Identification of Opinion Leaders in Social Networks Based on Sentiment Analysis: Evidence from an Automotive Forum

Maoran Zhu[1], Xingkai Lin[1,*], Ting Lu[2] and Hongwei Wang[1]
[1]School of Economics and Management, Tongji University, Shanghai 200092, China
[2]Saic General Motors Sales Co., LTD, Shanghai 201206, China
[*]Corresponding author

***Abstract*—Opinion leadership of social network plays an important role in the fields of knowledge spreading, public sentiment controlling, marketing, etc. The opinions of users are derived from the reviews of topics, and the analysis of users' sentiment is helpful in recognizing users' emotional preference of opinion leaders. Thus, it is necessary to classify the forum posts and refine the posts with highly professional knowledge. We then improve the sentiment analysis due to the imbalanced datasets, and establish a comprehensive attention and emotion weight matrix. Accordingly, in this paper, we are going to propose a Leader-PageRank algorithm, which is based on the social network structure and emotional tendency. We do the comparative experiments on the automotive forum, and the results show that the Leader-PageRank algorithm can identify the positive opinion leaders in the professional fields effectively through connecting with the interactions in social networks.**

Keywords-social network; opinion leader; sentiment analysis; short text classification; automotive forum

I. INTRODUCTION

With the increasing influence of the social network, the network center influential nodes which called opinion leaders have become more and more important in public opinion, knowledge dissemination and business recommendation. Rice [1] finds STD/HIV stigmatizing attitudes can be reduce through community popular opinion leaders. Cho's [2] research shows the opinion leaders with high sociality are the best marketing choice for fast diffusion, and the opinion leaders with high distance centrality are the best ones for the maximum cumulative number of adopters. On the other side, some opinion leaders may cause bad influence because of their own benefit, Sismondo's [3] research shows the pharmaceutical industry often turns to key opinion leaders to disseminate scientific information in its marketing efforts. These opinion leaders have accumulated the power to shape the information on which many others base their decisions, which may cause corruption problem.

The traditional method of identifying opinion leaders pays less attention to the professional and positive opinion leaders. And it will dig out some certain controversial opinion leaders. Based on this, this paper will combine sentiment analysis and topic classification to find professional and positive opinion leaders.

II. LITERATURE REVIEW

A. *Influential nodes of Social Network*

The influential nodes of social network play an important role in public opinion, advertising, marketing and knowledge dissemination. The opinion leader identification is affected by many factors. It's the reason why there are many different methods of opinion leader recognition. By using YouTube data, Yoganarasimhan [4] identifies the opinion leaders through the size and structure of the social network around the nodes. Zhang [5] extracts the communities by analyzing the replies of each post in the bulletin board system, and he proposes an opinion leader community mining method based on the level structure. Bakshy [6] finds the largest cascades tend to be generated by users who have been influential in the past and who have a large number of followers by using Twitter Data.

The existing methods focus the factors which include social structure and user attribute itself, but the interaction factors such as sentiment are less involved.

B. *Sentiment Analysis*

Sentiment analysis can identify sentiment of the text is positive, negative or neutral. And the comment text emotion usually shows the attitude of the user.

The common sentiment classifiers which use natural language processing techniques including Support Vector Machine (SVM), Naive Bayes (NB) and Maximum Entropy (ME). Pang uses three machine learning methods to find whether a movie review is positive or negative. The SVM has best accuracy rate, reaching 82.9% [7]. O'Keefe evaluates a range of feature selectors and feature weights with both Naive Bayes and Support Vector Machine classifiers by using IMDB data. The results show SVM classifiers maintain a state-of- the art classification accuracy of 87.15% while using less than 36% of the features [8]. Therefore, this paper uses SVM algorithm as the classifier to find out user comments emotion, and combines the social network structure to establish a comprehensive attention and emotion weight matrix.

