Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 125 (2018) 313–320

Procedia Computer Science

www.elsevier.com/locate/procedia

6th International Conference on Smart Computing and Communications, ICSCC 2017, 7-8 December 2017, Kurukshetra, India

SoftMax based User Attitude Detection Algorithm for Sentimental Analysis

Bhavish Khanna N [a], Sharon Moses J[b], Nirmala M[b]*

[a]*School of Computer Science and Engineering, VIT University, Vellore 632014, India*
[b] *School of Information Technology and Engineering, VIT University, Vellore 632014, India*

Abstract

The social microblogging sites empower users to more boldly express their views and reviews on various topics. This has led to the development of a new subdomain in the field of sentiment analysis which can be used for the benefit of various business. The business values associated with mining the user sentiments havefueled a profusion of researchers, companies etc. to get involved in the process to build, update and even promote their products. A majority of the existing methods utilize only the semantics in the user comments to mine the sentiments. The reliability of the comment is not taken into account in the previously existing methods which can severely undermine the sole purpose of sentiment analysis. In this work, SoftMax based attitude detection algorithm is proposed to identify the user nature efficiently. The reliability and the accuracy of the sentiment prediction can be substantially increased with taking user attitude into account. The proposed algorithm is evaluated on tweets fetched from micro blogging website twitter.

Peer-review under responsibility of the scientific committee of the 6th International Conference on Smart Computing and Communications

Keywords:Sentiment Analysis; User Attitude; Softmax Classifier; Reliability; opinion mining

1. Introduction

Pervasive developments in the internet domain combined with affordable smart devices has empowered millions with access to internet. In today's world, Internet and associated services are extensively used for a multitude of purposes viz. social networking, entertainment, transactions, etc. A plethora of social networking websites exist in the internet pool among which few have a broad user base, hence, a greater reach.

* Corresponding author. Tel.: +0-000-000-0000 ; fax: +0-000-000-0000 .
*E-mail address:*nirmaladhinesh@gmail.com

1877-0509
Peer-review under responsibility of the scientific committee of the 6th International Conference on Smart Computing and Communications
10.1016/j.procs.2017.12.042

Twitter, one among them, helps the users to express their opinions on various domains ranging from, but not limited to, politics, new products, scientific breakthroughs etc. Twitter serves as a enormous pool of user opinions which can be exhausted for the sentiments by business analysts, marketing teams, etc. The sentiment in the tweets of a given domain is mined and analyzed using an opinion detection algorithm. The existing algorithms take only the tweets into account ignoring the nature of the users which is essential to understand the reliability of the tweets. The major flaw with the existing methods is that the reliability of the tweets can become dubious. For instance, consider a user whose tweets are predominantly negative. There's a slightest possibility that the user can tweet negatively about a product even though it is actually good. Hence, the user can be, to some extent, considered to be pessimistic and this factors in additional complexities to the analysis as the reliability of the tweet will be questionable. It is essential to take into account the nature of the user in order to minimize the user's attitude's influence on the tweet and to predict the actual sentiment of the tweet with a greater accuracy. In this paper, a methodology is proposed for improving the accuracy of the prediction of the sentiment of the tweet aided by user nature detection. Users can be classified into multiple classes depending upon their attitude. Some tend to be optimistic and whatever they express will be positive while the pessimistic people tend to express it in a negative way. Moreover, the users can also be neutral. Hence, it is essential to identify the nature of the user before analyzing the tweets in order to enhance their reliability. User attitude can be determined by analyzing the tweets posted by the user and apply SoftMax regression to it in order to identify it more accurately. The proposed user attitude identifier algorithm is evaluated on corpus of twitter data. The paper is constructed as follows. In section 1, introduction and need for identifying the user nature is elaborated. In section 2, existing works on user nature and importance of mining the user nature is discussed. Section 3 explains the proposed SoftMax classified based user attitude detector. Section 4 and 5 elaborates about the evaluation of the proposed algorithm and concludes the paper with references.

