ISSN: 2277-9655
Impact Factor: 4.116

IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

MULTI-FACTOR REVIEW CLASSIFICATION ALGORITHM FOR PRODUCT PROFILE BUILDING USING PROBABILISTIC CLASSIFICATION

Mandeep Kaur*, Dr. Raman Chadha, Er. Pravneet Kaur
* Research scholar, CGCTC,Jhanjeri,Mohali
Professor, Head(CSE), Chandigarh Group of Colleges Technical Campus, Jhanjeri, Mohali
Assistant Professor(CSE), Chandigarh Group of Colleges Technical Campus, Jhanjeri, Mohali

DOI: 10.5281/zenodo.60104

ABSTRACT

The online shopping portals have been grown their popularity over the recent years and have gained the multi-billion dollar business by selling the variety of the goods from their portals. The online shopping portals post the major problem to customers in the case of the pre-purchase verification of the goods, which is the major reason behind the problem arising with the higher number of after sales product returns. The touch and feel of the product during purchasing gives the customers metal satisfaction as they completely investigate the product condition before paying the amount for purchasing. This factor is entirely missing from the online sales and the only factor available the customer review, which must be read entirely or the rating can be followed for the product quality. There are several cases come to fore for the false rating in the past, which also reduces the trust of the customer over such ratings. The customer review analysis algorithm with automatic reporting for the easy to access customer review factors can ease out the product quality evaluation process. In this paper, the proposed model entirely focuses upon the extraction and evaluation of the product quality based factor based upon the individual features, which gives the detailed review analysis and facilitates the users to easily decide upon their products. The proposed model has undergone the various kinds of the experiments and can be easily considered as the robust model. The proposed model has also outperformed the existing models in the terms of product review assessment and evaluation.

KEYWORDS: Multi-factor classification, sub-category classification, automatic categorization, N-gram analysis.

INTRODUCTION

The primary question for text summarization regarding customer reviews is to predict the quality of product. Customers want to know whether to buy product or not by reading summary of reviews generated by a text summarization approach. The problem is that most of the existing methods focus on opinion mining while processing customer reviews, which is used to determine reviewer's attitude either positive or negative with respect to the various features of product. But there are some reviews which cannot be labeled as positive or negative but are still valuable. So, for encountering such reviews various methods are used. One of the appropriate approaches is to provide a summary of the product quality which could help customers to take decision accordingly.

Another major problem is that there are hundreds or thousands of reviews are present online and it is not possible to reach each and every review in order to make a buy or leave decision. In order to solve such problems, summarization approaches are widely used to determine the overall quality of product.

The difficulty lies in the fact that there could be mixed opinions in a document, and people may express the same opinion in vastly diverse ways. Hence, there are several techniques used in generating summary of hundreds or thousands of reviews in order to predict whether the product is good or not. Several methods are using several different approaches. The author of the paper has decided to develop a new technique by combining several existing techniques to produce summary in an efficient and effective way.

The **main objective** is to develop an optimized summarization approach that provides an effective summary of product on the basis of customer reviews so that it would help the interested customers to take the decision whether to buy the product or not to buy.

The research is focused on following objectives:

1. To help the customers in taking right decisions regarding purchase of the product.
2. To provide an effective summary of the product reviews to the customers who are interested in buying that particular product.
3. To analyze the different aspects of product i.e. what is negative about the product? What are the features that attract the customers to purchase?
4. To develop an optimized approach of providing the summary of the product on basis of customer reviews.

In order to achieve the objective of the method, this approach require the customers to select what kind of review, he or she is going to enter; the subject of the review and then detailed review of the product. This approach uses the pattern matching techniques to match the subject of the review with detailed review. Many different factors like review length, sentence position and many more, are considered for review scoring as well as sentence scoring in order to provide high rank to good reviews as well as useful sentences within the good reviews.

