Journal of Applied Sciences

ISSN 1812-5654

science alert

ANSInet
an open access publisher
http://ansinet.com

Journal of Applied Sciences 14 (14): 1633-1637, 2014
ISSN 1812-5654 / DOI: 10.3923/jas.2014.1633.1637

Three Tier Securities for Opinion Mining Tool

G.R. Brindha, B. Santhi and P. Swaminathan
School of Computing, SASTRA University, Thanjavur, India

Abstract: The exponential growth of Web 2.0 concept influences the size of Internet information storage very much. People all over the world can easily interact with each other and share their emotions and experiences. This study aims to provide a way for secured opinion polling to improve the reliability of online reviews. While considering a set of data for review, careful selection of data set is needed, since there are wide chances for fake or spam opinions. Spam opinions can be posted by any persons who purposely post unreal information or by the business firm itself to improve their productivity by cheating users. Mining these reviews has become a challenging research nowadays. Many researchers concentrated on reviewing opinions for different domains. This study proposes new three tier security architecture to poll the reviews about holiday resorts. First tier creates secured customer id. Second tier is responsible for polling reviews with authorized entry with alert messages. Third tier provides restricted view to new customer who seeks reviews and read-only option of all reviews to the firm. Category wise reviews concept is useful for the customer to check their taste and for resort people to get more profit by enhancing the features to the customer. With this outcome the variations in reliability (92-62%), confidentiality (98-22%) and agility (89-42%) of opinion mining tool, with security and without security is analyzed.

Key words: Opinion mining, three tier security, preference ranking, facet

INTRODUCTION

The sudden increase of user-generated substance on the Internet has provided a different opportunities and also challenges to the companies which are caring more about their products and services (Feng *et al*., 2011) blogs, forums and social networks take part more in this regard. On commercial blogs the reviews based on bloggers' experiences about product or service can spread vary fast in cyberspace. In such situation negative comments are harmful to a business firm (Chen *et al*., 2011). Many researchers have taken variety of review data set such as news (Wilson *et al*., 2005), cars (Schumaker *et al*., 2012), bikes (Brindha and Santhi, 2012a), education, stock, computers (Wang *et al*., 2011), restaurant (Tan and Wang, 2011), MP3 players (Kang *et al*., 2012; Chen and Tseng, 2011), audio, videos players, DVD and digital camera (Chen *et al*., 2011) and more.

Some studies concentrated on new enhanced methods based on existing methods and some more compared the traditional methods performance on different data sets for classification purpose. New methods such as effective feature selection method (Brindha and Santhi, 2012a), SentRank process (Ghose and Ipeirotis, 2011), random walk algorithm (Wu and Tan, 2011), heuristic ssearch-enhanced Markov blanket model (Tan and Wu, 2011), new self learned feature extraction (Bai, 2011), new Pref-rating method and Collaborative Filtering (CF) algorithms (Hu *et al*., 2011a), Basilisk-bootstrapping algorithms (Wiebe and Riloff, 2011) are implemented.

The traditional methods Naïve Bayes (NB) and Support Vector Machine (SVM) are compared which gave the outcome that NB performs well compared to SVM (Leung *et al*., 2011; Tan and Wang, 2011). Whereas another researcher showed the influence of SVM (Wiebe and Riloff, 2011; Kang *et al*., 2012; Zhang *et al*., 2011) on different data set through comparison. Even in music industry they hire professional marketers to post comments about new albums in online chat rooms and fan sites (Saleh *et al*., 2011; Mayzlin, 2006). Another two studies (Brindha and Santhi, 2012a, b) include preference based Opinion Mining (OM) tool architecture which insist that review corpus user category will differ based on their taste which in turn influence the review result. In a detailed study (He and Zhou, 2011) reviews manipulation are posted by publishers, vendors or any third-party on behalf of customers to improve business firm's sales. The above said study deals with proving the existence of on line manipulations and numerical detection systems but ignored the textual content of online reviews (Hu *et al*., 2011b).