2. Similar Works

Before the evolution of web 2.0 and invention of social networking websites, first hand opinions from other individuals, also known as word of mouth, played a significant role in influencing the buying decision [1]. Later, the advancement of social networking has led way to the development of a new way of communication known as the electronic word of mouth [2]. Henning and others define electronic word of mouth (eWOM) as the statement made by the users and former users of a product or a brand that is made available to multitude of people through internet [3]. The authors states that even though the eWOM is not a face to face communication, it remains powerful because of its accessibility, being in print nature and its immediate reach. eWOM in the form of user opinions, sentiments, deeply felt emotions, and expressing ideas has an ever-growing impact on the ecommerce industry [4][5]. Bernard and others who made the investigation on the eWOM concluded that consumers depend and trust the social networks for opinions and insights [2]. Moreover, the author also states that in future, twitter will be one among the key social networking application that user will look up for trustable information. Twitter, which allows the users to post a 140-character message, is one among the rapidly growing microblogging website. Currently, many companies analyze twitter data to garner the reviews of their product, market the product and as well as to get in touch with the consumer. The business perspective of mining opinion from the twitter context makes sentiment analysis on twitter one of the trending researches in the recent times [6]. Twitter based sentiment analysis is done for various fields ranging from product review to stock market prediction [7] [8] [9] [10] [11]. Sentimental analysis is classified into three types namely lexicon based opinion mining, machine learning based methods, and hybrid methods [6] [12].

Generally, in lexicon based sentimental analysis detection method, the context of the tweet is compared with the dictionary of the words to mine the polarity of the tweet. Based on the polarity, the tweet is classified as positive or negative [12]. In machine learning methods, machine learning algorithm are trained to classify the polarity of the tweets [13] [14]. Hybrid approach combines both lexicon and machine learning methods. In all these methods, only the context is analyzed and mined for the sentiments. The reliability of the content and the reliability of the user who tweeted is not taken into account. Mining sentiment on unreliable information will return hazy results which may affect the accuracy of the sentiment prediction. In order to enhance the reliability and prediction accuracy of the sentiment analysis, SoftMax based attitude detection algorithm (SAD) is proposed. The SAD method will act as pre-processing stage of the sentimental analysis methods. Based on the history of user tweets, the attitude of the user is classified using SoftMax Regression. To the best of our knowledge, none of the existing sentiment analysis method

concentrate on generating user attitude before finding the sentiments. In the next section, the proposed SAD algorithm in identifying the user attitude is explained.

3. SoftMax Based Attitude Detection Algorithm

Finding the user attitude involves fetching the tweets posted by the user and analyzing it. Given a domain from the medley of tweets that exist, the relevant tweets are fetched to extract the user information from it. Based on the user details, history of tweet posted by each user is analyzed to predict the user attitude. User attitude is classified into three categories, namely, user with neutral attitude, optimistic attitude and pessimistic attitude. As in the algorithm, based on the user tweets over a domain, the user id of the user u_i is fetched and a saved to a user list. From the user list, tweets posted by each user is analyzed whether the user has tweeted more than threshold F. In this algorithm, the threshold F is fixed at fifty tweets. Hence, only the users who have tweeted more than fifty tweets are only considered for attitude detection. Since the proposed SAD algorithm works on the history of the tweets, it is essential to have a significant number of tweets to mine the user attitude. Once the identification phase is complete, each tweet $t_j$made by the user $u_i$is fetched and saved to list $L(t)$. From the list $L(t)$, tweets $u_i(t_j)$ made by the user $u_i$are tokenized, stopwords are removed and the polarity $P(t_j)$ of the tweet is extracted using lexicon based sentiment analysis. Each word in the tweet is compared against the bag of positive (b_P) and negative words (b_N) to compute the polarity $P(t_i)$ of the tweet.

SAD Algorithm

```
For each user ui in the user list
  Count the number of tweets n(ui)
  If n(ui)>F:
    Fetch the Tweets t and Save to List L(t)
    For ui in L(t):
      For j in ui(tj):
        Tokenize each word in tweet tj
        Remove Stop Words
        Calculate P(tj)
EndFor
      Save to List L(t)
      Fetch n(ui) and P(ti) for each user
      Get the nP,nN,nM of each User
      Input n(ui), nP,nN,nM to C
EndFor
    Get the probability
Analyze and Assign Attitude
  Else:
    If End of List==False
      Try for next User
EndIf
EndFor
```

Consider $w_k$to be the word belong to tweet $u_i(t_j)$ posted by the user where k is the total number of words in the tweet. Each of the tweet words are compared with b_P and b_N and if a match is found in positive bag b_P, the count of score $S(b_P)$ will be incremented by one. Similarly, if a match is found in the negative bag of words b_N, the score of $S(b_N)$ will be incremented by one. The computation of scores $S(b_P)$ and $S(b_N)$ is depicted in the equation (1) and (2) where T_W is the total number of words belong to the tweet.

