Product Aspect Ranking, Sentiment Analysis and Classification of Product Reviews on E-Commerce Sites- A Review

Ranjeet Singh
Assistant Professor Department of CSE, Chandigarh Engineering College Landran, Mohali.

Jagbir Singh Gill
Assistant Professor Department of Computer Science and Engg, Chandigarh Engineering College Landran, Mohali.

Tejpal Sharma
Assistant Professor Department of Computer Science and Engg, Chandigarh Engineering College Landran, Mohali.

Gaurav Goel
Assistant ProfessorDepartment of CSE, Chandigarh Engineering College Landran, Mohali.

Abstract –Customers reviewing their desired product online has become commonplace for people to express their opinions and sentiments toward the products bought or services received. Like social networks, Sentiment classification has been classified as the problem of training a binary classifier using reviews annotated for positive or negative sentiment. The vital product aspects are identified based on two factors: 1) the vital issues are usually commented on by a large number of consumers and 2) consumer opinions on the vital aspects greatly influence their overall desirability of product. Our analysis shows that both the sentiments expressed in the reviews and the quality of the reviews have a significant impact on the future sales performance of products in question.

Index Terms – Review mining, sentiment analysis, Aspect Ranking,

1. INTRODUCTION

Millions of products from various vendors are being offered online. For example, Bing Shopping1 archives more than five million products. Amazon.com sells a total of more than 36 million products.Shopper.com records more than five million productsfrom over 3,000 merchants. Alibaba.com dominates sales in china. In India Flipkart, Amazon, Shopper-stop are leading online retailers. Wide access to internet and ubiquitous mobile devices facilitate this trend. India's internet user base 355 million, registers 17.5 % growth in first 6 months of 2015: IAMAI (Internet and Mobile Association of India.) report. The base had grown to 303 million by the end of 2014 after clocking its fastest rise of 32% in a year, as per IAMAI, which includes members such as Google, Microsoft, Yahoo, eBay, IBM, Flipkart, Ola, Myntra, Uber and LinkedIn. Thus number of active users is booming.

While it took more than a decade for the user base to swell from 10 million to tenfold, and three years to cross the 200 million mark, it took only a year for the user base to rise to 300 million from 200 million.

2.SENTIMENT ANALYSIS

Sentiment analysis of natural language texts is a large and rising domain. Sentiment analysis or Opinion Mining is the computational analysis of opinions, sentiments and viewpoints of text [1]. Sentiment analysis is a Natural Language Processing and knowledge Extraction task that aims to quantify customer's feelings expressed in positive or negative comments, questions and requests, by analyzing a large numbers of documents on web. Transforming a piece of text to a feature vector is the fundamental step in any data driven approach to Sentiment analysis.

2.1 Naive Bayes for Ranking

Naive Bayes (simply NB) [4] has been popular algorithm in machine learning and data mining for effective classification. Because its conditional independence assumption is rarely true, researchers have modified naive Bayes. The related research work can be broadly divided into two categories: eager learning and lazy learning. Depending on time of major computation occurs. Different from eager approach, the key idea for extending naive Bayes from the lazy approach is to learn a naive Bayes for each testing example.

Classification is one of the most essential task in machine learning and data mining. Learning Bayesian classifiers aims to construct a special Bayesian networks from a given set of pre-classified instances, each of which is represented by a

vector with numerical values. Let$A_i, i = 1,2,3, \dots, n$ are n attributes which take values $a_i, i = 1,2,3 \dots, n$ respectively. These attributes help us to predict beforehand value c of the class C. Thus, the Bayesian classifier represented by a Bayesian network can be defined as:

$$arg_{c\in C}^{max} P(c)P(a_1, a_2, \dots\dots, a_n|c) \quad \text{(a)}$$

Let all attributes are independent given the class.

i.e.

$$P(a_1, a_2, \dots\dots, a_n|c) = \prod_{i=1}^{n} P(a_i|c) \quad \text{(b)}$$

The resulting classifier is known as naive Bayesian classifier, or simply naive Bayes:

$$arg_{c\in C}^{max} P(c) \prod_{i=1}^{n} P(a_i|c) \quad \text{(c)}$$

Naive Bayes is a probability-based classification model which is based on the fact that attributes are conditionally independent given the class label. Despite its advantages, such as conceptual and computational simplicity, its attribute independence assumption misleads the result adversely [3], if there are strong attribute dependencies.

