Sensors & Transducers, Vol. 159, Issue 11, November 2013, pp. 374-378

www.sensorsportal.com

Sensors & Transducers

http://www.sensorsportal.com

The Study of Sensors Market Trends Analysis Based on Social Media

[1] Shianghau Wu, [2] Jiannjong Guo

[1] Faculty of Management and Administration, Macau University of Science and Technology, Avenida Wai Long, Taipa, Macau, China

[2] Graduate Institute of Mainland China Studies, Tamkang University, No.151 Xuefu Road, 251,New Taipei City, Taiwan

[1] Tel.: (853)88972399, fax: (853)28823281

E-mail: shwu@must.edu.mo

Received: 30 October 2013 /Accepted: 20 November 2013 /Published: 30 November 2013

Abstract: The study aimed at analyzing the sensors related tweets on twitter in order to comprehend the market trend. The contribution of the study included the following two points. First, the study used the text mining method in order to explore the content of sensors related tweets. Second, the study applied the classification analysis to explore the relationship of the keywords. *Copyright © 2013 IFSA.*

Keywords: Sensors, Twitter, Random forests, AdaBoost algorithm, Text mining, Classification.

1. Introduction

In this study, the author applied to the text mining analysis in the beginning. The study analyzed the sensors related tweets from twitter in order to get keywords and grasp the trend of sensors market trend.

The rest of the paper was organized as follows. First, the study began with the introduction to text mining method. Second, the overall research design was outlined, the research sites were described and data collection methods were exposed. Third, the text mining results were then presented. Fourth, in order to comprehend the relationship between keywords of the policy addresses, the study applied the classification analysis to the text mining results. The paper concluded with implications and future research avenues.

2. Methodology

In the beginning, the study purposed to use the text mining method to analyze the keywords of the sensors related tweets. Text mining is one of the data mining methods, which learn from samples of past experience. In the text mining method, the text will be processed and transformed into a numerical representation. The text mining method is widely applied to information management on websites, biological data and customer relationship management [1].

2.1. Research Design

The research steps were as follows,

The study used the "tm" (text mining) and "twitteR" packages of the R language to explore the keywords of sensors related tweets from one of the famous social media twitter (http://twitter.com) according to the keywords frequencies.

The study got 220 sensors related tweets and found the keywords of sensors related tweets were'future','internet','track','amp', 'technology','data', 'diabetes', hcare','iot','iphone','GIS', 'infrared', 'weightless', 'techzone', 'wearable tech', 'digikey', 'innovation', 'webforms','blood'and 'servicesphere'. The goal was to make the classification of keywords

and attempted to find the market trend of sensors on the twitter.

Step 2: The study categorized the first 110 keywords' *idf* (inverse document frequencies) data and categorized as "Type 0", and the following 110 keywords data as "Type 1".

Then the study used different algorithms to make the classification analysis in order to comprehend the relationship among keywords and the importance ranking of keywords.

2.2. Random Forests Classification Analysis

The study also applied the random forests classification analysis to explore the relationship of Type 0 and Type 1 data and the importance of keywords. The random forests classification included the following steps [2, 3],

Step (1): Draw the n_{tree} bootstrap samples from the original data.

Step (2): For each of the bootstrap samples grow an unpruned classification or regression tree, with the following modification: at each node, rather than choosing the best split among all predictors, randomly sample m_{try} of the predictors and choose the best split from among those variables.

Step (3): Predict new data by aggregating the predictions of the n_{tree} trees (i.e., majority votes for classification, average for regression).

The study categorized the keywords *idf* data and made the first 110 tweets keywords data as "Type 0", the following 110 tweets keywords data as "Type 1". The number of trees was set as 500, and the number of variables tried at each split was set as 4. The "rattle" package of the R software randomly chose 33 validated keywords data as the test data (18 Type 0 data, 15 Type 1 data) and 187 keywords data as the training data. The error matrix of the random forests model for test data is shown in Table 1.