Corresponding Author: G.R. Brindha, School of Computing, SASTRA University, Thanjavur, India

To examine the textual content of reviews another researcher proposed a statistical 'Runs' test method to spot out products with opinions that are manipulated (Hu *et al*., 2012). But providing a foolproof way to have original real opinion is always better than identifying manipulations, or fake opinions. So this study proposes a secured three tier architecture for opinion polling. The proposed algorithm to mine the reviews is working on user preference based analysis for the chain of holiday resorts reviews.

IMPORTANCE OF SECURITY IN OPINION MINING

Basically a review exemplifies the emotions or sentiments or experiences of a person about a product, service or a person. Opinions are vibrant based on time and person. The research in OM domain is in vast expansion phase and many studies showed that it is not easy even with efficient methods and algorithms. Opinions are non separable element of decision making and so it is essential also. A fake or spam opinions can change the decision support information to its opposite sense. A fake review may be posted by the business people to increase their profit or intentionally by an individual. So a review corpus or a mining tool should be robust to fake or spam mixing.

In this fast moving world, people don't have time to recheck the decision. Ultimately people need that tool which provides data which is safe in all angles. Though an OM tool is providing perfect and efficient information and analytical data, if it is within the range of intruder's attack, then value for the tool becomes less. Reliability of the data is essential for the busy users. This is designed in three phases (1) Opinion entry phase, (2) Intruders phase and (3) Corpus phase. Hence, the proposed method includes securities in three phases to have reliable decision support information.

PRE-PROCESS OF OM

The customers are segregated into four categories such as family, couples, solo and business, since the requirements and expectations will differ from each other. Opinions about the stay are entered by the customer in on-line system through linguistic format. This raw data has to go through a pre-process to fine tune which is needed for classification process. First the sentences are separated into terms. To extract the facet from term set, stop words, such as 'a', 'the' etc., will be removed. The next process is part of speech tagging which separates adjectives and nouns. From this the terms which describes about hotel stay will be extracted. Now the facet set goes for case folding punctuation and short form process. To get the root word of dictionary, lemmatization will be done. After this the significant part of analysis i.e., negation detection and ranking will be implemented. Now the facet matrix is ready as an input for a classifier which may be as SVM or NB etc. This process provides an output with 'poor', 'normal' and 'good' (1, 2, 3) rating to the end user who wants to make decision.

PROPOSED OM TOOL

The proposed model is given in Fig. 1 with three tier security. The customer who stayed in resort will be given