$$S(b_P) = \frac{no\ of\ (w \in b_P)}{T_w} \qquad (1)$$

$$S(b_N) = \frac{no\ of\ (w \in b_N)}{T_w} \quad (2)$$

$$P(t_j) = \begin{cases} (t_j)^P : S(b_P) > S(b_N) \\ (t_j)^N : S(b_P) < S(b_N) \end{cases} \quad (3)$$

The scores $S(b_P)$ and $S(b_N)$ is compared with each other to find the polarity of the tweet. If $S(b_P)$ is greater than $S(b_N)$ then the polarity of the tweet is classified as positive. On the contrary value of $S(b_N)$ is higher than the positive score then the tweet polarity $P(t_j)$ is assigned as negative this is depicted in the equation (3). When the scores return zero or no words present in the positive and negative bag of the words then the tweet is classified as neutral and $P(t_j)$ takes the value $(t_j)^M$. Similarly, the polarity of all the tweets posted by each user is detected and saved to list $L(t)$.

Following the polarity computation phase, the number of positive n_P, negative n_N and neutral n_M polarity tweets posted by the user are in the equation (4) (5) and (6). Where the count for total number of tweets posted by user $u_i$having positive polarity is taken as the value for n_P as in the equation (4). Similarly, the count for negative and neutral tweets of the user u_i is computed as shown in equation (5) and (6).

$$n_P = count\ of\ (t_j)^P \in u_i \quad (4)$$

$$n_N = count\ of\ (t_j)^N \in u_i \quad (5)$$

$$n_M = count\ of\ (t_j)^M \in u_i \quad (6)$$

After finding the sentiment of the tweets and number of tweets posted by the user, users are classified into three classes namely optimistic, pessimistic and neutral using SoftMax regression. The SoftMax regression based analysis consist of two phases. In the first phase, the evidence of the input being in a certain class is added up, and then, the evidence is converted into probabilities. The evidence, also known as the membership value, is calculated by taking the n_P, $n_N$andn_M into account. For any given user, the count of total tweets and the sum of the n_P, $n_N$andn_M should be equal. The tweets of the user will be distributed in one of these three categories. In order to compute the evidence, a weight as well as a bias is added for a better estimate. The weight, $W_{i,j}$, where i is the user and j is the class, is negative if the tweet class is pessimistic. If the tweet class is optimistic, the weight will be positive. The bias, also known as extra evidence, is added to express that the evidence in some cases can be independent of the input. Hence, the evidence for a user i given an input x is show in equation (7) where x takes the value the n_P, $n_N$andn_M. where W_i is the weights and bi is the bias for class i, and j is an index for summing over the user data.

$$evidence = \sum_j W_{i,j}, x_j + b_i \quad (7)$$

$$y = softmax(evidence) \quad (8)$$

$$Softmax(x) = normalize(\exp(x)) \quad (9)$$

$$Softmax(x_i) = \frac{\exp(x_i)}{\sum_j \exp(x_j)} \quad (10)$$

The evidence tallies are then converted into predicted probabilities y using the "softmax" function as in equation (8). Here, the softmax serves as an activation function by shaping the output of our linear function into a probability distribution over 3 classes. The above process can be thought as converting tallies of evidence into probabilities of input being in each class. This is depicted in the equation (9) and the expansion of normalization is shown in equation (10). Once the probabilities are assigned for each class at the end of Softmax regression, these values can be treated as the extent to which the user exhibits the traits of that class. For instance, consider a user, ux, with the following values, np = 0.45, nn = 0.25 and nm = 0.3. From the above example, it can be inferred that the user ux exhibits optimistic traits 45% of the times, pessimistic 25% of the times and is neutral 30% of the times. In the following section, the user attitude obtained by SAD algorithm is evaluated user twitter data and validated by statistical method.

4. Evaluation of SAD Algorithm

For the purpose of evaluating SAD algorithm, the tweets posted by nearly 250 users are fetched from the twitter. In the total 250 of users, min number of tweets made by the user is 50 and maximum number of tweet is approximately 500. The SAD algorithm is evaluated on the fetched tweets and the results are shown in figure 1. Where 1, 2, and 3 in the z axis corresponds to optimistic, neutral and pessimistic attitudes of the users.