Table 1. Error Matrix of the Random Forests Model.

Observed	Predicted		
	Type No.0	***Type No.1***	***Percentage Correct***
Type No. 0	17	1	94.44
Type No. 1	10	5	33.33
Overall Error Percentage	33.33 %		

The study also used the ROC (Receiver Operating Characteristic) curve to determine whether the model is the suitable model. The ROC curve plots the true positive rate against the false positive rate. The method is to consider the square measures of areas under the ROC curves. If the square measure approaches to 0.5, it would be the less corresponding model. If the square measure equals to 1, it would bet the model with perfect accuracy. According to the calculation, the square measure of the area under the ROC curve was 0.7556. The ROC curve of the random forests model in the study is shown in Fig. 1.

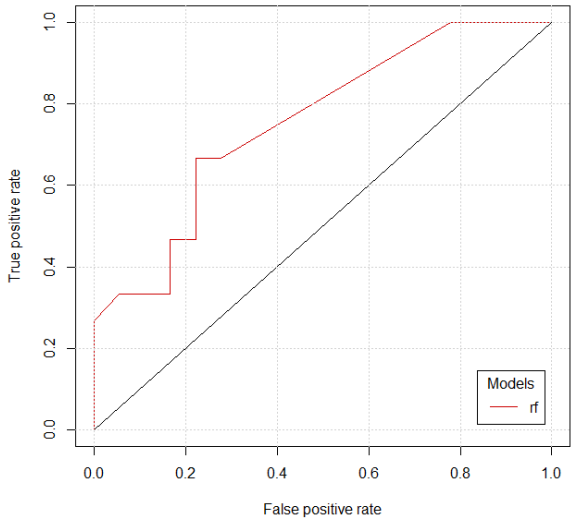

Fig. 1. The ROC curve of the Random Forests model.

The random forests model calculated the variable importance, mean decrease accuracy and mean decrease Gini of keywords which were listed as Table 2 and Fig. 2.

Table 2. Valuable Importance, Mean Decrease Accuracy and Mean Decrease Gini in Random Forests Model.

Variables	Valuable Importance (Type 0 Data)	Valuable Importance (Type 1 Data)	Mean Decrease Accuracy	Mean Decrease Gini
technology	15.85	21.71	21.05	2.51
webforms	11.66	9.36	11.56	0.92
innovation	9.58	11.19	11.48	0.89
track	3.56	9.82	8.50	1.06
diabetes	6.31	6.92	8.30	0.69
blood	7.61	7.00	8.29	0.45
infrared	5.94	7.76	7.64	0.31
hcare	6.01	2.98	6.12	0.58
servicesphere	8.37	1.33	5.34	0.58
GIS	1.92	5.07	4.96	0.59
future	3.83	3.27	4.51	0.90
weightless	2.95	4.17	4.36	0.63
iphone	-0.33	4.01	2.18	0.67
internet	-4.30	4.24	0.12	0.61
iot	-0.14	-0.70	-0.48	0.59
data	-4.43	-0.23	-2.96	0.42
Wearable tech	-4.43	-1.26	-3.14	0.30
digikey	-4.38	-2.11	-4.16	0.22
amp	-8.22	-0.79	-5.83	0.38
techzone	-8.56	-4.97	-8.04	0.32

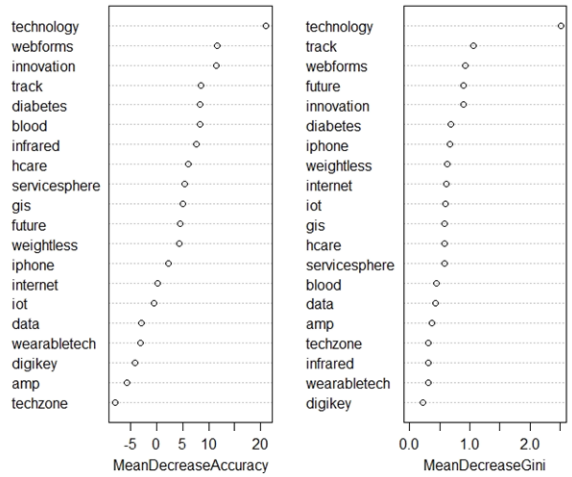

Fig. 2. Mean Decrease Accuracy and Mean Decrease Gini of Random Forests Model.