III. METHODS

A. *Model Overview*

Research framework is shown in Figure I

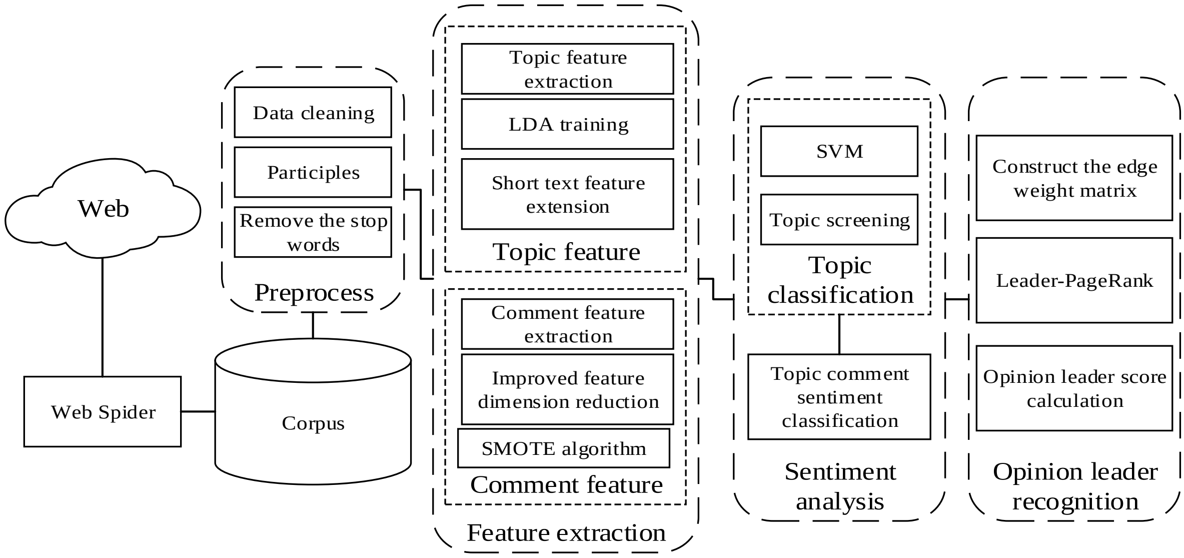

FIGURE I. THE OVERALL IMPROVED METHOD STRUCTURE.

B. Short Text Classification Based on Expanding Feature of LDA

Typically, opinion leaders get people's support through the topic post they made. In the social network, the theme of user attention is limited. For example, the users' comments show their area of interest in social networks. The social networks object of this study is an automotive forum, and the forum topics posts are divided into four categories: 1) Vehicle-related discussions, like vehicle information, repair, maintenance, modification and vehicle accessories; 2) Life trivia, including introduce trivia of life or share the road journey landscape. 3) Forum activities; 4) Forum services. This study assumes that the professional opinion leader's authority comes from the first kind of topics posts, the vehicle-related discussions.

This research uses post names as the classification of objects to extract the professional topics posts, because the post names are able to summarize the content of topic posts, and text content may be constituted by the image, which will cause low classification accuracy. However, the post names are usually short text which will face feature sparse issues on classification.

In order to improve the accuracy of short text classification, this study uses Latent Dirichlet Allocation (LDA) model to expand the feature of short text [9].

The LDA-SVM short text classification's basic idea is to pick a sample of topic posts main body as a long document to train LDA topic model. Then the corresponding topic post names are set as a short text into the model to match the keywords, and use these keywords to expand the feature. Algorithm pseudo-code is shown in the below.

Algorithm 1. The topic classification pseudo-code

```
Input: Long text data: $L = \{l_1, l_2, \ldots\ldots, l_i\}$,  Short text data: $S = \{s_1, s_2, \ldots\ldots, s_i\}$
Output: The category of topic
Model = LDA(topics = 50);
$M1$ = model.fit($L$);
For each $s_\mathrm{i}$ in $S$
        Get $s_\mathrm{i}$ topic features $Ft = \{Ft_1, Ft_2, \ldots\ldots, Ft_i\}$ and probability by using model $M1$;
        $F$ = ExactFeatures( $s_\mathrm{i}$ );
        $Fs$ = ExtendFeatures( $F, Ft$ );
End for
Divide the $S$ into train group $Tr = \{tr_1, tr_2, \ldots\ldots, tr_m\}$ and test group $Te = \{te_1, te_2, \ldots\ldots, te_n\}$ randomly;
Label $Tr$ manually;
Model = SVM;
$M2$ = Model.Train($Tr, Fs_{Tr}$)
For each $te_\mathrm{i}$ in $Te$
        $M2$.Classfy($te_\mathrm{i}$)
        Label $te_\mathrm{i}$
End for
Output
```

C. Reviews Affective Computing

The forum topic posts comments' emotions are mostly positive or neutral. For example, there are nearly 950 comments' emotion are positive or neutral, only 50 comments are negative. The imbalanced data will affect the accuracy of the classification. This paper proposes an improved imbalanced sentiment data classification based on SVM. As shown in Figure II.

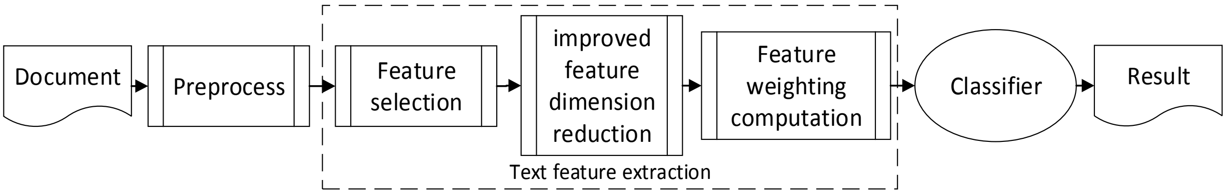

FIGURE II. SENTIMENT CLASSIFICATION PROCESS.

