Toward a Machine Learning Based Approach for Assessing the Credibility of Online Comments

Otto K. M. Cheng and Raymond Y. K. Lau
Department of Information Systems, City University of Hong Kong, Hong Kong SAR
Email: raylau@cityu.edu.hk

***Abstract*—Even though many incidents about fake online consumer reviews have been reported, very few studies have been conducted to date to examine the credibility of online consumer comments. One of the reasons is the lack of an effective computational method to deal with the huge number of online comments which are not embedded with explicit features for a spam detection system to separate the untruthful comments (i.e., spam) from the legitimate ones (i.e., ham). To improve the hygiene and the usefulness of online comments, there is a pressing need to develop a robust methodology for an objective and systematic assessment of the quality of online comments. The main contribution of this paper is the design, development, and evaluation of a novel machine learning based methodology for the assessment of the credibility of online comments. Our preliminary experiments show that the proposed quality assessment methodology is more effective than other baseline methods such as a peer-review based quality assessment method.**

***Index Terms*—opinion credibility, opinion analysis, spam detection, SVM, machine learning**

I. Introduction

With the rapid growth of the Social Web, there has been increasingly more user-contributed comments posted to the Internet on a daily basis. These online comments refer to consumer products, financial products, social events, political figures, and so on. These low-cost, massively available online comments provide organizations with unprecedented opportunities to extract valuable business intelligence and market intelligence to strengthen business strategy development. Moreover, the sheer volume of peer-contributed comments can considerably improve consumers' comparison shopping experience. Nevertheless, the widespread sharing and utilization of uncontrolled user-generated online comments such as product reviews has also raised the concerns about the quality and trustworthiness of these items [1]-[3]. New York Times reported a settlement case about shill comments not long ago. The news report revealed that a U.S. based cosmetic surgery company had requested its employees to pretend to be satisfied customers and post glowing comments of its products on its own Web sites and other third party opinion sharing sites. As a matter of fact, firms or individuals have the financial or political motivations to strategically manipulate online comments [1]. Accordingly, there is a pressing need to develop advanced method to detect and identify low-quality comments so that firms or individuals are not misled by these comments.

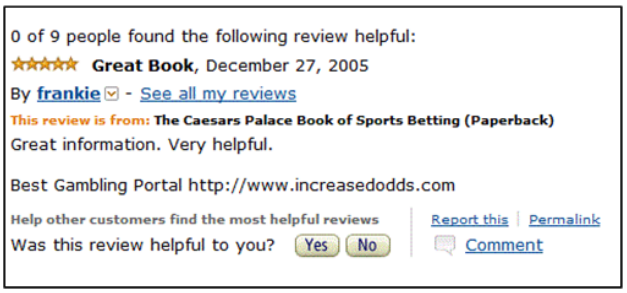

Figure 1. Note how the caption is centered in the column.

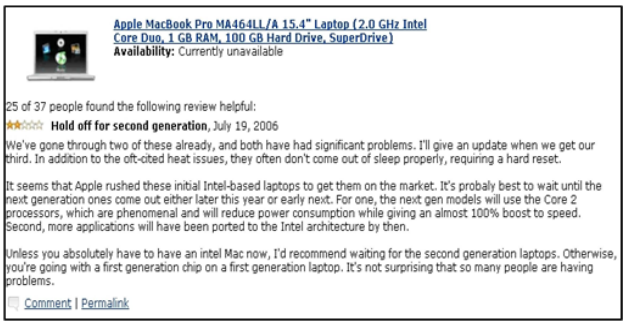

Figure 2. The first example suspicious comment.

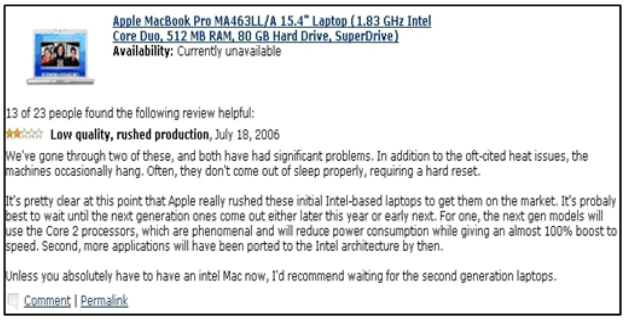

Figure 3. Another example of suspicious comment.