2.3. AdaBoost Algorithm Classification Analysis

AdaBoost model is a machine learning algorithm which builds a strong classifier from a small set of efficient but weak classifiers. The idea is to choose the weak classifiers in such a way that when combined they perform much better. In the result, the final strong classifier builds a model that is able to predict the class of a new observation given a data set [4, 5]. Viola and Jones (2001) also developed the AdaBoost algorithm further to boost the classification performance by combining collections of weak classifiers to form a stronger classifier. In the beginning, a set of weak classifiers are chosen with the lowest classification error. Then the sequence of machine learning problems is solved and the final strong classifier which takes a weighted combination of the weak classifiers is determined. The final strong classifier determines the optimal threshold classification function for each feature [6].

The general procedure of AdaBoost algorithm is shown as Fig. 2 [7].

Input: Data set $D = \{(x_1, y_1), (x_2, y_2), \cdots, (x_m, y_m)\}$;
Base learning algorithm $\mathcal{L}$;
Number of learning rounds T.

Process:

1. $\mathcal{D}_1(i) = 1/m$. % Initialize the weight distribution
2. for $t = 1, \cdots, T$:
3. $h_t = \mathcal{L}(D, \mathcal{D}_t)$; % Train a learner h_t from D using distribution $\mathcal{D}_t$
4. $\epsilon_t = \Pr_{\boldsymbol{x} \sim \mathcal{D}_t, y} I[h_t(\boldsymbol{x}) \neq y]$; % Measure the error of h_t
5. if $\epsilon_t > 0.5$ then break
6. $\alpha_t = \frac{1}{2} \ln\left(\frac{1-\epsilon_t}{\epsilon_t}\right)$; % Determine the weight of h_t
7. $\mathcal{D}_{t+1}(i) = \frac{\mathcal{D}_t(i)}{Z_t} \times \begin{cases} \exp(-\alpha_t) & \text{if } h_t(\boldsymbol{x}_i) = y_i \\ \exp(\alpha_t) & \text{if } h_t(\boldsymbol{x}_i) \neq y_i \end{cases}$
$= \frac{\mathcal{D}_t(i) \exp(-\alpha_t y_i h_t(\boldsymbol{x}_i))}{Z_t}$ % Update the distribution, where
% Z_t is a normalization factor which
% enables $\mathcal{D}_{t+1}$ to be a distribution
8. end

Output: $H(x) = \text{sign}\left(\sum_{t=1}^{T} \alpha_t h_t(x)\right)$

Fig. 2. The AdaBoost Algorithm.

The study also applied the AdaBoost classification analysis to make the classification analysis. The "rattle" package of the R software randomly chose 33 data as the test data (18 Type 0 data, 15 Type 1 data) and 187 keywords data as the training data. The maximized depth was set as 30, the minimum split was set as 20 and the iterations were set as 50. The error matrix of the AdaBoost model for test data is as Table 3.

According to the calculation, the square measure of the area under the ROC curve was 0.6630. The ROC curve of the AdaBoost model classification is shown in Fig. 4.

Table 3. Error Matrix of the AdaBoost Model.

