Detecting Fake Reviews Utilizing Semantic and Emotion Model

Conference Paper · July 2016

DOI: 10.1109/ICISCE.2016.77

3 authors, including:

Yuejun Li
Shandong Jianzhu University
7 PUBLICATIONS **17** CITATIONS

SEE PROFILE

Detecting Fake Reviews Utilizing Semantic and Emotion Model

Yuejun Li, Xiao Feng, Shuwu Zhang
Institute of Automation
Chinese Academy of Sciences
Beijing, China
273253612@qq.com, xiao.feng@ia.ac.cn,
shuwu.zhang@ia.ac.cn

Yuejun Li
School of Computer Science & Technology
Shandong Jianzhu University
Jinan, China
273253612@qq.com

***Abstract*—As people are spending more time to shop and view reviews on line, some reviewer write fake reviews to earn credit and to promote (demote) the sales of product and stores. Detecting fake reviews and spammers becomes more important when the spamming behavior is becoming damaging. This paper proposes three types of new features which include review density, semantic and emotion and gives the model and algorithm to construct each feature. Experiments show that the proposed model, algorithm and features are efficient in fake review detection task than traditional method based on content, reviewer info and behavior.**

Keywords—Fake Review Detection; Spammer Detection; Semantic Model; Emotion Model

I. INTRODUCTION

Most of the online B2C and professional review web sites provide service of writing reviews of the product, stores and services. Due to the importance of reviews to the product and stores, some fake reviewers emerge quickly to post fake reviews to promote sales or increase credit of the reviewer. Such individuals are called opinion spammers and the activities are called opinion spamming [1,2]. Although the web site carried out some operation to reduce review spamming, it's still a long way to go to prohibit reviewers posting deceptive reviews due to the complexity of detecting fake reviews.

Researches have been done since Jindal [1] first proposed opinion spam. On the whole, the review spam (opinion spam) detecting method can be classified to two types: supervised learning and unsupervised methods. Supervised learning method [1, 3, 4] has relatively good scalability and performance given the right feature and labeling training data. [4,5] utilize review content part-of-speech(POS) ,LIWC text features and language model to find deceptive opinion spam, but did not consider the user behavior which is very useful to detect fake review and reviewer. Unsupervised method has been used to detect group spammers [6, 8] and review burstiness [7] and behavioral footprints [9].

In this paper we propose review density, semantic and emotion related model and feature to identify fake reviews from professional review web site. Experiments show that the model and features we proposed outperform the usually used features which include behavior, reviewer information and content based feature. The main contributions of our work include:

1. A labeled review dataset with fake and non-fake reviews from a professional review site. Although labeling fake reviews is difficult, when given sufficient information of reviewer, review and related store, it becomes easier for human.
2. New features of review density which include category density, store density and time density.
3. Semantic models of review similarity and emotion models of review to get emotion diversity.

II. FEATURES OF DETECTING FAKE REVIEWS

Various features have been used in previous work like the content of the review, the reviewer and the product being reviewed [1]. The most used features can be classified to three types: review behavior related feature, reviewer basic and characteristic information related feature, content related feature. We proposed review density related features to capture the category, store and time character of fake reviews. The feature definition is defined below:

A. User Behavior Diversity Related Feature

- Good Review Ratio (GRR, F1): the ratio of number of reviews that ranks with a relatively high rank (for example 4 star and 5 star rating) divides number of reviews the reviewer posted.
- Bad Review Ratio (BRR, F2): the ratio of number of reviews that ranks with a relatively low level (for example ranks less than 4 star and 5 star rating) divides number of reviews the reviewer posted.
- Reviewer Reviews Ratio (RRR, F3): the ratio of number of reviews that reviewer u posted divides the maximum number of reviews that all reviewers posted.
- Average Review Density (ARD, F4): number of reviews the reviewer posted divides number of days that has at least one review.
- Maximum Review Density (MRD, F5): the maximum of number of reviews the reviewer posted in a day divides number of days that has at least one review.

- Rating Variance (RV, F6): the mean variance of ratings of reviewer.

$$RV(r) = \sqrt{\frac{1}{n}\sum_{j=1}^{5}(pr_j - \overline{pr})^2} \quad (1)$$

pr denotes the proportion of rating in every star rating.