Fig. 1 shows an example of low quality online comment posted to amazon.com. Basically, this short comment (review) is not informative and misleading at all. The title of the comment is about "Great Book", but the contents of the comment are trying to lure readers to traverse to an online gambling site. However, a long

Manuscript received January 20, 2014; revised April 18, 2014.

comment with rich contents is not necessarily with high-quality either. Fig. 2 depicts an example of long product comment but with low-quality (a near-duplicated comment). It is about the product "Apple MacBook Pro MA463LL". As can be observed based on Fig. 2, by examining all the possible features (e.g., occurrence of product features, frequency of sentiment indicators, occurrence of URLs, number of helpful votes, etc.) associated with the comment, it may still be difficult to judge if the comment is with low-quality or not. In fact, this near-duplicated comment attracts 25 out of 37 helpful votes! However, if another product comment depicted in Fig. 3 is presented, a reader or a quality detection program can easily determine that the comment shown in Fig. 2 is a near-duplicated of the comment shown in Fig. 3. In fact, both comments contain low-quality comments since each one is not informative given the presence of another comment.

Based on these real-life examples, it is not difficult to find that a multi-facets approach is required to judge the quality of online comments such as online product comments. For instance, an effective quality metric should take into account both the intrinsic factors of a comment and the relative informativeness of the comment with respect to other existing comments. In order to develop a robust multi-facets quality assessment framework for online comments, we investigate into the theoretical ground on "information quality" in general [4], [5]. There are 16 quality metrics applied to evaluate the quality of information collected from the World Wide Web (Web), namely, "subject", "breadth", "depth", "cohesiveness", "accuracy", "timeliness", "source", "maintenance", "currency", "availability", "authority", "presentation", "credibility", "noisy", "writing style", and "popularity" [5]. Similarly, the dimensions such as "content", "source", "presentation", and "format" are being evaluated in a qualitative study related to the quality assessment of Web-based information. Since not all the aforementioned facets are relevant for the quality assessment of online comments, we will illustrate a novel framework for the quality assessment of opinionated expressions in Section 2.

The main contribution of the research reported in this paper is the design, development, and evaluation of a novel information theory based quality assessment framework for online comments. In particular, our proposed method differs from other existing approaches in that both the intrinsic properties and the external associations of the targeted online comments are taken into account. The practical implication of our research is that the proposed information theory based quality metric can be applied to assess and filter out low-quality online comments. As a result, the effectiveness of opinion mining or opinion retrieval processes can be improved. The direct business implication is that more accurate and timely business intelligence can be extracted from online comments to enhance organizations' business strategy development. The rest of the paper is organized as follows. Section 2 illustrates the framework and the computational method of the proposed information theory based quality assessment of online comments. Section 3 reports the preliminary evaluation of our proposed quality metric and discusses the experimental results. Finally, we offer concluding remarks and discuss future directions of our research work.

II. Information-Based Feature Selection

Quality assessment of online comments is divided into two stages. First, the non-informative duplicated online comments are detected by invoking a duplication detection mechanism [6]. Since there are many existing duplicated content detection methods, this paper will not go into details about this topic and assume that we can use an existing method to carry out this task. Second, a classifier is applied to analyze the intrinsic features of online comments so that low-quality expressions can be filtered out. The set of intrinsic features for quality assessment of online comments are defined according to the dimensions for information assessment [4], [5]. In particular, the dimensions of subject, breadth, depth, timeliness, source, presentation, accuracy, writing style, credibility, and popularity are applied to our study.