3. PORPOSED MODELLING

Product Aspect Ranking framework is composed of three main parts: (1) aspect identification; (2) sentiment classification on aspects; and (3) probabilistic aspect ranking. After collecting comments, one should first identifythe aspects in the reviews and then analyze consumer views on the aspects via a sentiment classifier. In this paper we propose a probabilistic aspect ranking algorithm to judge vital aspects of given product by taking intoaccount aspect frequency and the influence of consumers'opinions given to each aspect over their opinions.The overall opinion in a review is an sum total of the opinions given to different aspects in the review, and specific aspects have varied contributions in the aggregation calculation which is provided to algorithm as an input. The opinions on vitalaspects have strong impact on the generation of overall opinion and vice versa. To model such aggregation, by formulate that the overall rating$O_r$in each review r is generated based on the weighted sum of the opinions on specific aspects, as $\sum_{k=1}^{m} \omega_{rk}, O_{rk}$ or in matrix form as $\omega_r^T O_r$. $O_{rk}$is the opinion on aspect $a_k$and the importance weight ω_{rk}reflects the emphasis placed on a_k. Larger ω_{rk}signifies$a_k$is more important, and vice versa. $\omega_r$denotes a vector of the weights, and *or* is the opinion vector with each dimension indicating the opinion on a particular aspect. Specifically, the observed overall ratings are letd to be generated from a *Gaussian Distribution*, with mean $\omega_r^T O_r$ and variance σ^2 as:

$$p(O_r) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left\{-\frac{(O_r - \omega_r^T O_r)^2}{2\sigma^2}\right\}. \quad (1)$$

In order to take the uncertainty of $\omega_r$into consideration, we let$\omega_r$as a sample drawn from a *Multivariate Gaussian Distribution* as:

$$p(\omega_r) = \frac{1}{(2\pi)^{m/2}|\sum|^{1/2}} exp\left\{-\frac{1}{2}(\omega_r - \mu)^T \sum^{-1} (\omega_r - \mu)\right\}, \quad (2)$$

where $\mu$and $\sum$ are the mean vector and covariance matrix, respectively. They are both unknown and need to be estimated.

Thus, the aspects that are frequently commented by consumers are likely to be valuable. Hence, we exploit aspect frequency as the prior knowledge to assist learningω_r. In particular, we expect the distribution ofω_r, i.e.,$\mathrm{N}(\mu, \sum)$is close to the distribution$\mathrm{N}(\mu_0, \mathrm{I})$. Each element in $\mu_0$is the frequency of a specific aspect:$frequency(a_k) \Big/ \sum_{i=1}^{m} frequency(a_i)$. Thus, we formulate the distribution $\mathrm{N}(\mu, \sum)$ based on its Kullback-Leibler (KL) divergence to $\mathrm{N}(\mu_0, \mathrm{I})$as

$$\mathrm{p}(\mu, \sum) = exp\{-\varphi . KL(\mathrm{N}(\mu, \sum)||\mathrm{N}(\mu_0, \mathrm{I}))\} \quad (3)$$

where $\varphi$is a weighting parameter.

Base on the above formula, the probability of generating overall opinion rating $O_r$in review *r* is given as

$$P(O_r|r) = P(O_r|\omega_r, \mu, \sum, \sigma^2) = \int p(O_r|\omega_r^T o_r \sigma^2).p\,(\omega_r|\mu, \sum) d\omega_r, \quad (4)$$

Where $\{\omega_r\}_{r=1}^{|R|}$ are the importance weights and $\{\mu, \sum, \sigma^2\}$ are the model parameters. While $\{\mu, \sum, \sigma^2\}$ can be estimated from review corpus$R = \{r_1, \dots, r_{|R|}\}$ using the maximum likelihood (ML) estimation, $\omega_r$in review *r* can be optimized through the maximum a posteriori (MAP) estimation. Since $\omega_r$and $\{\mu, \sum, \sigma^2\}$ are coupled with each other, we here optimize them using a EM-style algorithm. We iteratively optimize $\{\omega_r\}_{r=1}^{|R|}$ and $\{\mu, \sum, \sigma^2\}$ in each E-step and M-step respectively as follows.