Observed	**Predicted**		
	Type No.0	***Type No.1***	***Percentage Correct***
Type No. 0	15	3	83.33
Type No. 1	10	5	33.33
Overall Error Percentage	39.39 %		

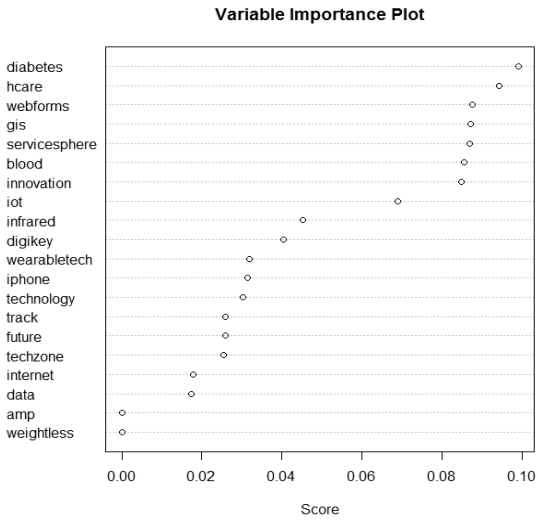

Fig. 3. Mean Decrease Accuracy and Mean Decrease Gini of AdaBoost Model.

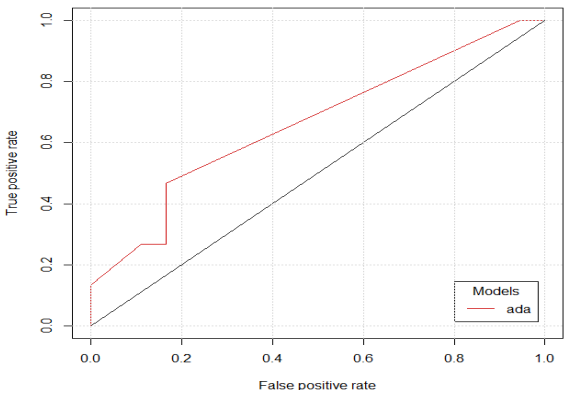

Fig. 4. The ROC curve of the AdaBoost model.

2.4. Decision Tree Classification Analysis

Decision tree analysis is useful for logical induction in the data mining process. Decision tree induction is the learning of decision trees from class-labeled training tuples. A decision tree is a flowchart-like tree structure, where each internal node represents a test on an attribute, each branch represents an outcome of the test, and each leaf node (terminal node) holds a class label. The topmost node is the root node [8]. The "rattle" package of the R software randomly chose 33 data as the test data (18 Type 0 data, 15 Type 1 data) and 187 keywords data as the training data. The configuration parameters is originally set by the Rattle 2.6.26 software, including min. split=20, max. depth=30 and min. bucket=7. The error matrix of the AdaBoost model for test data was as Table 4. And the decision tree model classification was shown as Fig. 5.

Table 4. Error Matrix of the Decision Tree Model.

Observed	Predicted		
	Type No.0	*Type No.1*	*Percentage Correct*
Type No. 0	16	2	88.89
Type No. 1	11	4	26.67
Overall Error Percentage	39.39 %		

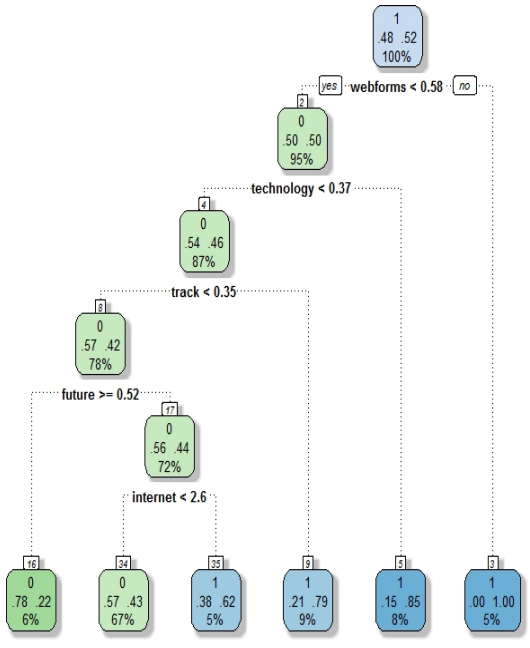

Fig. 5. Decision Tree of the Classification.