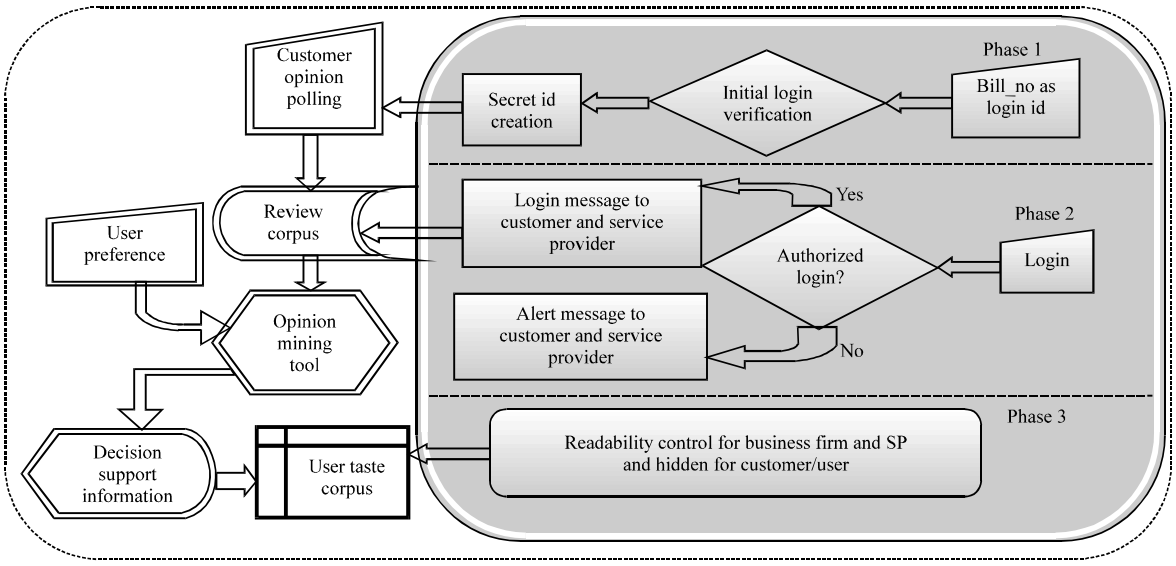

Fig. 1: Three tier architecture for OM tool

an initial login id based on bill number. To poll their reviews they have to use this id for first login. After that the login id (bill_no) will be combined with four letter word given by the customer. This enables secured entry of opinion into the corpus and prevents fake polling. Customer can post their reviews in corpus. The service provider who is managing this corpus and OM tool takes these reviews and does preprocess as explained in section III. Now any user who wants to know about holiday resort can enter their preference from 1 to 5 for the facets, 'cost, service, food, comfort and cleanliness'.

The comfort will includes, view to enjoy, nearby shopping, safety and many more. Since, the user preference may vary based on their taste, the above said 5 facets are given weights. Now the preprocessed matrix which is ready with ranking (1, 2 or 3) for 5 facets is used with weighted preference to know the analyzed result of stored reviewer's opinion. This result is stored in a separate storage which is used by the service provider to know about new user taste and the resort firm to enhance their service. For customer and the tool user this storage is hidden. The proposed algorithm is given below.

Algorithm:

Step 1: Set facet matrix as input
Step 2: Get preference from the user
Step 3: Preference based weight updates the matrix
Step 4: Individual review can be retrieved through matrix position
Step 5: No. of 'poor', 'normal', 'good' ranking is counted
Step 6: Suggested opinion is displayed

PROPOSED SECURITY TIERS IN OM

The main focus of the study is to provide a sentiment analysis tool which offers service with security. To be a robust system, the proposed architecture includes 3 tier security phase (Fig. 1). The OM tool and security are controlled and managed by OM service provider.

Phase 1: The customers who stayed in the holiday resort will be provided a customer id and password. The id is based on their bill number, so that only those who stayed in resort can post reviews. This is to avoid fake postings by others who have not stayed. Only with this id customer can become a member and enter into the port. After the initial login, customer should update the login id with eight digits, in which first four will be bill number. This is to prevent the changes or fake postings by resort people.

Phase 2: In this phase, the authentication mechanism allows customer to modify the review content. Whenever the customer logs in, message will be sent to customer's mobile phone and to the service provider. If anybody other than customer tries to log in, an alert message will be sent to the customer and service provider and entry will be blocked.

Phase 3: The final outcome i.e., decision support information is given to the new user to decide whether to utilize the service or not. And the user taste corpus is also used by the service provider to know the usage of OM tool and by the business firm to improve its profit by updating facilities. Only the current preference outcome checked by the new user can be viewed by them, not previous preferences. Those results are hidden to the user but service provider can know about the OM tool user and the resort people can use this information for further enhancement on the facet to improve profit. Hence, this corpus is a read only to service provider and business firm.

RESULTS ANALYSIS

Factor analysis was done to check main security issues in E-commerce (Folorunso *et al*., 2006), Video data security is achieved by symmetric key encryption mapping with assymmetric key encryption (Salem *et al*., 2011), data storage security is proved by novel number system compared with existing number system (Barati *et al*., 2008). In this study to analyze the proposed tool around 15 preferances were taken as input to analyze the stored reviews (200) and the overall result indicates in Fig. 2, it is good to stay in the resort. This is apparent in the comparitive chart with 88% for Good opinion, 80% for Normal opinion and Poor rating is only 32%. Figure 3. compares the categorywise mean difference of preferances. Among all, family people liked the stay compared to others. This is visible from the Poor rate reviews (less than 25) by family category compared to other categories (around 30). And also number of good reviews (100) are more by family category. In general 70-80