B. Reviewer Characteristic and Basic Information Related Feature

- Active Time Frame (ATF, F7): Number of reviews reviewer posted divides number of Active Time Frame since reviewer registered in the website.
- Ratio of Favorite Review (RFR, F8): ratio of number of favorite stores that reviewer selected divides number of reviewer reviews.
- Ratio of Focus Count (RFC1, F9): ratio of number of stores that reviewer focus divides number of reviewer reviews.
- Ratio of Fans Count (RFC2, F10): ratio of number of fans that reviewer has divides number of reviewer reviews.
- Ratio of Flower Count (RFC3, F11): ratio of number of flower that reviewer get from other reviewer divides number of reviewer reviews.

C. Content Related Features

- Review content similarity in a single review (RCS, F12) which equals number of different words divides length of review.
- Review content similarity between review and other reviews of reviewer (RCS2, F13).

III. REVIEW DENSITY RELATED FEATURES

Fake reviewers often post deceptive reviews with special density character in category, store and time. We propose three review density related feature which are defined below.

A. Category Density

Reviewers usually post different reviews of stores in different categories like hotel, restaurant, and car service etc., because reviewer's interest may change from one category of stores to other category of stores over a period of time. 'Professional' fake reviewers sometimes post lots of reviews in a single category to accomplish his task and make a profit. So we propose category density of review r to represent this feature (F14).

$$categoryDensity(r) = \frac{Num\,Review(c)_r}{|R_u|} \quad (2)$$

Where $Num\,Review(c)_r$ denotes number of reviews which are in the same category as the review r of this reviewer u. And $|R|_u$ is number of reviews reviewer u posted.

B. Store Density

Many fake reviewers comment several reviews over the same store or product to enhance the influence of the credit of the store. So detecting this behavior will be helpful to find fraudulent reviews which are focused on the same store. The store density (F15) is defined as follows:

$$sameStoreDensity(r) = \frac{num\,Review(s)_r}{|R_u|} \quad (3)$$

where $num\,Review(s)_r$ denotes number of reviews which refer to the same store as the review r of this reviewer u. And $|R|_u$ is number of reviews reviewer u posted.

C. Time Density

If the reviewer posts reviews very frequently (almost post reviews every day), the reviewer is likely to be a spammer or fake reviewer. Time density (F16) can be used to formulate the frequency of reviews reviewer post.

$$timeDensity(r) = \frac{numDays(u)}{|R_u|} \quad (4)$$

where numDays(u) denotes the number of days the reviewer u involve.

IV. SEMANTIC SIMILARITY OF REVIEWS

In previous section, traditional content based features (F13) computes the similarity of reviews using the cosine similarity between words of reviews. But sophisticated spammers often post reviews using different expression style of distinct words and sentences to avoid being regarded as spammer. We propose a semantic method to compute review similarity utilizing the word2Vec tool provided by Google.

Word2vec models can be used to map each word to a vector of typically several hundred elements. We generate the semantic representation of review by use of word embeddings and compute the similarity of review using cosine function. The function is defined below.

$$semanticSimilarity(r_{u,k}, R_u) = \frac{1}{|R_u| - 1} \sum_{\substack{i=1 \\ i \neq k}}^{|R_u|} cosine(wordVector(r_{u,k}), wordVector(r_{u,i})) \quad (5)$$

where $wordVector(r_{u,k})$, $wordVector(r_{u,i})$ is the vector representation of review k and i of reviewer u. We use the word2vec tool to compute the vector representation of each word of review, and then combine each vector of the word of the review together to formulate the vector representation of sentences of reviews. The representation of review r of vector style can be denoted as:

$$wordVector(r) = (wordVector(r))_{k=1}^{|K|} = (\sum_{i=1}^{|r|} vector(word_i)_k)_{k=1}^{|K|} \quad (6)$$

The vector of review r is defined as a vector of dimensions of $|K|$ which is also the dimensions of each word vector representation. $|r|$ is the number of words in review r in which dimensions of each word i is summarized as a whole to represent the review representation.

The algorithm of computation of semantic similarity between reviews can be described as follows.