A. The Dimension of Accuracy

For the dimension of accuracy, the following elementary features are applied to quantitatively estimate the score for this dimension: normalized appearance frequency of the target entity (e.g., a product) in an online comment (e.g., a product comment) (f1) and normalized appearance frequency of the features (e.g., product features) of the target entity (f2). The normalized frequency of a token t is derived by the fraction $(freq(t))/(|d|)$, where freq(t) is the raw appearance frequency of t and |d| is the length (number of tokens) of the online comment d. The related features of a product comment are extracted from its online product description.

B. The Dimension of Subject

For the dimension of subject, the following elementary features are applied to quantitatively estimate the score of this dimension: percentage of opinionated words in the expression (f3), percentage of positive words in the expression (f4), percentage of negative words in the expression (f5).

C. The Dimension of Breath

For the dimension of breadth, the fraction of the total frequency of all entities mentioned in the online comment d over the frequency of the target entity (e.g., a product) mentioned in d is computed (f6). Entities are identified based on an extended version of the GATE named entity recognition (NER) module [7]. Moreover, the frequency of all positive words and the frequency of all negative words mentioned in d are summed (f7).

D. The Dimension of Depth

For the dimension of depth, the fraction of the total frequency of the target entity mentioned in the online comment d over the frequency of all the entities mentioned in d is computed (f8). Moreover, the percentage of all the related features of the target entity

(e.g., product features) over all the features mentioned in d is applied (f9).

E. The Dimension of Timeliness

For the dimension of timeliness, the ratio of the longest elapsed time of all the online comments related to a target entity (e.g., a product) over the elapsed time of the online comment is computed (f10).

F. The Dimension of Source

For the dimension of source, the normalized rank of the reviewer (f11) and the percentage of expressions written by the writer over all the expressions in the collection D (f12) are computed.

G. The Dimension of Presentation

For the dimension of presentation, the following features are applied: average length of sentences in the expression (f13), average length of paragraphs in the expression (f14), length of the expression (f15).

H. The Dimension of Writing Style

For the dimension of writing style, the following features are used: percentage of valid words in the expression (f16), percentage of alphabetical words in the expression (f17), percentage of numerical tokens in the expression (f18), percentage of nouns (f19), percentage of verbs (f20), percentage of adjectives (f21), percentage of adverbs (f22), percentage of pronoun (f23), percentage of capitalized tokens (f24), and percentage of special characters (f25).

I. The Dimension of Credibility

For the dimension of credibility, the following features are applied: the standard deviation of the rating of the online comment (f26), the difference of the positive words found in the online comment d and the average number of positive words found in comments (f27), and the difference of the negative words found in the online comment d and the average number of negative words found in comments (f28).

J. The Dimension of Popularity

For the dimension of popularity, the normalized helpful votes of the online comment (f29) is computed.

K. The Prediction Models

Support vector machines have been successfully applied to text categorization tasks, and it seems that they outperform probabilistic classifiers such as Naive Bayes [8]. We applied the aforementioned features to a SVM classifier to separate the high-quality expressions from the low-quality expressions. For the two-class problem (e.g., high-quality versus low-quality), the basic principle is to find a hyperplane represented by the weight vector $\vec{\omega}$ (a normal vector perpendicular to the hyperplane), which not only separates the comment vectors into two classes, but also guarantees the separation or margin, is as large as possible. This search corresponds to a constrained optimization problem with the following form:

$$\vec{\omega} = \sum \alpha_i c_i \vec{d}_i, \ \forall \alpha_i \geq 0 \quad (1)$$

$$b = c_i - \vec{\omega}^T \vec{d}_i, \ \forall \vec{d}_i : \alpha_i \neq 0 \quad (2)$$

where each $\vec{d}_i$ represents a labeled opinionated expression, and $c_i \in \{1, -1\}$ denotes a class of this binary classification problem; α_i is the Lagrange multiplier and each nonzero α_i indicates that the corresponding $\vec{d}_i$ is a support vector; b is the intercept of the binary classifier. Based on the parameters such as $\vec{\omega}$, α_i, and b learned from the training data, the large margin classification function $f(\vec{d})$ can then be defined, where the signum function (sgn) is used to determine the class of a given test sample $\vec{d}$ (i.e., an unlabeled opinionated expression). For instance, a positive margin value suggests the high-quality class, and a negative margin value indicates the low-quality class.

