A Supervised Learning Technique for Classifying Amazon Product Reviews based on Buyers Sentiments

Richa Chunekar Kokje
CSE Dept.
MIST, Indore

Gajendra Singh Chouhan
Assistant Professor
CSE Dept. MIST, Indore

ABSTRACT
A number of applications using internet provide vary essential services. Among them social media and ecommerce platforms are very common. These platforms include a large amount of data which is generated by the users. That data is available in the form of opinion about some post or review. Means the text with emotions which is contain the buyers or user sentiment about some kind of product or service. In this presented work a data mining model is introduced that offers sentiment based text classification for the Amazon product reviews. The proposed data model first preprocesses data and extract the actual review text, in next phase of preprocess the stop words and special characters are removed. The refined text is further utilized with two feature selection techniques first is based on TF-IDF which is used for selecting intense keywords from the review text. Additionally the second feature is selected using NLP text parser. That parser basically performs the POS (Part Of Speech) tagging of review text. Using the obtained POS tags the NLP feature is contracted. Both the features are combined in next and two supervised learners are used namely SVM (support vector machine) and SVR (support vector regression). The experimental results of both the model is measured and compared. The performance study demonstrates the proposed SVM based classifier performs classification accurately and efficiently as compared to SVR based classifier.

Keywords
Sentiment analysis, text classification, NLP, Amazon product review, hidden emotions on text.

1. INTRODUCTION
In recent years the internet based services are become very common. Among e-commerce, social media and video publishing websites are much common. These applications generating a bulk amount of data and it contains significant amount of information. Mining such data with some context is a task of Natural Language Processing (NLP). In this work the NLP based applications are investigated. Additionally to demonstrate the functional aspects of NLP, An application of review based text classification is proposed to implement and analyze. In this application a product review based text data is used for training and recognizing the buyer sentiments about the product. Basically in ecommerce platforms the product reviews are very influential factor which can motivate a buyer to buy a product or reject it. On the other hand supplier or product designers are always interested to know the buyer sentiments about the product. But reading all the product reviews is a complicated task.

Thus the review classification techniques help us to identify the negative sentiments among entire text. It is working like a filter to identify the reviews which are needed to be attention because the negative reviews can reduce the demand of target product or service. Thus this work is motivated to design an Amazon product review analysis system to classify the reviews into two classes negative and positive. Using this classification a product vendor can work with the negative feedbacks and improve their quality of work and product. In order to achieve this goal we proposed an extended feature model which combines two kinds of features for identifying more crucial reviews about the product and services.

2. PROPOSED WORK
The proposed work is intended to analyze the Amazon product review and classify the data according to the hidden sentiments in terms of positive and negative. This chapter offers the detailed understanding about the proposed work carried out for designing the required data model.

2.1 System Overview
Communication technology is growing frequently additionally the consumption of IT services are also growing. In this context various real world applications (Facebook, Amazon, twitter and others), are used in every age groups. During this a significant amount of data is generated by users. Such kind of data analysis needs significant efforts therefore computational algorithms are used for analyzing the data. That data can be used for various business intelligence application purposes. Basically, the users express their emotions by using social media post or any product review. During that time what the key emotions it is need to be understand. The understanding about the emotions is a research domain of NLP (natural language processing) and machine learning.

Therefore to classify the text according to their sentiments we need to implement the NLP based technique for analysis of Amazon product review. The proposed technique extracts the product reviews and preprocesses the content. During the preprocessing the noise contents are removed from input text. After preprocessing the feature extraction technique is taken place. During this the keywords are selected which are having potential sentiments. Further the features are transformed into a fixed amount of sentiment attributes using a NLP parser. These attributes are used for training and testing of the employed classifier. The classifiers are after learning able to classify the text, according to emotions positive or negative. This section provides the basic overview of the proposed work, in next section explains the required data model.

2.2 Proposed Methodology
The proposed data model for Amazon product review classification technique is demonstrated in figure 1. In this diagram the different components of sentiment classification is associated for performing required task.

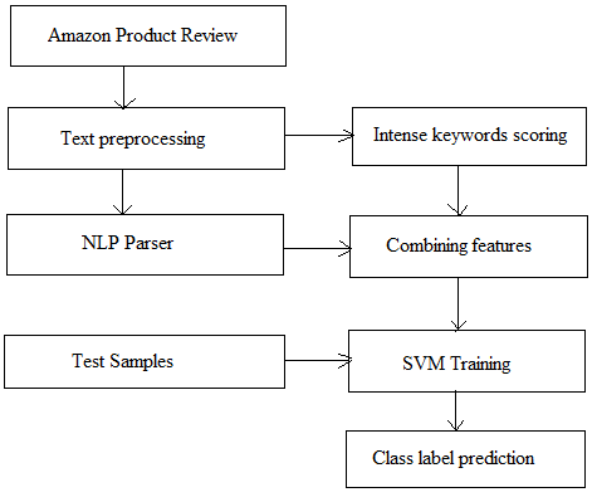

Fig.1 Proposed System

2.2.1 Amazon Product Review

The proposed system is a machine learning model for classifying the product review according to their sentiments. The model is a supervised learning model which is needed to be learnt first. The learned model is used to identify the similar class data instances. In this context the provision is made to accept the Amazon product review dataset which contains the product review relevant various attributes.