LITERATURE REVIEW

W Cheet al. Have projected a sentence compression model for that's for side based mostly sentiment analysis. During this model sentence is compressed before the sentiment analysis step. Sentence compression 'SENTCOMP' is proposed for the discarding of the gratuitous words before sentiment analysis method and therefore compacts the sentence that's uncomplicated to dissect and have an outrageous performance for sentiment analysis. The accuracy achieved by victimization this model is ninetieth. As 'Sent comp' model discards the perennial duplicate words, unwanted words however it sustain polarity based mostly info that's necessary for sentiment analysis. A. Esuli et al. Have projected associate senti wordnet that that's associate lexical resource for mining opinion. Wordnet system have scores the subsequent 'obj(s)','Neg(s)','Pos(s)' that have some numerical values that presents however negative, positive, objective is hold in set. It comes with an commendable interface. Eight Ternary classifiers are trained in order that 3 scores for set are obtained. Coaching set and learning device for several ternary classifier is restricted and thus specific classification outcome of the related system is made. Scores for opinions are calculated by standardization on the premise of appointed labels. Wordnet system can have highest score once each ternary classifier assigns same label to wordnet system otherwise the score are going to be proportional to ternary classifiers.

Minqing Hu et al. Have projected a model that summarizes opinion supported their options. Within the projected work, they have worked on the options that the shoppers have reviewed. First of all the projected work identifies what area unit the options of the merchandise that the patron have reviewed on area unit known then their scores area unit appointed on the premise of the days of frequency within the review. For each feature had been reviewed; the positive and negative opinions are determined. Than last opinion account is enforced by summarizing the opinions for extracted characteristics in 2 categories that's positive and negative. B. Pang et al. Had worked on the moving-picture show reviews for sentiment analysis. They worked upon 3 machine learning ways Naïve mathematician, most Entropy Classification and Support vector Machine. They need centered on the Sentiment classification drawback instead of the subject based mostly classification. The Support vector Machine had outperformed the 3 algorithms however the un-similarity isn't a lot of in comparison with the human generated baselines. Even in comparison with topic based mostly classification the accuracy isn't the maximum amount. Here authors have Janus-faced the matter of feature identification.

S Mukherjee et al. Has projected sentiment analysis for feature based mostly product reviews. Their main goal is to see the attainable feature by utilizing their correlation. The model get backs opinion expression that illustrates the options given by the user. The model has showed higher accuracy and had outturned in performance altogether the elective baselines. Underneath supervised classification model performs far better with large margin. The most defect of the system is that it doesn't train information that's domain specific and thus can't analyze the domain dependent sentiments. X Fang et al. have worked on sentiment analysis of product reviews on Amazon information set. Essentially their work focuses on polarity categorization of sentiments. For this polarity categorization of sentiments

they need conferred feature vector categorization methodology. The tactic is enforced at each sentence and review level. They need used 3 totally different classifiers that area unit naïve based mostly, random forest and support vector machine. The review level categorization must be improved because it doesn't perform well for implicit sentence; the sentence with neutral words. Therefore here review level categorization must be improved by extracting a lot of options and grouping them into feature vector.

EXPERIMENTAL DESIGN

The product review classification model in the proposed model has been developed for the automatic review categorization and the feature based evaluation. The proposed model has been completely developed to work upon the English language based review data and has been empowered with the WORDNET dictionary from the Princeton University. The proposed model has been designed to work with the text data, which has been obtained in the form of the excel format for the live source or the local source. The proposed model design can be further subdivided into the three major parts:

1. The application programming interfaces (APIs)
2. Automatic Feature based Categorization
3. Feature Assessment and Polarization

Application Programming Interface (API)

The API has been developed to extract the data from the online shopping portal in the form of the product reviews. The dual-stage API model has been defined in order to obtain the information from the online source and to store it in the local database after the restructuring. The dual-stage API is utilized for the obtaining the data for the experimentation. The API is capable of extracting the 40% data from the source review database in the one round, as per defined in the data release restrictions over the data source.

