

An E-commerce Feedback Comment Mining Using SentiWordNet Tool and K-Means Clustering Method

Tinku Varghese¹, Subha Sreekumar²

¹Department of Computer Science & Engineering,
M. G. University,
Mangalam College of Engineering,
Ettumanoor, Kerala, India
tinkuvarghese90@gmail.com

²Department of C. S. E., M. G. University,
Mangalam College of Engineering, Ettumanoor,
Kerala, India
subha.sreekumar@mangalam.in

Abstract: Reputation based trust models are widely used in e-commerce applications. The feedback comments are used to compute sellers' reputation trust score. This system is based on the observation that the buyers can express their opinions openly and honestly in free text feedback comments. If there are no feedback comments available then the buyer need to consider the features specified for each product. The tool SentiWordNet is used for the extraction of feedback comments into positive, negative and neutral. K-means clustering method is used to group the data obtained after sentimental analysis. Each sentence in a feedback comment is considered as a document. This calculation is lead to obtain a sellers trust profile. The problem faced by all the reputation system is an *all good reputation* problem where reputation score are universally high for sellers and this will be difficult for a potential buyer to identify the potential buyer. This system will provide a solution for this system.

Keywords: E-commerce, SentiWordNet, Text mining, K-means.

1. Introduction

Reputation systems [1], [2], [4], give a proper path for developing the trust through social control without interference of trusted third parties. Most of the research on reputation based trust uses information such as community-based feedbacks about past experiences of users or peers. This is done to help making recommendation and judgment on quality and reliability of the transactions, products and delivery. The trust context changes from communities to communities and from transactions to transactions. This is important to build a reputation based system that is able to adapt to different communities and different situations.

A well-reported issue with the all e-commerce sites [1], [2] reputation management system is the *all good reputation* problem where feedback ratings are over 99% positive on average. Such strong positive bias can hardly guide buyers to select sellers who are not trustworthy. These sites detailed seller ratings for sellers (DSRs) [1], [11] on four aspects of transactions, namely *item as described*, *communication*, *postage time*, and *postage and handling charges*. DSRs are aggregated rating scores on a 1- to 5-star scale. Still the strong positive bias is present – aspect ratings are mostly 4.8 or 4.9 stars. One possible reason for the lack of negative ratings at e-commerce web sites is that users who leave negative feedback ratings can attract retaliatory negative ratings and that will damage their own reputation.

The three challenges [4], [7], of reputation systems are: (1) provide information that will allow buyers to differentiate between trustworthy and non-trustworthy sellers (2) make the sellers to be trustworthy, and (2) avoid the participation of unverified users. In the terminology of asymmetric information, the number 2 and number 3 challenges are that a reputation system must prevent moral hazard and adverse selection on the part of sellers.

E-commerce [5], [9], is expanded as electronic commerce, is trading in products or services using computer networks, such as the Internet. Electronic commerce draws on technologies such as mobile commerce, electronic funds transfer, supply chain management, Internet marketing, online transaction processing, electronic data interchange (EDI), inventory management systems, and automated data collection systems. Modern electronic commerce typically uses the World Wide Web for at least one part of the transaction's life cycle, although this may also use other technologies such as e-mail.

The comments from buyers and sellers about each other are collected by a reputation system. The system collects comments after each transaction [1], [2]. Examination of a huge data set from 1999 reveals several interesting features of this system. It facilitates many millions of sales each month. At very first, without incentives to complimentary ride, more than half the time response was supplied. Further, it is always positive beyond expectations. Third, future performance of reputation profiles was predictive. The best predictor available is far from it. Fourthly, although sellers with better reputations

were more likely to trade their things, they benefited from no boost in price, at slightest for the two sets of objects that were examined. Finally, an elevated correlation between buyer and seller feedback as it suggests that the players reciprocate and retaliate is present. The previous systems are mostly deal with the concept topic modeling but topic modeling is not that much effective in case of all systems. So this system uses SentiWordNet tool for opinion mining and calculate the score on three different values.