This study uses Synthetic Minority Over-Sampling Technique algorithm [10] to expand the minority class, and propose an improved Chi-square Statistic feature selection method to solve the imbalanced data problem. The calculation formula of CHI score is as shown in (1).

$$\chi^2(t,c) = \begin{cases} \frac{N(AD-BC)^2}{(A+C)(B+D)(A+B)(C+D)} \times \ln\frac{M}{m_t}, AD-BC > 0\ and\ \alpha < 0.5 \\ 0, AD-BC < 0\ \ and\ \alpha < 0.5 \\ \alpha = \frac{A}{A+D} \end{cases} \quad (1)$$

Here, N represents the number of training dataset's documents, A is the number of documents which belong to class c and contain feature t, B is the number of documents which don't belong to class c s but contain feature t, C is the number of documents which belong to class c but do not contain feature t, D is the number of documents which do not belong to class c and do not contain feature t either, α is the choice tendency factor for small sample class, M is the number of all categories, m_t is the number of category which contain feature t.

This improved method can effectively keep features which in the small category, and the identify factor $\ln\frac{M}{m_t}$ can promote the weights of feature which can do larger contribution to the small category classification.

D. Social Network Leader-PageRank Algorithm

PageRank considers the structure of the social network, but it ignores the social interaction properties. So this paper proposes a Leader-PageRank algorithm, based on the combination of sentiment analysis and weighted PageRank algorithm [11].

Based on prospect theory, the influence of the negative comments are much bigger than positive comments. Thus negative comments emotional weight should be set higher than that of positive emotional weight, which can better reflect the real influence of social network interactions. This study sets the value of a single negative comment to -2 points, neutral comment to 0.5 point and positive comment to 1 point. And then get a sentiment matrix E (2).

$$E = \begin{bmatrix} E_{11} & \cdots & E_{1n} \\ \vdots & \ddots & \vdots \\ E_{\mathrm{m}1} & \cdots & E_{mn} \end{bmatrix}, E_{ij} = \frac{\sum e_{ij}}{n_{ij}} \quad (2)$$

Here, $\sum e_{ij}$ represents the sum of the emotional value of all comments which are user j to user i; n_{ij} is the number of comments which are user j to user i, E_{ij} represents the emotional tendencies which is user j to user i.

Equation (3) is the algorithm comprehensive edge weights.

$$W_{ij} = \frac{\sum e_{ij}}{n_{ij}} + F_{ij} \quad (3)$$

Here, W_{ij} is the comprehensive weight which is user j to user i, the range of W_{ij} is [2,-2], F_{ij} represents whether user j follow user i.

Then the Leader-PageRank algorithm formula (4) shown below.

$$\mathrm{LPR}(i) = \frac{1-d}{N} + d\sum_{j\in R(i)} LPR(j)\frac{W_{ij}}{\sum_{k\in T(j)}|W_{kj}|} \quad (4)$$

LPR(i) is the value of the user i Leader-PageRank; d is damping factor; this study sets d to 0.85; N is the number of users; $R(i)$ is a set which is contain all users that focus user i; $\sum_{k\in T(j)}|W_{kj}|$ is the sum of absolute value of edge weight which is come from user j. Ultimately, we can get the effective ranking of opinion leaders by using this equation.

IV. EXPERIMENTS

A. Data Sets

A set of data are used in this study. This web data are come from a Chinese automotive forum by using web spider. The information of dataset includes user's information, user focus situation, topic data and comment data. The data consist of 5871 user information, 36772 topics and 171193 comments in total.

B. Results

After topic classification, we get 4987 vehicle-related discussion topics and 31914 related comments. Then we calculate the comment sentiment value by using different classifiers. The results are shown in Table I.

TABLE I. DIFFERENT CLASSIFICATION RESULTS OF THE ALGORITHM

Classifier	Improved NB			Improved ME			Traditional SVM			Improved SVM		
Sentiment	P Positive	N Neutral	Neg Negative	P	N	Neg	P	N	Neg	P	N	Neg
Precision (p)	0.77	0.81	0.93	0.89	0.75	0.92	0.89	0.73	0.82	0.91	0.78	0.91
Recall(r)	0.88	0.80	0.70	0.76	0.91	0.80	0.77	0.92	0.35	0.79	0.91	0.85
F-measure($\frac{2pr}{p+r}$)	0.82	0.80	0.80	0.82	0.82	0.86	0.82	0.81	0.49	0.85	0.84	0.88
Accuracy	0.808			0.826			0.795			0.850		

The Table 1 shows the improved imbalanced sentiment classification based on SVM result is better than other algorithms. And this paper uses the trained model to classify all comment samples and calculated the emotional matrix E (2).