Automatic Feature Categorization

The automatic feature categorization is the process of classifying the input review text according to the pre-defined categories. The knowledge-based categorization process is based upon the assessment of the noun and pronoun based keywords for the evaluation of the category, which are extracted by using the supervised keyword extraction model. The proposed model can be defined in the detailed manner in the following algorithm:

Algorithm 1: Automatic Feature Categorization

1. *Perform the data acquisition*
2. *Count the number of the messages*
3. *Run the iteration for each message*
    a. *Apply the STOPWORD elimination module over the message body*
    b. *Load the knowledge list containing the keyword database*
    c. *Extract the keywords from the input message body*
    d. *Assess the keywords by analyzing them against the knowledge data*
    e. *Verify the weights of each of the defined category*
    f. *Return the result for highest category weight*
4. *Return the message category to the feature assessment and polarization model*

Feature Assessment and Polarization Model

The proposed model has been further defined with the polarization and feature quality assessment model, which once again evaluates the each of the input message under the supervised polarization module. The polarization module is responsible for the weight computation from the given list of the keywords defined with the N-gram terminology. The proposed model has been defined under the multivariate forest model for the classification of the message polarity by assessing the individual keyword based weight calculation of each of the message in the given database for overall assessment and feature specific assessment. The following algorithm deeply defines the complete working of the proposed model:

Algorithm 2: Feature Assessment and Polarization Model

1. *Load the messages data obtained in the form of excel file*
2. *Count and run the iteration for each of the message in the database*
3. *Load the standard STOPWORD list from the local source*

4. *Perform the STOPWORD elimination to remove the nouns and pronouns and other conjugations or symbols, which does not carry any polarity.*
5. *Obtain the keyword list by using the N-gram analytical module for the keyword extraction from the given database*
6. *Load the WORDNET dictionary from the local source*
7. *Compute the polarity of the message*
8. *If the message polarity is higher than 0, add it the positive messages*
9. *Else If the message polarity is lower than 0, add it the negative messages*
10. *Else If the message polarity is equal to 0, add it the neutral messages*
11. *Load the message category*
12. *Under the category listings for review classification, do the following*
    a. *If the category defined message polarity is higher than 0, add it the positive messages*
    b. *Else If the category defined message polarity is lower than 0, add it the negative messages*
    c. Else If the category defined message polarity is equal to 0, add it the neutral messages
13. Prepare the review report from the data analysis statistics
14. Return product review report

RESULT ANALYSIS

The proposed model has been undergone the various experiments and the overall accuracy of the proposed model has been assessed in the terms of recall and precision. The precision gives the accuracy of the system in the existence of the false positive cases, where the recall calculates the overall accuracy in the presence of the false negative cases. The proposed model has been found better than the existing polarization model based upon the Modified LexRank algorithm, which utilizes the ROUGE model for the product review assessment and evaluation. The following table (Table 4.1) defines the robustness of the proposed model in comparison with the Modified LexRank model. The existing model hasn't been recorded with the precision, which has been left blank in the table 4.1. On the basis of the recall value, the proposed model has been clearly outperformed the existing model.

Table 4.1: Modified LexRank vs Proposed Model

Sr. No.	System	Recall Value	Precision
1.	Modified LexRank	45.56%	-
2.	Proposed model	97.15%	95.17%

The proposed model has been compared with the existing model called hierarchical classification model. The proposed model has been recorded with the higher accuracy on the basis of the multivariate analysis as per defined in the table (Table 4.2). The table 4.2 clearly defines that the proposed model has been consistently performed better in all of the defined domains, where the keyword data has been processed differently.

Table 4.2: The multivariate sentence compression based accuracy evaluation

Method	System	Accuracy
Pang et. al's	Without compression	89%
	Manually compressed data	88%
	Automatic multilayered compression	88%
Mohammad et. al's	Without compression	91.50%

ISSN: 2277-9655
Impact Factor: 4.116

	Manually compressed data	91%
	Automatic multilayered compression	91%
Proposed	Without compression	94%
	Manually compressed data	94%
	Automatic multilayered compression	96%

The table 4.2 has been graphical represented in the following figure. The proposed model has been defined with the multivariate sentence compression for the multi-faceted assessment of the proposed model. The following figure graphical assess the quality of the proposed model as per defined in the table 4.2.