2. Related Works

Topic modeling [9], [10], [12] is a key tool for the discovery of hidden structure in large collections of documents. Probabilistic methods, such as Latent Dirichlet allocation (LDA) [1], [9] are often employed by using tools such as the Java MALLET library. In natural language processing, Latent Dirichlet allocation (LDA) [1], [9], [12] is a generative model that allows sets of observations to be explained by unobserved groups that explain why some parts of the data are similar. For example, if observations are words collected into documents, it posits that each document is a mixture of a small number of topics and that each word's creation is attributable to one of the document's topics.

3. Proposed System

Reputation based trust models are widely used in e-commerce applications and feedback ratings are aggregated to compute sellers' reputation trust scores. The *all good reputation* problem is present in current reputation systems. Here reputation scores universally high for sellers and this is difficult for a potential buyer to select the trustworthy sellers. The system is based on the observation that buyers often express their opinions openly and honestly in free text feedback comments. This system consists of a multi dimensional trust model for computing reputation scores from users feedback comment. This system uses a tool SentiWordNet for opinion mining and the K-means clustering method for grouping the feedback comments on different aspects. SentWordNet is a lexical resource where each WORDNET is synset denoted as s is associated to three numerical values that are $obj(s)$, $pos(s)$ and $neg(s)$ which describes how positive, negative and objective the terms contained in the synset. SentiWordNet is based on the quantitative analysis of the glosses associated to synset. Extensive experiments on Amazon data demonstrate that this multi dimensional system can effectively address the *all good reputation* issue and rank sellers effectively.

3.1 Comment Based Multi Dimensional Trust Evaluation

The text feedback comments are the input for the system where the users might express their opinion in mixed format based on different aspects like transaction, delivery and shipping time. So we call this salient aspect as dimensions of e-commerce feedback comments therefore the comment based trust evaluation is multi dimensional. Here we consider an example *bad communication will not buy from again, super slow shipping, item as described* where the buyer gave a positive feedback rating for a transaction but he left negative feedback towards communication and delivery.

3.2 Procedure

The frontend of the system collect the feedback comments from the users and this feedback comments are the input for the system.

The steps involved in the system:

- Initially the system selects the data set needed. The data set is the feedback from the users.
- In this step, data set training is done by using preprocessing, stemming and tagging.
 - The preprocessing technique is used for eliminating the garbage's in the feedback comments.
 - After completing the preprocessing step. Stemming is used for information retrieval for reducing the infected word or removes the stems which have no importance in the further processing.
 - Then in tagging, the system tag the remaining words into different classes like adjective, adverb, noun, verb etc
- After completing the data set training, we uses SentiWordNet tool for quantitative analysis to obtain the values such as objective, positive and negative. Using this tool we calculate the positive and negative values for the different classes of words.
- In the next step, we use K-means clustering method. Here we partition n observations into k clusters. Here we are considering the distance method for that considers a centroid and consider objects which have smallest distance to the centroid and take that object into a particular cluster. Here each cluster can be the sellers and we allocate positive and negative bias to each seller.
- The trust score is calculated with the equation (1).
- From this score, we can identify the trustworthy seller.

3.3 Trust Score Calculation

We calculate the scores related to positive, negative and objective values using SentiWordNet tool. Then we use K-means clustering method to group the different terms into different sellers and then we calculate the trust score using the following equation (1).

$$T = \sum_{d=1 \dots m} t_d * w_d \quad (1)$$

$T \rightarrow$ Trust score for sellers

$d \rightarrow$ Dimensions

$t_d \rightarrow$ Trust score for dimension d

$w_d \rightarrow$ Weight for dimension d

This equation obtains a score which will be computed from the feedback comments and help to identify the importance of different sellers.

3.4 System Architecture

The following figure (Fig. 1) is the detailed framework of comment based system where the system takes feedback comments as the input. The data set training is done on this input data using the techniques like preprocessing, stemming and tagging [6]. After completing data set training, unwanted data are documented as not applicable and the relevant data are taken and using SentiWordNet tool quantitative opinion mining

is done. The scores of each word classes are calculated, if there is no sufficient data available then the system again go back to the input session and load another feedback set. Then K-means clustering method is applied. The trust score is calculated using equation (1). Finally compute the trust profile for each seller.