We get an edge weight matrix W which combine the emotional matrix E and focus matrix F. Then we calculate the users' leader score by using Social Network Leader-PageRank Algorithm. The result is showed in Table II.

TABLE II. COMPARED WITH TRADITIONAL PAGERANK AND IMPROVED LEADER-PAGERANK

	PageRank				Leader-PageRank			
	NO	Follow	Fan	Score	NO	Follow	Fan	Score
1	2759	268	290	33.75373	2759	268	290	34.10796
2	2758	297	288	22.48873	2768	286	171	25.62410
3	2760	145	246	22.30012	2761	187	231	23.14146
4	2761	187	231	21.67382	2760	145	246	23.02614
5	2768	286	171	21.46459	2758	297	288	21.81042
6	2772	68	131	18.51679	2817	7	73	20.51843
7	2794	96	95	16.93851	2772	68	131	20.36013
8	2766	33	172	16.10599	2766	33	172	19.99249
9	2762	316	235	15.83468	2788	47	98	19.29265
10	2763	382	218	15.51417	2781	100	121	18.49039

The results of the two algorithms are quite different. According the raw data, we can find the follow and focus are not the only factors will affect the result of opinion leader score. Some users get lower score than the traditional method because their topics are less concern the automobile information or their comments get much opposition. User No.2817 only has 73 fans, which ranked 53 in the traditional method, but we find this user has many automobile-related topics and get a number of approval. So he get ranked 6 in the Leader-PageRank method. In contrast, due to a large amount of users against its opinion, user No.2763 drop to rank No.21 in the improved method. Due to considering the user's emotional factors, the result of this paper improved method is more reasonable.

V. CONCLUSION

This paper proposes an improved method about opinion leaders of social networks recognition. We identify the topic which contain professional knowledge, construct a comprehensive attention and emotion weight matrix, and propose a Leader-PageRank base on the comprehensive weighting matrix to calculate the leader score. Compared with previous study, the proposed method can recognize opinion leaders which are influential in certain professional fields, and this method consider user emotional interactions situation in social networks which help to identify positive and active opinion leaders.

ACKNOWLEDGMENT

This work is partially supported by the Natural Science Foundation of China [71371144, 71601119, 71601082] and Shanghai Philosophy and Social Science Planning Projects [2013BGL004].

REFERENCES

[1] Rice, R. E., Z. Wu, L. Li, R. Detels, and M. J. Rotheramborus, Reducing STD/HIV stigmatizing attitudes through community popular opinion leaders in Chinese markets, Human Communication Research, 2012, v. 38, p. 379-405.

[2] Cho, Y., J. Hwang, and D. Lee, Identification of effective opinion leaders in the diffusion of technological innovation: A social network approach, Technological Forecasting & Social Change, 2012, v. 79, p. 97‑106.

[3] Sismondo, S., Key Opinion Leaders and the Corruption of Medical Knowledge: What the Sunshine Act Will and Won't Cast Light On, Journal of Law Medicine & Ethics, 2013, v. 41, p. 635‑643.

[4] Yoganarasimhan, H., Impact of social network structure on content propagation: A study using YouTube data, Quantitative Marketing and Economics, 2012, v. 10, p. 111-150.

[5] Zhang, W., H. He, and B. Cao, Identifying and evaluating the internet opinion leader community based on k -clique clustering, Neural Computing and Applications, 2014, v. 8, p. 595-602.

[6] Bakshy, E., J. M. Hofman, W. A. Mason, and D. J. Watts, Everyone's an influencer: quantifying influence on twitter, ACM International Conference on Web Search and Data Mining, 2011, p. 65-74.

[7] Pang, B., L. Lee, and S. Vaithyanathan, Thumbs up?: sentiment classification using machine learning techniques, Acl-02 Conference on Empirical Methods in Natural Language Processing, 2002, p. 79--86.

[8] O'Keefe, T., I. Koprinska, and T. O'Keefe, Feature Selection and Weighting Methods in Sentiment Analysis, Adcs, 2009, v. 24, p. 181.

[9] Tan, C., Short text classification based on LDA and SVM, International Journal of Applied Mathematics & Statistics, 2013, v. 51, p. 205-214.

[10] Ramentol, E., Y. Caballero, R. Bello, and F. Herrera, SMOTE-RS B * : a hybrid preprocessing approach based on oversampling and undersampling for high imbalanced data-sets using SMOTE and rough sets theory, Knowledge and Information Systems, 2012, v. 33, p. 245-265.

[11] Jain, A., R. Sharma, G. Dixit, and V. Tomar, Page Ranking Algorithms in Web Mining, Limitations of Existing Methods and a New Method for Indexing Web Pages, International Conference on Communication Systems and Network Technologies, 2013, p. 640-645.