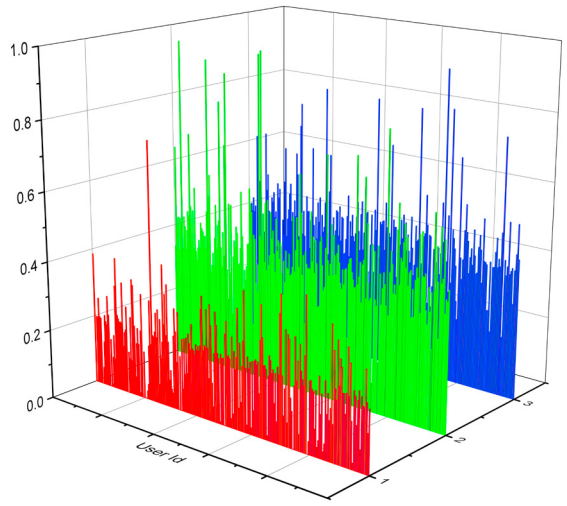

Fig. 1. User Nature Classification

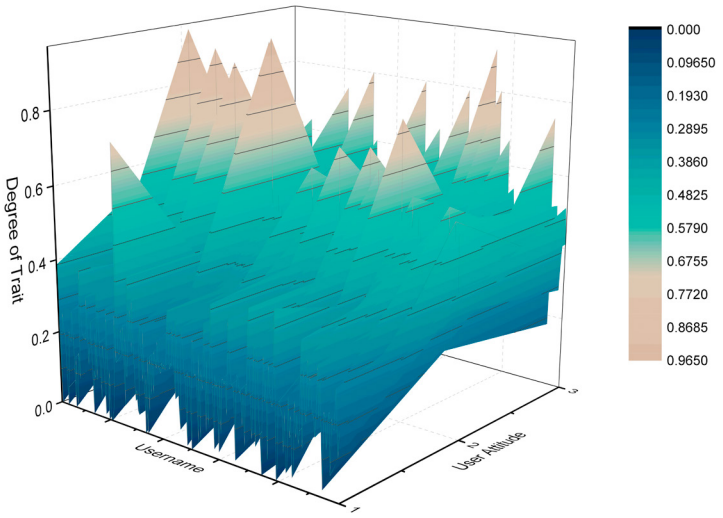

Fig. 2.Comparison of Degree of Trait and User Nature

In figure 2, the possible attitude of user is plotted. From which, it is evident that some users are totally optimistic while some are totally pessimistic in expressing their opinion. This variation in the user attitude will strongly affect the sentiment of the tweet. In order to evaluate the SAD algorithm, the results of the algorithm are

verified statistically using regression analysis based on the manual classification. The analysis of the SAD algorithm are depicted in the figures 3,4, and 5; where reference line indicates the expected results and observed results are denoted by percentiles.