Optimizing $\omega_r$given $\{\mu, \sum, \sigma^2\}$:

Assuming we are given the parameters$\{\mu, \sum, \sigma^2\}$, we use the maximum a posteriori (MAP) estimation to get the optimal value ofω_r. The object function of MAP estimation for review *r* is defined as:

$$\mathcal{L}(\omega_r) = logp(O_r|\omega_r^T o_r \sigma^2) p(\omega_r|\mu, \sum) p(\mu, \sum) \quad (5)$$

By substituting Eq. (1) - (3), we get

$$\mathcal{L}(\omega_r) = -\frac{(O_r - \omega_r^T O_r)^2}{2\sigma^2} - \frac{1}{2}(\omega_r - \mu)^T \sum^{-1}(\omega_r - \mu) - \varphi . KL\big(\mathrm{N}(\mu, \sum)||\mathrm{N}(\mu_0, \mathrm{I})\big) - log\left(\sigma|\sum|^{1/2}(2\pi)^{\frac{m+1}{2}}\right). \quad (6)$$

$\omega_r$can thus be optimized through MAP estimation as follows:

$$\widehat{\omega}_r = arg_{\omega_r} max \mathcal{L}(\omega_r) = arg_{\omega_r} max \left\{ -\frac{(O_r - \omega_r^T O_r)^2}{2\sigma^2} - \frac{1}{2}(\omega_r - \mu)^T \sum^{-1}(\omega_r - \mu) \right\} \quad (7)$$

We take the derivative of $\mathcal{L}(\omega_r)$with respect to ω_r and let it vanish at the minimizer:

$$\frac{\partial \mathcal{L}(\omega_r)}{\partial \omega_r} = \frac{(\omega_r^T O_r - O_r).o_r}{\sigma^2} - \sum^{-1}(\omega_r - \mu) = 0, \quad (8)$$

which results in the following solution:

$$\widehat{\omega}_r = \left(\frac{O_r {O_r}^T}{\sigma^2} + \sum^{-1}\right)^{-1}\left(\frac{O_r O_r}{\sigma^2} + \sum^{-1}\mu\right). \quad (9)$$

Optimizing $\{\mu, \sum, \sigma^2\}$ givenω_r: Given$\{\omega_r\}_{r=1}^{|R|}$, we optimize the parameters $\{\mu, \sum, \sigma^2\}$ using the maximum-likelihood (ML) calculation over the review corpus *R*. The probability of observing all the overall ratings on the corpus *R*is maximized by parameters. Therfore, they are estimated by maximizing the log-likelihood function over the whole review corpus *R*. Taking a simple case we denote $\{\mu, \sum, \sigma^2\}$ as Ψ.

$$\widehat{\Psi} = arg_{\Psi} max \mathcal{L}(R) = arg_{\Psi} max \sum_{r \in R} logp(O_r|\mu, \sum, \sigma^2). \quad (10)$$

By putting values Eq.(1) - (3), following equation is obtain

$$\widehat{\Psi} = arg\ max_{\Psi} \sum_{r \in R} \left\{ -\frac{1}{2}(\omega_r - \mu)^T \sum^{-1}(\omega_r - \mu) - \frac{(O_r - \omega_r^T O_r)^2}{2\sigma^2} - \varphi . KL\big(\mathrm{N}(\mu, \sum)||\mathrm{N}(\mu_0, \mathrm{I})\big) - log\left(\sigma|\sum|^{1/2}(2\pi)^{\frac{m+1}{2}}\right) \right\} \quad (11)$$

The derivative of *L(R)* with respect to each parameter in $\{\mu, \sum, \sigma^2\}$ is taken and let it vanish at the minimizer:

$$\frac{\partial \mathcal{L}(R)}{\partial \mu} = \sum_{r \in R}\{-\sum^{-1}(\omega_r - \mu)\} - \varphi(\varphi_0 - \mu) = 0$$

$$\frac{\partial \mathcal{L}(R)}{\partial \sum} = \sum_{r \epsilon R}\{(\sum^{-1})^T + ((\sum^{-1})^T)(\omega_r - \mu)(\omega_r - \mu)^T(\sum^{-1})^T\} + \varphi.((\sum^{-1})^T - \mathrm{I}) = 0$$