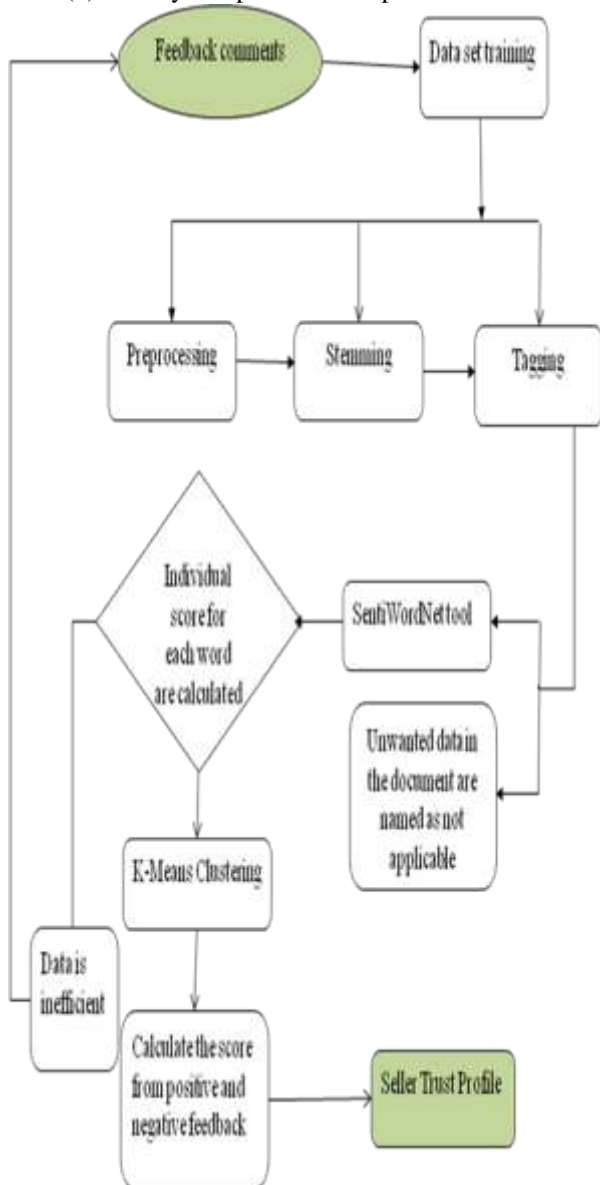


Fig. 1. Architecture of comment based model

4. Experimental Results

The following graphical representation (Fig. 2) shows the comparison between the score obtained for each seller. The score will be taken from the seller's trust profile. This graphical representation shows that the *seller1* is more trustworthy than *seller2* as per the seller's trust profile.

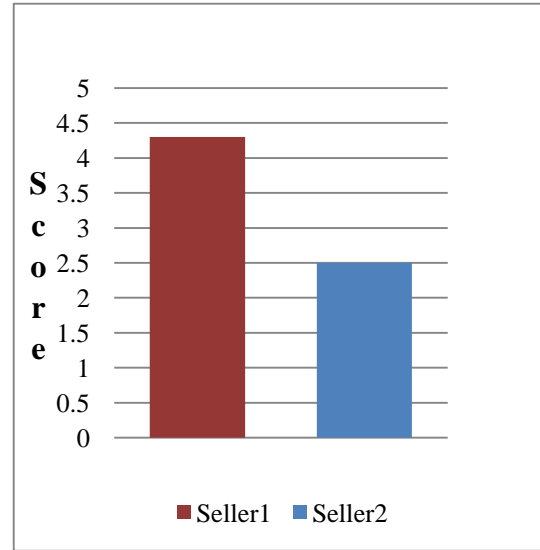


Fig. 2. Comparison between two sellers

The figure (Fig. 3) depicts the comparison between the qualities of product with respect to scores. The scores obtained in each product are taken and calculate the average of this product scores to compute the seller trust profile.

