tweets could be analyzed in less than one second. In order to guarantee the statistical significance, we set the cycle of collecting tweets as 30 minutes, which means *MoodLens* would download nearly 12,000 tweets in one monitoring cycle. Then these tweets would be categorized into different sentiment classes in less than 1 minute. As shown in Figure 5, we present a sample cycle from the real-time monitoring, which starts from 19:30

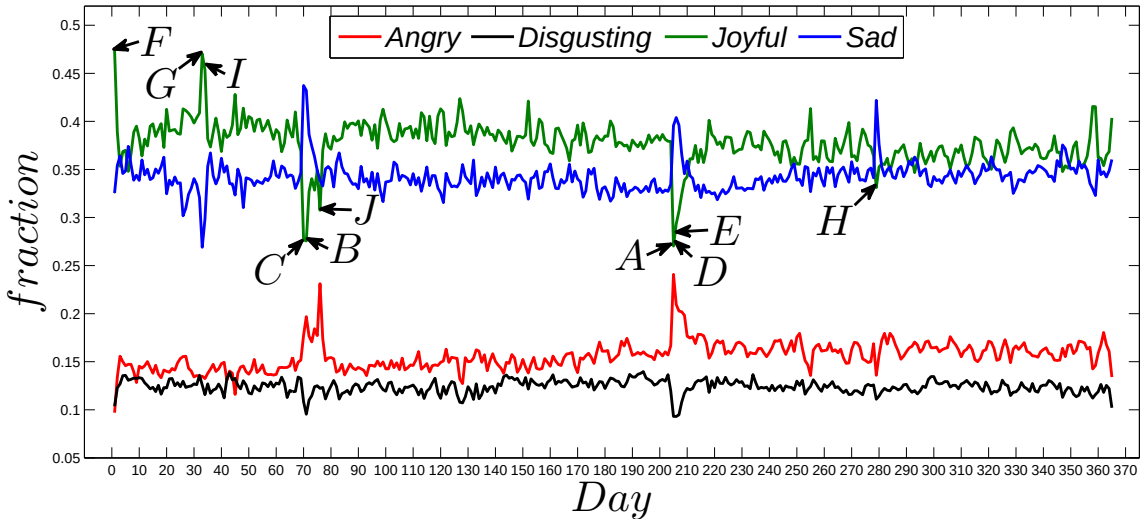


Figure 4: Abnormal event detection of 2011. We also provide an interactive version on the following link: http://gana.nlsde.buaa.edu.cn/hourly_happy/timeline_2011.html.

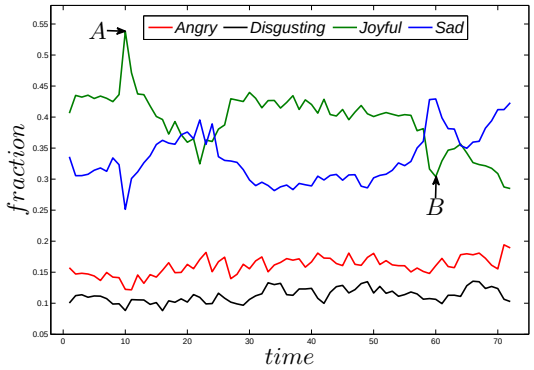


Figure 5: A sample from real-time monitoring. The online application could be accessed through the following link: gana.nlsde.buaa.edu.cn/hourly_happy/line_36h.html.

PM of Dec. 24, 2011 and ends to 7:00 AM of Dec. 26, 2011, the delay between two points is 30 minutes and 72 points are drew in all. As can be seen, the real-time event could be indeed monitored. For instance, regarding to the time marked by A, which is at 0:00 PM of Dec. 24, the sentiment of *joyful* reaches peak because everyone is celebrating the coming of the Christmas Day. While with respect to B, at the time of 1:00 AM of Dec. 26, the fraction of *sad* tweets rises suddenly, the bi-gram terms reveal that it is caused by the earthquake happened in Chengdu at that time.

4. CONCLUSION

MoodLens is an online sentiment analysis system for Chinese tweets in Weibo. It employs the emoticons for the generation of sentiment labels for tweets, and builds an incremental learning Naïve Bayes classifier for the categorization of four types of sentiments: *angry*, *disgusting*, *joyful*, and *sad*. *MoodLens* is now available online for temporal and spatial sentiment pattern discovery, abnormal events detection and illustration, and online real-time monitoring of sentiment fluctuations.

5. ACKNOWLEDGEMENTS

This works was supported by the fund of the State Key Laboratory of Software Development Environment (SKLSDE-2011ZX-02) and the Research Fund for the Doctoral Program of Higher Education of China (Grant No. 20111102110019). The third author was supported in part by the National Natural Science Foundation of China (NSFC) under Grants 71171007, 70901002, and 90924020.

6. REFERENCES

[1] Sho Aoki and Osamu Uchida. A method for automatically generating the emotional vectors of emoticons using weblog articles. ACACOS'11, pages 132–136, Stevens Point, Wisconsin, USA, 2011.

[2] Johan Bollen, Huina Mao, and Xiaojun Zeng. Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1–8, 2011.

[3] Johan Bollen, Alberto Pepe, and Huina Mao. Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. 5th ICWSM, 2011.

[4] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. Liblinear: A library for large linear classification. *J. Mach. Learn. Res.*, 9:1871–1874, June 2008.

[5] Scott A. Golder and Michael W. Macy. Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures. *Science*, 333(6051):1878–1881, September 2011.

[6] http://gana.nlsde.buaa.edu.cn/hourly_happy/line_cycle.html.

[7] http://gana.nlsde.buaa.edu.cn/hourly_happy/map_pie.html.