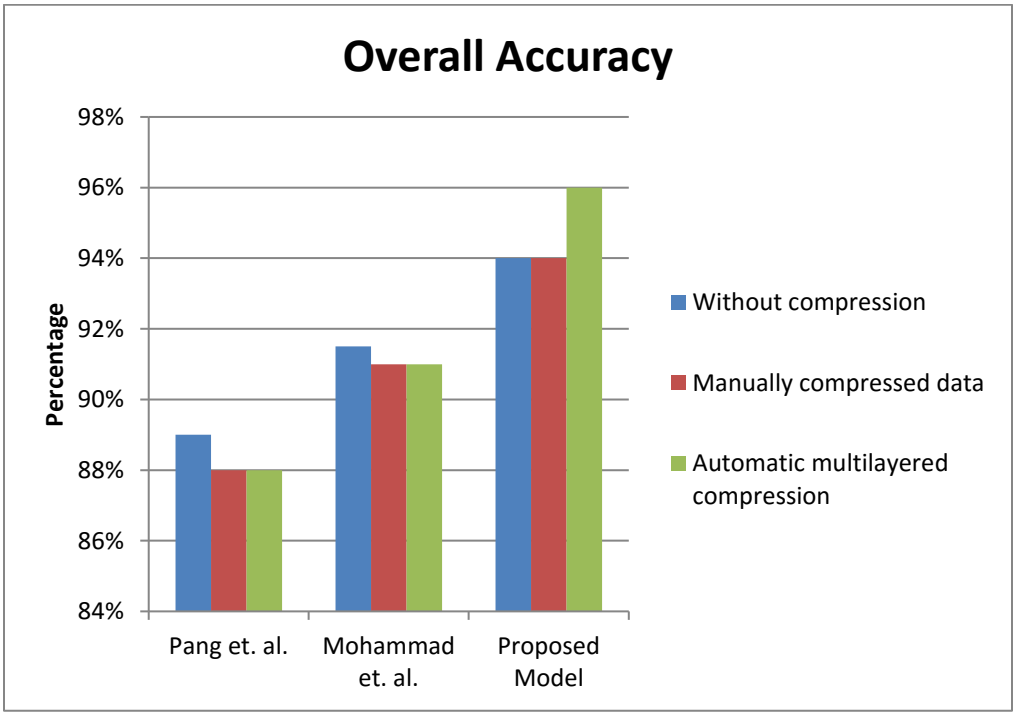

Figure 4.1: The overall accuracy based assessment in the multi-domain environment

CONCLUSION

The performance of the proposed model has been assessed in the multi-domain experimentation. The proposed model has been found very efficient in the terms of the overall accuracy. The proposed model has been found efficient in all of the domains it has been tested with. The proposed model has been recorded with nearly 96% of recall and 95% of precision in the overall testing of the message categorization and its impact assessment by using the polarization. The proposed model has been clearly outperformed the existing models as per described in the table 4.2 in the results section. It has marked the overall improvement of at-least 3-5% in all of the domains defined for the sentence compression. The experimental results have justified the robust performance of the proposed model.