$$f(\vec{d}) = \text{sgn}\left(\sum_i \alpha_i c_i K(\vec{d}_i \vec{d}) + b\right) \quad (3)$$

The kernel function of Eq. 3 has the form: $K(\vec{d}_i \vec{d}) = (1 + \vec{d}_i^T \vec{d})^z$. When $z = 1$ is established, a linear kernel is applied. It becomes a quadratic kernel when $z = 2$ is established. For the radial basis kernel, it has the following form: $K(\vec{d}_i \vec{d}) = \exp\left(-\frac{(\vec{d}_i - \vec{d})^2}{2\sigma^2}\right)$. We applied Joachim's SVM package[1] and the Weka package[2] for the implementation of various classifiers.

III. Experiments and Results

Since a benchmark data set for the assessment of the quality of online comments is not available, we first downloaded online comments from amazon.com and then human annotators were invited to examine the candidate sets of high-quality comments and low-quality comments respectively to construct the evaluation data set. Similar approach was adopted by a previous study [9]. However, manually inspecting all the downloaded online comments is extremely difficult and time consuming. We applied several heuristic to automatically construct candidate sets of high-quality and low-quality comments respectively. The heuristic we used include the length of a comment, the percentage of special characters of a comments, the percentage of valid words of a comments, etc. Human annotators then evaluated the candidate sets to confirm the positive and negative cases. Comments from six Amazon product categories were retrieved using the Amazon Web services (October 2009 version). A total of 1,218 high-quality and 1,208 low-quality comments were manually annotated from among the downloaded comments of our evaluation data set.

Given a balanced class distribution in our experimental data set, we simply used accuracy as the performance measure. We examined SVM with a radio basis kernel

[1] http://svmlight.joachims.org/
[2] http://sourceforge.net/projects/weka/files/weka-packages/

(SVM-RB), a SVM with linear kernel (SVM-LN), Naive Bayes (NaiveB), and Neural Network (NeuralN) for quality classes prediction. A ten-fold cross validation was then applied to our experiment. The results of our experiments were shown in Table I. The SVM with radio basis kernel achieved the best accuracy (79%). The accuracy of our method is better than that reported in a previous study [9] even though we did not applied any sophisticated feature extraction and selection methods in our experiment. We performed a follow-up study by isolating subsets of features and found that the dimension of *presentation* is the best individual dimension for the prediction of high-quality comments. This finding is basically consistent with the results of the previous studies [9], [10]. The best three dimensions validated in our preliminary experiment are *presentation*, *writing style*, and *credibility*. In contrast, the dimension of *popularity* is one of the poorest dimensions for the prediction of high-quality comments. Fig. 4 shows the helpful votes distribution of the low-quality "Books" comments from a separate data set. As can been seen, while around 700 low-quality comments attract zero helpful votes, there are similar number of low-quality comments receiving 10 to 100 helpful votes. Therefore, using such a feature to classify high-quality comments is not effective.

TABLE I. PERFORMANCE OF VARIOUS QUALITY PREDICTION METHODS

Category	SVM-RB	SVM-LN	NaiveB	NeuralN
Apparel	792	732	751	789
Baby Items	806	769	758	811
Music	785	755	736	763
Food	782	743	703	779
VHS	812	759	739	806
Books	791	737	711	785
Average	795	753	736	789

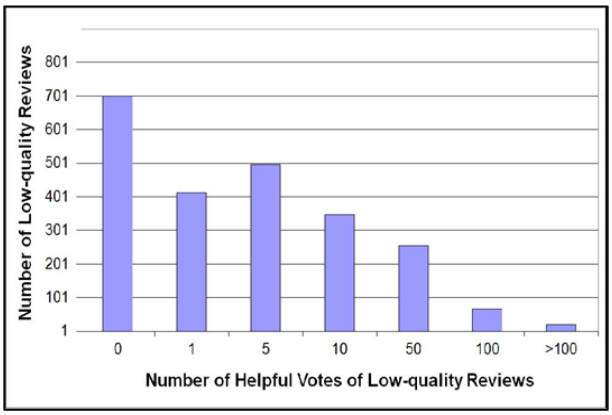

Figure 4. The helpful votes distribution of low-quality comments.