TABLE I. ALGORITHM OF COMPUTATION OF SEMANTIC SIMILARITY BETWEEN REVIEWS

```
Input: review r1, r2
Output: similarity of r1 and r2
Begin
    R1_word[] = word segment of review r1;
    lengthA = length(R1_word);
    R2_word[] = word segment of review r2;
    lengthB = length(R2_word);
    For every dimension k in vector space
      For every word in R1_word[i]
        R1_word_vector[i] = word2Vec(R1_word[i]);
        R1_review_vector[k] += R1_word_vector[i][k];
      END
      R1_review_vector = R1_review_vector[k] divide lengthA;
      For every word in R2_word[j]
        R2_word_vector[j] = word2Vec(R1_word[j]);
        R2_review_vector[k] += R2_word_vector[j][k];
      R2_review_vector = R2_review_vector[k] divide lengthB;
      END
    END
    similarity(r1,r2) = cosin(R1_review_vector, R2_review_vector);
    Return similarity(r1,r2);
End
```

V. EMOTION MODELING

Experienced spammers often post reviews that look like real reviews but can't hide their intention of spamming. If the spammer is hired to promote the sales of the product or to increase the credit of stores, the emotion that is hidden in the content of the review is likely to be positive because of the economic reason. As we observe in the corpus of review set, reviews of balanced emotion are less likely to be faked. We first model the emotion of the review and then compute the diversity of the emotion in reviews.

$$\text{Index}_{\text{positive}}(r) = \frac{\text{Num}(\text{positiveWord})_r}{\text{wordNum}(r)} \quad (7)$$

Where Num(positiveWord) denotes number of positive word in review r which the negation word situation is considered. wordNum(r) equals to the word number of review r except irrelevant stop words. Other two scalars are defined as follows similarly:

$$\text{Index}_{\text{negative}}(r) = \frac{\text{Num}(\text{negativeWord})_r}{\text{wordNum}(r)} \quad (8)$$

$$\text{Index}_{\text{neutral}}(r) = \frac{\text{Num}(\text{neutralWord})_r}{\text{wordNum}(r)} \quad (9)$$

The negative index of review r denotes the level of negative emotion and neutral index implies the neutral emotion of review r.

In order to measure the diversity of emotion in review, we utilize the mean square deviation function (F18) to compute emotion diversity:

$$\text{emotionDiversity}(r) = \sqrt{\frac{1}{3}\sum_{k=1}^{3}(\text{Index}_k(r) - \overline{\text{Index}_k(r)})^2} \quad (10)$$

VI. EXPERIMENTS AND EVALUATION

A. Data Set and Human Evaluation

Before we evaluate the proposed feature and model, we first collect the review dataset using robot from www.dianping.com which is the most famous review web site in China which is just like yelp.com. The datasets comprises of 21,255 reviews, 504 reviewers and 14,187 stores.

As the evaluation process is taken in supervised setting, we need to annotate the reviews manually which are fake or not. The annotation process needs lots of labor especially when nothing context information is provided but only the content of the reviews is provided. To solve the problem we provide human experts enough contexts of the reviews which include all reviews of the reviewer posted, posted time, related store, reviewer basic information and store basic information. Human experts select 3600 reviews to be annotated because of limit time to review every item in the datasets.

B. Experiment and evaluation

We use three different types of classifier to perform the detection of fake reviews and Use 5 fold cross-validation to train and test. The result is shown below.

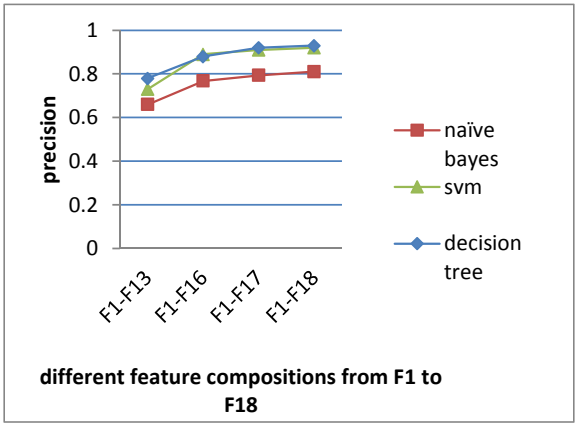

Fig.1 precision of fake review detection using different features and classifiers

F1-F13 denote that the feature F1 to F13 are used in the classify process which include reviewer behavior, reviewer basic information and content similarity related features. F1-F16 contain more review density features we proposed. F1-F17 and F1-F18 add semantic similarity feature and emotion diversity feature we proposed.