REFERENCES

[1] A.h. K. Walaa Medhat, "Sentimental analysis algorithms and application:A survey," *Ain Shams Engineering Journal*, 2014.
[2] B. C. a. C. L. Azevedo, "A Sensitivity - Analysis - Based Approach for the Calibration Of Traffic Simulation Models," in *IEEE TRANSACTIONS ON iNTELLIGENT SYTEMS*, 2014.
[3] Che, Wanxiang, Yanyan Zhao, HongleiGuo, Zhong Su, and Ting Liu. "Sentence Compression for Aspect-Based Sentiment Analysis." Audio, Speech, and Language Processing, IEEE/ACM Transactions on 23, no. 12 (2015): 2111-2124.
[4] E. a. F. S. stefano BACCIANANELLA, "Star Track:The Next Generation (Of Product Reviw Management Tools)," in *new generation Computing 31(2013)47-70 Ohmsha Ltd and Springer*, Japan, 2013.
[5] E. a. F. Sebastianai, "Senti WORDNET: A publicly avaliable lexical Resource for opinion mining," in *5rd conference on language resources and evaluation*, Genova,IT, 2006.
[6] Esuli, Andrea, and FabrizioSebastiani. "Sentiwordnet: A publicly available lexical resource for opinion mining." In Proceedings of LREC, vol. 6, pp. 417-422. 2006.
[7] Fang, Xing, and Justin Zhan. "Sentiment analysis using product review data." *Journal of Big Data* 2.1 (2015): 1.
[8] Ghorashi, Seyed Hamid, Roliana Ibrahim, ShirinNoekhah, and NiloufarSalehiDastjerdi. "A frequent pattern mining algorithm for feature extraction of customer reviews." In IJCSI International Journal of Computer Science Issues. 2012.
[9] Hu, Minqing, and Bing Liu. "Mining and summarizing customer reviews."InProceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 168-177.ACM, 2004.
[10] Jin, Jian, Ping Ji, and Ying Liu. "Prioritising engineering characteristics based on customer online reviews for quality function deployment." Journal of Engineering Design 25, no. 7-9 (2014): 303-324.
[11] Jin, Jian, Ping Ji, and Ying Liu. "Translating online customer opinions into engineering characteristics in QFD: A probabilistic language analysis approach." Engineering Applications of Artificial Intelligence 41 (2015): 115-127
[12] K. H. M. N. A. R. a. J. R. Sasha Blair-Goldensohn, "Buliding a sentment summarizer for local Service Reviews," in *NLPIX 2008*, Beijing,china, 2008.
[13] K. L.-C. H.-C. L. a. C. -H. W. Hao-Chiang, "An emotion Recoginition mechanism based on the combination of mutal information and semantic cles," in *J Ambient Intell Human comuput @springer -verlag 2011*, verlag, 2012.
[14] k. P. a. H. Lim, "Acquiring lexical knowledge using raw corpora and unsupervised clustering method," in *cluster comput(2014) @springer sceince +business media newyork 2013*, New York, 2014.
[15] Kherwa, Pooja, AnishSachdeva, DhruvMahajan, NishthaPande, and Praveen Kumar Singh. "An approach towards comprehensive sentimental data analysis and opinion mining."In Advance Computing Conference (IACC), 2014 IEEE International, pp. 606-612.IEEE, 2014.
[16] Kim, Soo-Min, and Eduard Hovy. "Automatic identification of pro and con reasons in online reviews."In Proceedings of the COLING/ACL on Main conference poster sessions, pp. 483-490.Association for Computational Linguistics, 2006.
[17] Liu, Siyuan, Xiaoyin Cheng, and Fan Li. "TASC: Topic-Adaptive Sentiment Classification on Dynamic Tweets." (2015).
[18] Mukherjee, Subhabrata, and Pushpak Bhattacharyya. "Feature specific sentiment analysis for product reviews." *International Conference on Intelligent Text Processing and Computational Linguistics*.Springer Berlin Heidelberg, 2012.
[19] P. K. A. M. Simko, "Sentimental Analysis on Microblog Utilizing appraisal Theory," in *World Wide Web @Springer Science + Buiness social media*, New York, 2014.
[20] Pang, Bo, Lillian Lee, and ShivakumarVaithyanathan. "Thumbs up?: sentiment classification using machine learning techniques." *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*.Association for Computational Linguistics, 2002.
[21] R. M. D. K. R. A. Amir.H, "Dream Sentiment analysis using Second Order Soft co- occurance (SOSCO) and time Course representation," in *J Intell Inf Syst @Springer science+Business media* , NewYork, 2014.
[22] Suanmali, Ladda, NaomieSalim, and Mohammed Salem Binwahlan. "Fuzzy logic based method for improving text summarization." arXiv preprint arXiv:0906.4690 (2009).

[23] V. Y. a. H. E. PrabhU Planisamy, "Serendio:Simple and Practical Lexicon Based approach To sentiments Analysis," 2006.

[24] Yazhini, R., and Raja P. Vishnu. "Automatic summarizer for mobile devices using sentence ranking measure."In Recent Trends in Information Technology (ICRTIT), 2014 International Conference on, pp. 1-6.IEEE, 2014.

[25] Z. K. P. N. A. R. S. R. a. V. S. Theresa Wilson, "Sentiment Analysis in Twitter," in *7th international Workshop on semantics Evalutaion for Computing Lingustics*, 2013.

[26] Zha, Zheng-Jun, Jianxing Yu, Jinhui Tang, Meng Wang, and Tat-Seng Chua. "Product aspect ranking and its applications." Knowledge and Data Engineering, IEEE Transactions on 26, no. 5 (2014): 1211-1224.