IV. CONCLUSIONS AND FUTURE WORK

Many reports have indicated the quality issues of online comments. The main contribution of our work is the development of a novel information theory based quality framework to detect non-informative, low-quality online comments. In particular, our method does not rely on the user-generated helpful votes to establish the ground-truth of high-quality comments. Our preliminary experimental results show that SVM with a radio basis kernel outperforms other prediction methods under testing. The best three dimensions for quality class prediction are presentation, writing style, and credibility. Future work involves testing the proposed mechanisms based on a larger set of online comments and refines the set of quality dimensions for quality class prediction.

ACKNOWLEDGMENT

The work described in this paper was supported by a grant from City University of Hong Kong (Project No. 7008138) and Hong Kong RGC's GRF Grant (Project No. CityU 145712).

REFERENCES

[1] C. Dellarocas, "Strategic manipulation of internet opinion forums: Implications for consumers and firms," *Management Science*, vol. 52, pp. 1577-1593, 2006.
[2] R. Y. K. Lau, S. Y. Liao, *et al.*, "Text mining and probabilistic language modeling for online review spam detection," *ACM Transactions on Management Information Systems*, vol. 2, no. 4, pp. 1-30, 2011.
[3] G. Wu, D. Greene, and P. Cunningham, "Merging multiple criteria to identify suspicious reviews," in *Proc. Fourth ACM Conference on Recommender Systems*, 2010, pp. 241-244.
[4] S. Y. Rieh, "Judgment of information quality and cognitive authority in the web," *JASIST*, vol. 53, no. 2, pp. 145-161, 2002.
[5] X. Zhu and S. Gauch, "Incorporating quality metrics in centralized/distributed information retrieval on the world wide web," in *Proc. 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2000, pp. 288-295.
[6] C. Lai, K. Xu, R. Lau, Y. Li, and L. Jing, "Toward a language modeling approach for consumer review spam detection," in *Proc. IEEE 7th International Conference on e-Business Engineering*, 2010, pp. 1-8.
[7] D. Maynard, V. Tablan, C. Ursu, H. Cunningham, and Y. Wilks, "Named entity recognition from diverse text types," in *Proc. 2001 Conference on Recent Advances in Natural Language Processing*, 2001, pp. 257-274.
[8] T. Joachims, "Making large-scale SVM learning practical," in *Advances in Kernel Methods - Support Vector Learning*, B. Schölkopf, C. J. C. Burges and A. J. Smola, Eds. Cambridge, Massachusetts: MIT Press, 1999.
[9] J. Liu, Y. Cao, C. Y. Lin, Y. Huang, and M. Zhou, "Low-quality product review detection in opinion summarization," in *Proc. EMNLP-CoNLL*, 2007, pp. 334-342.
[10] Z. Zhang, "Weighing stars: Aggregating online product reviews for intelligent e-commerce applications," *IEEE Intelligent Systems*, vol. 23, no. 5, pp. 42-49, 2008.

Otto K. M. Cheng is a Senior Research Associate of the Department of Information Systems, City University of Hong Kong. His research interests include Text Mining, Social Media Analytics, and Business Analytics. He is a member of the IEEE.

Raymond Y. K. Lau is an Associate Professor in the Department of Information Systems at City University of Hong Kong. He has worked at the academia and the ICT industry for over twenty years. He is the author of more than 100 refereed international journals and conference papers. His research work has been published in renowned journals such IEEE Transactions on Knowledge and Data Engineering, IEEE Internet Computing, ACM Transactions on Information Systems, Decision Support Systems, etc. His research interests include Information Retrieval, Text Mining, and Social Media Analytics. Dr. Lau is a senior member of the IEEE and the ACM respectively.