TABLE II. PRECISION OF FAKE REVIEW DETECTION USING DIFFERENT FEATURES AND CLASSIFIERS

classifier / features	Naïve Bayes	SVM	Decision Tree
F1-F13 (basic behavior)	0.66	0.73	0.78
F1-F16 (above plus review density)	0.76	0.89	0.88
F1-F17 (above plus semantic)	0.79	0.91	0.92
F1-F18 (above plus emotion)	0.81	0.92	0.93

From the table 2 we can see that when only the features F1-F13 are imported, the precision is rather low and less than 0.8 whatever any of the three classifiers is used. While when the review density, semantic and emotion related features are used, the precision of every classifier jump high to more than 0.9. Naïve Bayes performs weak because of its simplicity. SVM and decision tree get quite good performance with precision of 0.92 and 0.93 when additional 3 features we proposed are incorporated. This demonstrates that review density related features have strong ability to promote precision in fake review detection and semantic and emotion related features have additional promotion during the detection process.

VII. CONCLUSION AND FUTURE WORK

This paper proposes three types of new features which include review density, semantic and emotion and gives the model and algorithm to construct each feature. Experiments show that the proposed model, algorithm and features are efficient in fake review detection task.

Future work includes collecting abundant review data from other review web sites, computer assisted labeling of reviews to reduce the workload of human experts, more efficient model of detecting the relationship of reviews, reviewers and stores

ACKNOWLEDGMENT

The paper is supported by National Key Technology R&D Program of China under Grant NO.2015BAH49F01 and Technology Plan of Beijing under Grant No. D161100005216001.

REFERENCES

[1] N. Jindal, B. Liu. "Opinion spam and analysis." International Conference on Web Search and Data Mining ACM, 2008, pp. 219--230.

[2] A. Mukherjee, A. Kumar, B. Liu, et al. “Spotting opinion spammers using behavioral footprints”, ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2013:632-640.

[3] S. Feng, R. Banerjee, Y. Choi, “Syntactic Stylometry for Deception Detection”, ACL (2011), pp. 171-175.

[4] M. Ott, Y. Choi, C. Cardie, J.T. Hancock, “Finding Deceptive Opinion Spam by Any Stretch of the Imagination”, ACL (2011), 309–319.

[5] RYK. Lau, SY. Liao, RCW. Kwok, K. Xu, Y. Xia, Y. Li, “Text mining and probabilistic language modeling for online review spam detection”, ACM Trans. on Management Information Systems (TMIS), 2011,2(4):25

[6] A. Mukherjee, B. Liu, N. Glance, “Spotting Fake Reviewer Groups in Consumer Reviews”, International Conference on World Wide Web ACM, 2012, pp. 191-200.

[7] G. Fei, A. Mukherjee, B. Liu, M. Hsu, M. Castellanos, R. Ghosh, R. “Exploiting Burstiness in Reviews for Review Spammer Detection”, ICWSM, 2013.

[8] A. Mukherjee, B. Liu, J. Wang, N. Glance, N. Jindal, “Detecting Group Review Spam”, International Conference on World Wide Web ACM, 2011, pp. 93-94, DOI: 10.1145/1963192.1963240.

[9] A. Mukherjee, A. Kumar, B. Liu, et al, “Spotting opinion spammers using behavioral footprints” ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2013, pp. 632-640.

[10] F. Li, M. Huang, Y. Yang, X. Zhu, “Learning to Identify Review Spam” IJCAI, 2011, pp. 2488–2493.

[11] E.P. Lim, V.A. Nguyen, N. Jindal, B. Liu, H.W. Lauw, “Detecting product review spammers using rating behaviors”, Acm International Conference on Information & Knowledge Management ACM, 2012, pp. 939-948.

[12] T. Mikolov, I. Sutskever, K. Chen, et al, “Distributed Representations of Words and Phrases and their Compositionality”, Advances in Neural Information Processing Systems, 2013,vol. 26, pp. 3111-3119.

[13] B. Liu and L. Zhang. “A Survey of Opinion Mining and Sentiment Analysis”, Jour. Mining Text Data, 2012.

[14] Y. Zhang, G. Lai, M. Zhang, Y. Zhang, Y. Liu, and S. Ma, “Explicit Factor Models for Explainable Recommendation based on Phrase-level Sentiment Analysis”, SIGIR, 2014.