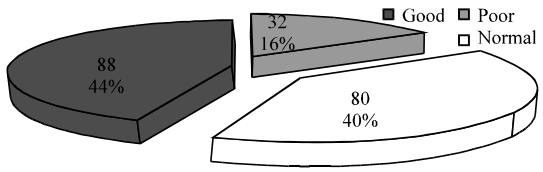

Fig. 2: Over all opinion rates from 200 reviews about holiday resort

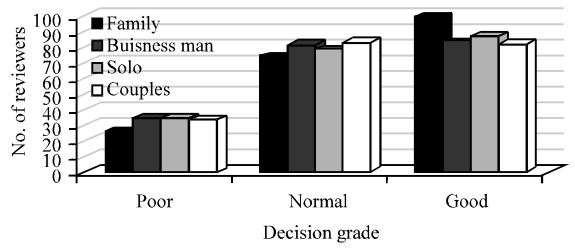

Fig. 3: Category-wise comparison of opinion range using mean difference of 200 reviews

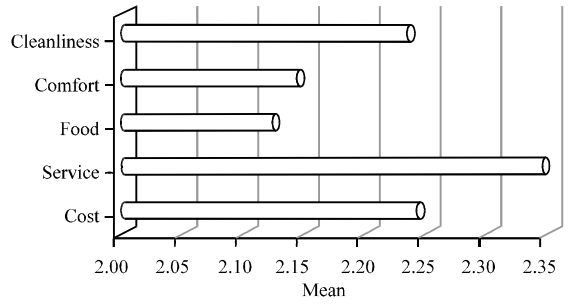

Fig. 4: Feature wise rating comparison using mean difference of 200 customer reviews

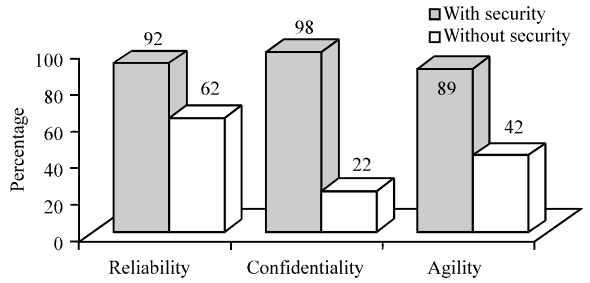

Fig. 5: Comparison of security attributes to prove the significance of security tiers

reviews rate is Normal and 80-100 reviews said it is good which is a favorable positive review for the resort. Figure 4 shows the criteria in which the resort should be improved. The mean difference of 200 reviews, depicts that service is excellent (mean 2.35), cost is nominal and cleanliness is good (range: 2.2-2.25). But the firm has to improve its service of comfort and food taste (less than 2.15) with quality.

Figure 5 depicts the significance of security tiers for opinion mining tool with the comparison of three attributes. The first attribute is Reliability which talks about trustworthyness and consisitancy of data, the data given by the customer is preserved with out third person knowledge, hence customer rely on service. This is proved by the reliability variation of 30% (92-62). And also corpus contains consistant updated data. Keeping the corpus and user account secretly, increases the confidentiality and also the user preferances are hidden to others prove this attribute. The significance of confidentiality is visible in the difference rate of 76% (98-22) which is higher than reliability. Since, the corpus is loaded with new customer pollings, the user can be provided with fresh updated reviews that takes care of agility which has variation of 47% (89-42) between with and without security.

CONCLUSION

This study deals with three tier secured architecture for an Opinion Mining tool. People used to believe information when it comes from authorized source even though the information is unbelievable or unusual one. This study used this concept to provide a foolproof system to know about a service and decide whether to utilize that service or not. The proposed tool provides assured data about holiday resort with preference based information. Hence, the results are analyzed in three dimensions such as customer, service provider and business firm which satisfy the three groups based on 3 tier security. All decision making leads to an important fact. If it is not reliable enough then it would create chaotic system. Thus reliability of the OM tool is enhanced by this proposed architecture.

REFERENCES

Bai, X., 2011. Predicting consumer sentiments from online text. Dec. Support Syst., 50: 732-742.

Barati, A., M. Dehghan, A. Movaghar and H. Barati, 2008. Improving fault tolerance in ad-hoc networks by using residue number system. J. Applied Sci., 8: 3273-3278.

Brindha, G.R. and B. Santhi, 2012a. A novel opinion mining technique for product review based on preferences. Res. J. Applied Sci. Eng. Technol., 4: 5084-5088.