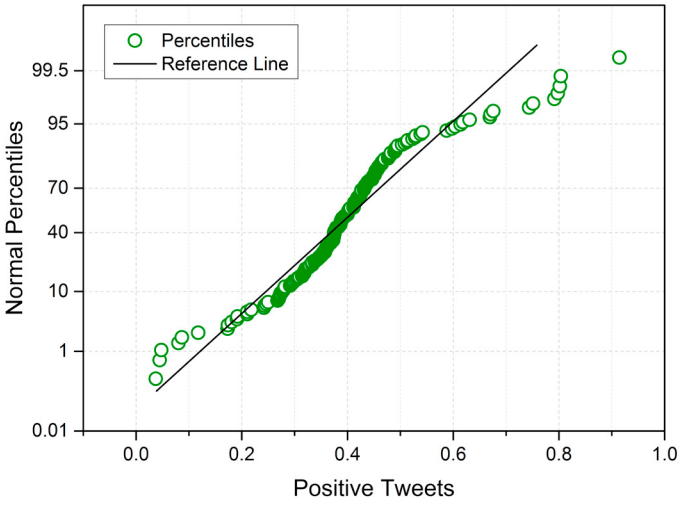

Fig. 3. Optimistic Users

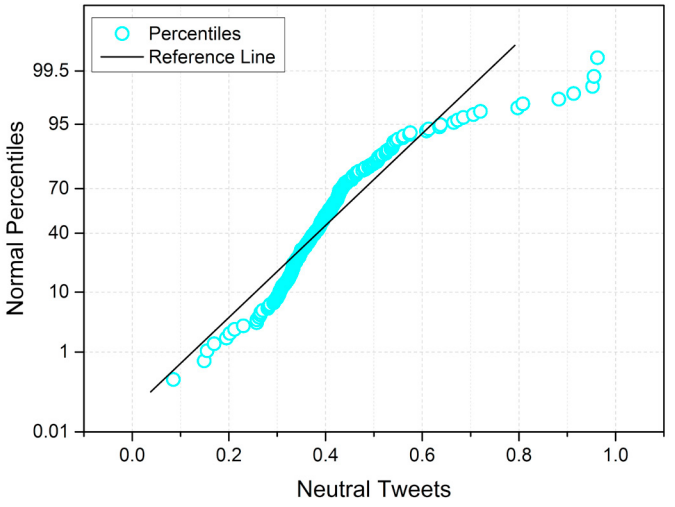

Fig. 4.Neutral Users

In detecting the positive nature, proposed algorithm exhibited high accuracy which can be witnessed in figure 3, where a majority of the detected attitude coincides with the manual classification outcomes. Also, similar trends can be found in determining the neutral and pessimistic attitude of the users.The error percentage of the proposed algorithm is computed by finding the mean absolute error as depicted in the equation (11). Where the prediction made by the SAD algorithm is denoted by P_i, $A_i$corresponds to the actual attitude of the user and n stands for the total user considered for calculating the error percentage.