According to the calculation, the square measure of the area under the ROC curve was 0.5981. The ROC curve of the decision tree model in the study is shown in Fig. 6.

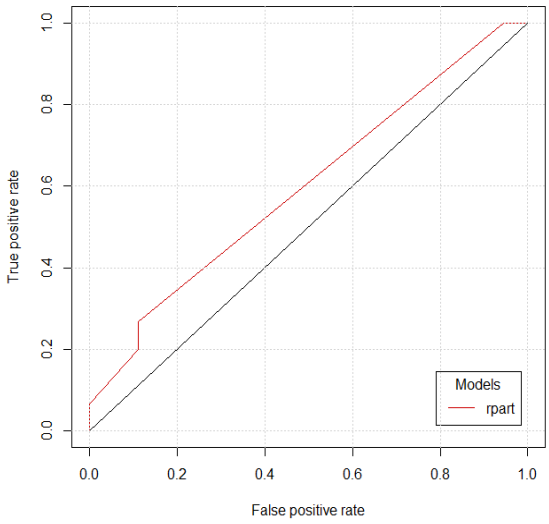

Fig. 6. The ROC curve of the Decision Tree model.

3. Discussion

In the beginning, the study applied the text mining method to get the keywords from sensors related tweets, and then used the random forests model, the AdaBoost Model, the decision tree model to analyze the classification results and major keywords in the classification process. The major results are listed below:

(i) The study found the top five major keywords in the random forests model classification were technology, webforms, innovation, track and diabetes. The random forests model had the best classification performance with the lowest error percentage and the largest square measure of the area under the ROC curve.

(ii) n the AdaBoost classification results, the study found the top five keywords were diabetes, health care (hcare), webforms, GIS(Geographic Information System) and serviceshpere. The classification performance of the AdaBoost model was the second best in three models according to the overall error percentage and the square measure of the area under the ROC curve.

(iii) As for the Decision Tree model, the top four terminal node were webforms, technology, track and future, while the model had the worst performance according to the overall error percentage and the square measure of the area under the ROC curve.

4. Conclusions

The contributions of the study were as follows. First, the study developed a new literature survey method to explore the sensors related tweets to comprehend the market trend of sensors on the social media. From the study, the study found the random forests classification had the best performance in three models of classification. The study also found

the major keywords in sensors related tweets including the realm of web technologies (such as technology, webforms, innovation, track) and health issues (such as health care and diabetes). It offers more insights on further research.

Acknowledgements

The authors were gratitude for the sponsorship of Faculty Research Grant funded by the Macau University of Science and Technology.

References

[1]. S. Weiss, N. Indukhya, T. Zhang, and F. Damerau, Text Mining: Predictive Method for Analyzing Unstructured Information, *Springer*, 2005.

[2]. A. Liaw and M. Wiener, Classification and Regression from Random Forest, *The R Journal*, Vol. 2, No. 3, 2002, pp. 18-22.

[3]. L. Breiman, Random Forests, *Machine Learning*, Vol. 45, No. 1, 2001, pp. 5-32.

[4]. Y. Freund, R. E. Schapire, A short introduction to boosting, *Journal of Japanese Society for Artificial Intelligence*, Vol. 14, No. 5, 1999, pp. 771-780.

[5]. Y. Shin, D. W. Kim, S. W. Yang, H. H. Cho, K. I. Kang, Decision Support Model using the AdaBoost Algorithm to Select Formwork Systems in High-Rise Building Construction, in *Proceedings of the 25th International Symposium on Automation and Robotics in Construction*, 2008, pp. 644-649.

[6]. P. Viola, M. Jones, Rapid Object Detection using a Boosted Cascade of Simple Features, in *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, 2001, pp. 511-518.

[7]. X. Wu, V. Kumar (eds.), The Top Ten Algorithms in Data Mining, *Taylor & Francis*, 2009.

[8]. J. Han, M. Kamber, Data Mining: Concepts and Techniques, *Elsevier*, 2006.