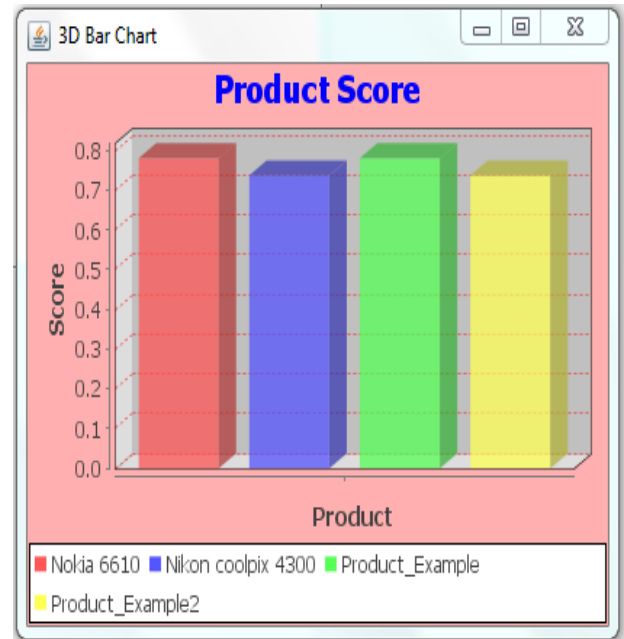


Fig. 3. Comparison between different products

5. Conclusion

This system acts as a data analyst in the e-commerce applications and this will collect the reviews given by the user and then using the SentiWordNet tool and K-means clustering method for mining the comments as in sentimental analysis likewise the data extraction is done and trust score is calculated. This system efficiently gives the solution to *all good reputation* problems so all e-commerce reputation systems are preferable to opt this system for better performance. This system computes two features such as trust score of dimension and weight of dimension to calculate the final trust score. SentiWordNet is a tool will efficiently replace and make easy to handle and calculate trust of each word in their word classes.

References

- [1] Xiuzhen Zhang, Lishan Cui, and Yan Wang, "CommTrust: Computing Multi-Dimensional Trust by Mining E-Commerce Feedback Comments," *IEEE Transactions on Knowledge And Data Engineering*, VOL. 26, NO. 7, JULY 2014
- [2] P. Resnick, K. Kuwabara, R. Zeckhauser, and E. Friedman, "Reputation systems: Facilitating trust in internet interactions," *Commun. ACM*, vol. 43, no. 12, pp. 45–48, 2000.
- [3] P. Resnick and R. Zeckhauser, "Trust among strangers in internet transactions: Empirical analysis of eBay's reputation system," *Econ. Internet E-Commerce*, vol. 11, no. 11, pp. 127–157, Nov. 2002.
- [4] J. O'Donovan, B. Smyth, V. Evrim, and D. McLeod, "Extracting and visualizing trust relationships from online auction feedback comments," in *Proc. IJCAI*, San Francisco, CA, USA, 2007 pp. 2826–2831, 2007.
- [5] M. De Marneffe, B. MacCartney, and C. Manning, "Generating typed dependency parses from phrase structure parses," in *Proc. LREC*, vol. 6., pp. 449–454, 2006.
- [6] M. De Marneffe and C. Manning, "The Stanford typed dependencies representation," in *Proc. CrossParser*, Stroudsburg, PA, USA, 2008.
- [7] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Found. Trends Inf. Ret.*, vol. 2, no. 1–2, pp. 1–135, Jan. 2008
- [8] B. Liu, *Sentiment Analysis and Opinion Mining*. San Rafael, CA, USA: Morgan & Claypool Publishers, 2012.
- [9] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Jan. 2003.
- [10] T. Hofmann, "Probabilistic latent semantic indexing," in *Proc. 22nd ACM SIGIR*, New York, NY, USA, pp. 50–57, 1999.
- [11] A. Jøsang, R. Ismail, and C. Boyd, "A survey of trust and reputation systems for online service provision," *DSS*, vol. 43, no. 2, pp. 618–644, 2007.
- [12] S. D. Kamvar, M. T. Schlosser, and H. Garcia-Molina, "The Eigen Trust algorithm for reputation management in P2P networks," New York, NY, USA, pp. 194–195, 2001.

Author Profile



Tinku Varghese received B. E. degree in Computer Science & Engineering from Swami Vivekanand Technical University, Bhilai in 2013. Pursuing M. Tech degree in Computer Science & Engineering from M. G. University, Kerala.



Subha Sreekumar is holder of B. Tech & M. Tech in Computer Science & Engineering. Now she is Assistant Professor in Mangalam College of Engineering, Kerala.