Brindha, G.R. and B. Santhi, 2012b. Application of opinion mining technique in talent management. Proceedings of International Conference on Management Issues in Emerging Economies, August 17-18, 2012, Thanjavur, India, pp: 127-132.

Chen, C.C. and Y.D. Tseng, 2011. Quality evaluation of product reviews using an information quality framework. Decis. Support Syst., 50: 755-768.

Chen, L.S., C.H. Liu and H.J. Chiu, 2011. A neural network based approach for sentiment classification in the Blogosphere. J. Inform., 5: 313-322.

Feng, S., J. Pang, D. Wang, G. Yu, F. Yang and D. Xu, 2011. A novel approach for clustering sentiments in Chinese blogs based on graph similarity. Comput. Math. Appl., 62: 2770-2778.

Folorunso, O., A.O. Gabriel, S.K. Sharma and J. Zhang, 2006. Factors affecting the adoption of E-commerce: A study in Nigeria. J. Applied Sci., 6: 2224-2230.

Ghose, A. and P.G. Ipeirotis, 2011. Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics. Trans. Knowl. Data Eng., 23: 1498-1512.

He, Y. and D. Zhou, 2011. Self-training from labeled features for sentiment analysis. Inform. Process. Manage., 47: 606-616.

Hu, N., I. Bose, N.S. Koh and L. Liu, 2012. Manipulation of online reviews: An analysis of ratings, readability and sentiments. Dec. Support Syst., 52: 674-684.

Hu, N., L. Liu and V. Sambamurthy, 2011a. Fraud detection in online consumer reviews. Dec. Support Syst., 50: 614-626.

Hu, N.I., Y.G. Bose and L. Liu, 2011b. Manipulation in digital word-of-mouth: A reality check for book reviews. Dec. Support Syst., 50: 627-635.

Kang, H., S.J. Yoo and D. Han, 2012. Senti-lexicon and improved Naive Bayes algorithms for sentiment analysis of restaurant reviews. Exp. Syst. Appl., 39: 6000-6010.

Leung, C.W.K., S.C.F. Chan, F.L. Chung and G. Ngai, 2011. A probabilistic rating inference framework for mining user preferences from reviews. Worldwide Web, 14: 187-215.

Mayzlin, D., 2006. Promotional chat on the internet. Market. Sci., 25: 155-163.

Saleh, R.M., M.T. Martin-Valdivia, A. Montejo-Raez and L.A. Urena-Lopez, 2011. Experiments with SVM to classify opinions in different domains. Exp. Syst. Appl., 38: 14799-14804.

Salem, Y., M. Abomhara, O.O. Khalifa, A.A. Zaidan and B.B. Zaidan, 2011. A review on multimedia communications cryptography. Res. J. Inform. Technol., 3: 146-152.

Schumaker, R.P., Y. Zhang, C.N. Huang and H. Chen, 2012. Evaluating sentiment in financial news articles. Dec. Support Syst., 53: 458-464.

Tan, S. and Q. Wu, 2011. A random walk algorithm for automatic construction of domain-oriented sentiment lexicon. Exp. Syst. Appl., 38: 12094-12100.

Tan, S. and Y. Wang, 2011. Weighted SCL model for adaptation of sentiment classification. Exp. Syst. Appl., 38: 10524-10531.

Wang, S., D. Li, X. Song, Y. Wei and H. Li, 2011. A feature selection method based on improved fisher's discriminant ratio for text sentiment classification. Exp. Syst. Appl., 38: 8696-8702.

Wiebe, J. and E. Riloff, 2011. Finding mutual benefit between subjectivity analysis and information extraction. Trans. Affect. Comput., 2: 175-191.

Wilson, T., J. Wiebe and P. Hoffmann, 2005. Recognizing contextual polarity in phrase-level sentiment analysis. Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing, October 2005, USA., pp: 347-354.

Wu, Q. and S. Tan, 2011. A two-stage framework for cross-domain sentiment classification. Exp. Syst. Appl., 38: 14269-14275.

Zhang, Z., Q. Ye, Z. Zhang and Y. Li, 2011. Sentiment classification of Internet restaurant reviews written in Cantonese. Exp. Syst. Appl., 38: 7674-7682.