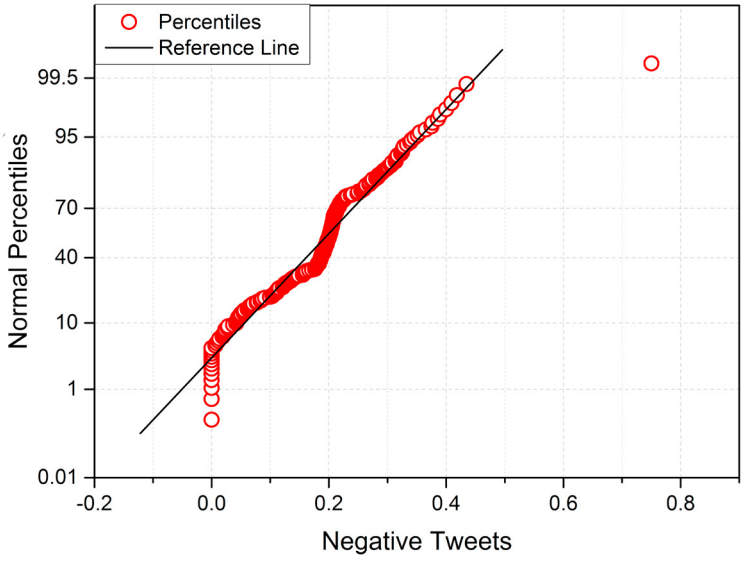

Fig. 5. Pessimistic Users

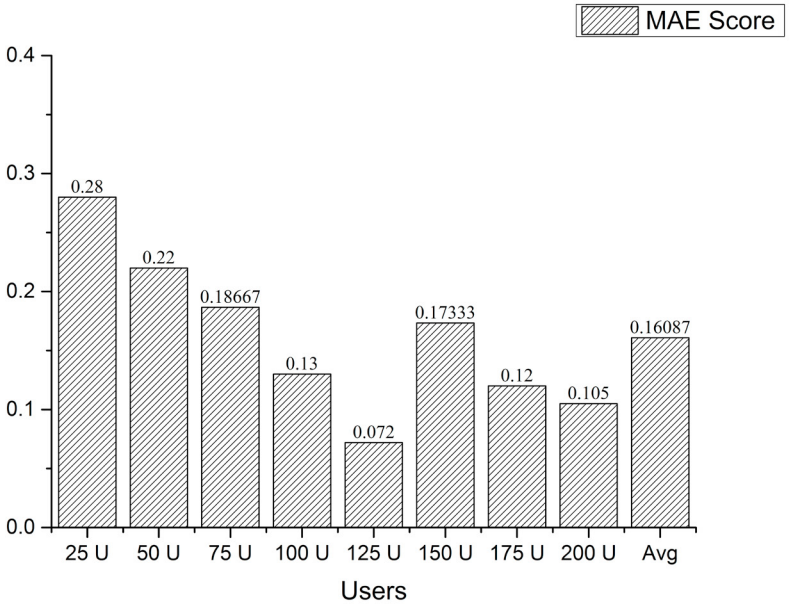

Fig. 6.MAE Scores

The computation of the mean absolute scores for different number of users is illustrated in the figure 6. The highest error accounted for twenty-five users whereas the average error percentage is 0.16087. From the error percentage, it can be conceived that the proposed SAD algorithm will have higher prediction accuracy since the highest error percentage itself marginal. The evaluation of the prediction followed by the computation of the MAE scores shows that the proposed algorithm will detect the user nature more accurately.

5. Conclusion

In this paper, a SoftMax regression based attitude detection algorithm is proposed to identify the user nature of the user. By finding the nature of the user, reliability of the tweet posted by them can be verified which will in turn help the data miners to obtain tweets that are trustworthy. The evaluation of the proposed work proves that the SoftMax regression based algorithm predicts the user attitude with great accuracy. Currently, only the nature of the user is rendered using SAD algorithm. In future, the work will be extended to effectively detect the user attitudes over the user views. Also, the proposed algorithm can open new possibilities in identifying the people with depression, ill intents thereby avoiding the unexpected happenings.

References

1. Marsha L. Richins and Teri Root-Shaffer. (1988) "The role of evolvement and opinion leadership in consumer word-of-mouth: An implicit model made explicit." *Advances in Consumer Research***15(1)**: 32-36.
2. Jansen Bernard J, Mimi Zhang, Kate Sobel, and Abdur Chowdury.(2009) "Twitter power: Tweets as electronic word of mouth." *Journal of the Association for Information Science and Technology***60(11)**: 2169-2188.
3. Hennig-Thurau, Thorsten, Kevin P. Gwinner, Gianfranco Walsh, and Dwayne D. Gremler.(2004) "Electronic word-of-mouth via consumer-opinion platforms: what motivates consumers to articulate themselves on the internet?." *Journal of interactive marketing***18(1)**: 38-52.
4. Davis, Alanah, and Deepak Khazanchi.(2008) "An empirical study of online word of mouth as a predictor for multi-product category e-commerce sales." *Electronic Markets***18(2)**: 130-141.
5. Park Cheol and Thae Min Lee.(2009) "Information direction, website reputation and eWOM effect: A moderating role of product type." *Journal of Business research***62(1):** 61-67.
6. Pandey, Avinash Chandra, Dharmveer Singh Rajpoot, and Mukesh Saraswat.(2017) "Twitter sentiment analysis using hybrid cuckoo search method." *Information Processing & Management***53(4)**: 764-779.
7. Thelwall, Mike, Kevan Buckley, and Georgios Paltoglou. (2011) "Sentiment in Twitter events," *Journal of the Association for Information Science and Technology***62(2)**: 406-418.
8. Alexander Pak and Patrick Paroubek (2010) "Twitter as a corpus for sentiment analysis and opinion mining." *inProceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta.
9. Kouloumpis, Efthymios, Theresa Wilson, and Johanna D. Moore (2011) "Twitter sentiment analysis: The good the bad and the omg!." *In Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, Barcelona, Catalonia, Spain.
10. Agarwal, Apoorv, Boyi Xie, Ilia Vovsha, Owen Rambow, and Rebecca Passonneau (2011) "Sentiment analysis of twitter data*," in Proceedings of the Workshop on Languages in Social Media*,Portland, Oregon.
11. Bollen, Johan, and Huina Mao. (2011) "Twitter mood as a stock market predictor." *Computer***44(10)**: 91-94.
12. Liu, Bing, and Lei Zhang (2012) "A Survey of Opinion Mining and Sentiment Analysis." inCharu C. AggarwalChengXiang Zhai (eds)*Mining text data*, Boston, MA, Springer.
13. Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan (2002) "Thumbs up?: sentiment classification using machine learning techniques." *in Proceedings of the ACL-02 conference on Empirical methods in natural language processing*, Stroudsburg, PA.
14. Ye, Qiang, Ziqiong Zhang, and Rob Law. (2009) "Sentiment classification of online reviews to travel destinations by supervised machine learning approaches." *Expert systems with applications***(36)3**: 6527-6535.