$$\frac{\partial \mathcal{L}(R)}{\partial \sigma^2} = \sum_{r \in R}\left(-\frac{1}{\sigma^2} + \frac{(O_r - \omega_r^T o_r)^2}{o^4}\right) = 0, \quad (12)$$

which lead to the following solutions:

$$\hat{\mu} = (|R|.\sum^{-1} + \varphi.\mathrm{I})^{-1}\left(\sum^{-1}\sum_{r \in R}\omega_r + \varphi.\mu_0\right)$$

$$\hat{\sum} = \left(\frac{1}{\varphi}\sum_{r \in R}\big((\omega_r - \mu)(\omega_r - \mu)^T\big) + \left(\frac{|R| - \varphi}{2\varphi}\right)^2.\mathrm{I}\right)^{1/2} - \frac{|R| - \varphi}{2\varphi}.\mathrm{I}$$

$$\hat{\sigma}^2 = \frac{1}{|R|}\sum_{r \in R}(O_r - \omega_r^T o_r)^2.| \quad (13)$$

These two optimization steps will be executed again and again till the likelihood value converges. After getting the importance weights ω_r for each review $r \in R$, the overall importance score $\bar{\omega}_k$ of each aspect a_k is calculated by combining its scores over the reviews as$\bar{\omega}_k = \sum_{r \epsilon R}\omega_{rk} \Big/ |R_k|$, where $R_k$is the set of reviews containing a_k. According to$\bar{\omega}_k$, the vital product aspects can be identified.

Algorithm 1
Algorithm for extracting Sentiment of Review Comment and implementing PAR Algorithm

Require: Product Review Document
Ensure: Sentiment of User comment.
1. Fetch the comment to review corpus R, each review r ϵ R is associated with overall ratingO_r, and a vector of opinion o_r on specific aspects.
2. Convert the unstructured comment data to structured document.
3. Tokenize the sentences into keywords.
4. Eliminate Stop words and tag the tokens using POS tagger.
5. If term is not in the dictionary check for the correct word.
6. Apply Nave Bayes classifier.
7. Calculate Precision Recall and F measure.
8. Apply decision tree algorithm.
9. **Output**: Importance score $\bar{w}_k|_{k=1}^m$for all the m aspects.
while not converged **do**
Update $\{w_r\}_{r=1}^{|R|}$ according to Eq.(9);
Update $\{\mu, \Sigma, \sigma^2\}$ according to Eq.(13);
end while

Compute aspect importance score $\{\overline{w}_k\}_{k=1}^{m}$
10. Return sentiment and sentiment score of review

4. CONCLUSION

In this paper, we have proposed a framework to quantify the vital aspects of commodities from large database of consumer opinions expressed online by users, using product aspect ranking in real time. The above mentioned framework consists of three parts, that is, product aspect identification, aspect sentiment classification, and aspect ranking. Firstly, we used the Pros and Cons reviews to improve aspect identification and sentiment classification on free-text reviews. Thereafter probabilistic aspect ranking algorithms were compared to give out the usefulness of various aspects of a product from numerous web reviews. The algorithm simultaneously compares aspect frequency and the influence of consumer opinions given to each aspect over the overall opinions. The product aspects are ranked in order of importance scores. It is found that out of Naïve Bayes (**NB**), Maximum Entropy (**ME**), and Support Vector Machine (**SVM**) the supervised approach is optimal. Preference can be further improved in future with implementing SVM by adding libSVM with linear kernel, NB for ranking can be coupled with Laplace smoothing, and ME can be used along with L-BFGS parameter estimation.

REFERENCES

[1] Song, H., Chu, J., Hu, Y., & Liu, X. (2013, December). Semantic Analysis and Implicit Target Extraction of Comments from E-Commerce Websites. In *Software Engineering (WCSE), 2013 Fourth World Congress on* (pp. 331-335). IEEE.

[2] Yadav, M. P., Feeroz, M., & Yadav, V. K. (2012, July). Mining the customer behavior using web usage mining in e-commerce. In *Computing Communication & Networking Technologies (ICCCNT), 2012 Third International Conference on* (pp. 1-5). IEEE.

[3] Kumar Singh, P., Sachdeva, A., Mahajan, D., Pande, N., & Sharma, A. (2014, September). An approach towards feature specific opinion mining and sentimental analysis across e-commerce websites. In *Confluence The Next Generation Information Technology Summit (Confluence), 2014 5th International Conference-* (pp. 329-335). IEEE.

[4] Yu, C., & Ying, X. (2009, December). Application of Data Mining Technology in E-Commerce. In *Computer Science-Technology and Applications, 2009. IFCSTA'09. International Forum on* (Vol. 1, pp. 291-293). IEEE.

[5] Sellam, T. (2010). Embedding Naive Bayes classification in a Functional and Object Oriented DBMS.

[6] Anitha, N., Anitha, B., & Pradeepa, S. (2013). "Sentiment classification approaches–A review". International Journal of Innovations in Engineering and Technology (IJIET), vol. 3(Issue 1), pp. 22-31.

[7] Bahrainian, S. A., & Dengel, A. (2013, November). "Sentiment Analysis Using Sentiment Features". In Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies. Vol. 3, pp. 26-29. IEEE.

[8] Balog, K., Mishne, G., & De Rijke, M. (2006, April). "Why are they excited?: identifying and explaining spikes in blog mood levels". In Proceedings of the Eleventh Conference of the European Chapter of the Association for Computational Linguistics: Posters & Demonstrations (pp. 207-210). Association for Computational Linguistics.

[9] Bollegala, D., Weir, D., & Carroll, J. (2013). "Cross-domain sentiment classification using a sentiment sensitive thesaurus". Knowledge and Data Engineering, IEEE Transactions on Knowledge and Data Engineering, vol. 25(no. 8), pp. 1719-1731.

[10] Cambria, E., Schuller, B., Xia, Y., & Havasi, C. (2013). "New avenues in opinion mining and sentiment analysis". IEEE Intelligent Systems Journal, vol. 2, pp. 15-21.

[11] Garcia-Moya, L., Anaya-Sanchez, H., & Berlanga-Llavori, R. (2013). "Retrieving product features and opinions from customer reviews". IEEE Intelligent Systems, vol. (3), pp. 19-27.

[12] Ghose, A., & Ipeirotis, P. G. (2011). "Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics". Knowledge and Data Engineering, IEEE Transactions on Knowledge and Data Engineering, vol. 23(no. 10), pp. 1498-1512.

[13] Guzman, E., & Maalej, W. (2014, August). "How do users like this feature? a fine grained sentiment analysis of app reviews". In Requirements Engineering Conference (RE), 2014 IEEE 22nd International Requirements Engineering Conference (Karlskrona, Sweden) pp. 153-162.

[14] Han, J., & Gao, J. (2009). "Research challenges for data mining in science and engineering". Next Generation of Data Mining, Chapter Research Challenges for data mining in Science and Engineering, pp.1-18.

[15] Indhuja, K., & Reghu, R. P. (2014, December). "Fuzzy logic based sentiment analysis of product review documents". In Computational Systems and Communications (ICCSC), 2014 First IEEE International Conference on Computational Systems and Communications (Trivandrum). pp. 18-22. IEEE.

[16] Jiang, L., & Zhang, H. (2005, November). "Learning instance greedily cloning naive bayes for ranking". In Data Mining, Fifth IEEE International Conference on Data Mining Sponsored by the IEEE Computer Society. Houston, Texas, USA (pp. 1-8).

[17] Khairnar, J., & Kinikar, M. (2013). "Machine learning algorithms for opinion mining and sentiment classification". International Journal of Scientific and Research Publications, vol. 3(Issue 6), pp. 1-6.

[18] Kumar Singh, P., Sachdeva, A., Mahajan, D., Pande, N., & Sharma, A. (2014, September). "An approach towards feature specific opinion mining and sentimental analysis across e-commerce websites". In 2014 5th IEEE International Conference, The Next Generation Information Technology Summit (Confluence), (pp. 329-335). IEEE

[19] Liu, X., Kale, A., Wasani, J., Ding, C., & Yu, Q. (2015, June). "Extracting, Ranking, and Evaluating Quality Features of Web Services through User Review Sentiment Analysis". In Web Services (ICWS), 2015 IEEE International Conference on (pp. 153-160). IEEE.

[20] Lizhen, L., Wei, S., Hanshi, W., Chuchu, L., & Jingli, L. (2014). A novel feature-based method for sentiment analysis of Chinese product reviews. Communications, China, ISSN: 1673-5447 Vol. 11(No. 3), pp. 154-164.