2.2.2 Text Preprocessing

The input dataset contains various attributes but we need only one attribute which contains the text reviews for target products. Thus the preprocessing technique is applied first which remove the attribute which are not necessary for the proposed system implementation. The remaining attributes are used further for keyword extraction and other task. Before this step the extracted review text is preprocessed for removing the noisy contents such as stop words and special characters. In this context the model accepts two lists first contain a list of stop words defined by the experimenter and a list of special characters. The system replaces each stop word and special character which is available in the input list. After this text processing the data is produced in next phase.

2.2.3 Intense Keyword Scoring

In a number of applications the feature selection techniques are used for extracting potential keywords from the text data. In this context the TF-IDF (Term Frequency-Inverted Document Frequency) concept is implemented. In order to calculate the term frequency the following function can be used:

$$TF = \frac{t_c}{T}$$

Where, t_c is the total count of a token, and T is termed as total word in a document.

Additionally to calculate the IDF the following function is used:

$$IDF = \log\left(\frac{N}{df_i}\right)$$

Where the N is the number of document and df_i number of document contains that token.

2.2.4 NLP Parser

After text refinement the intense keywords are preserved separately additionally the NLP parser is applied. The NLP features are used for tagging of the text in terms of POS (Part Of Speech) tags. The NLP parser accepts the Amazon product review text one by one and find out their relevant part of speech information. In order to understand we can take an example:

"The product was good"

Or

"Article Noun conjunction pronoun"

Using this example we can prepare a 2D vector such as:

Table.1 POS Tagging

Noun	**Pronoun**	**Verb**	**Adverb**	……
1	1	0	0	….

2.2.5 Combining Features

In this place two previously discussed features intense keywords which is selected on the basis of TF-IDF weight and the POS tag based extracted features are combined in a common 2D vector. Using this process the text data is transformed into a 2D vector or in a tabular format. That 2D vector is used further for training or learning process.

2.2.6 SVM Training

After reforming the text dataset the supervised learning classifier is applied with the approach of one-vs.-all. That is basically a technique where binary SVM classifier is utilized for classifying multiple class labels. Thus combined data is processed using the SVM classifier for training of classifier and after training of SVM (support vector machine) the test samples are classified for identifying the sentiment class labels.

2.2.7 Class Label Prediction

The proposed model accepts the input test dataset which contains the data instances and their class labels. The test data instances are produced to the trained SVM algorithm which computes their prediction of class. The predicted class labels and actual class labels are compared to find the prediction is correct or not. If the classifier prediction is correct then the accuracy is increases otherwise the error rate. In this manner the performance of a classifier is evaluated in terms of accuracy and error rate. At the same time the time required and memory utilized is also measured for demonstrating efficiency of the system.

2.3 Proposed Algorithm

This section provides the summary of the entire process involved for classifying the Amazon product review in terms of positive and negative classes. The proposed algorithm is demonstrated in table 2.

Table.2 Proposed Algorithm

Input: Amazon product review dataset D, List of special Character C, List of stop words S, Test Dataset T Output: Predicted class labels L
Process: 1. $D_n = ReadDataset(D)$

2. $P_n = preProcessData(D_n, C, S)$
3. $T_{model} = SVM.Train(P_n)$
4. $for(i = 1; i \leq T.length; i++)$
   a. $L = T_{model}.Classify(T_i)$
5. $endfor$
6. Return L

3. RESULT ANALYSIS

This chapter includes the results analysis and performance of the implemented algorithms which are used for developing model for Amazon text review sentiment analysis. Therefore a detailed discussion about the results and their measured parameters are reported.

3.1 Accuracy

The accuracy can be explained as the measurement of algorithm classification correctness. That can be measured using the ratio of total correctly classified and the total patterns to be classified. That can also be represented using the following equation.

$$accuracy = \frac{totalcorrectlyclassified}{totalpatternstoclassify} X100$$

Table.3 Accuracy (%)

Data instances	SVR	SVM
100	82.5	88.5
200	85.9	90.4
300	88.4	92.1
400	90.1	93.6
500	91.7	95.8
750	92.2	96.1
1000	93.9	98.3

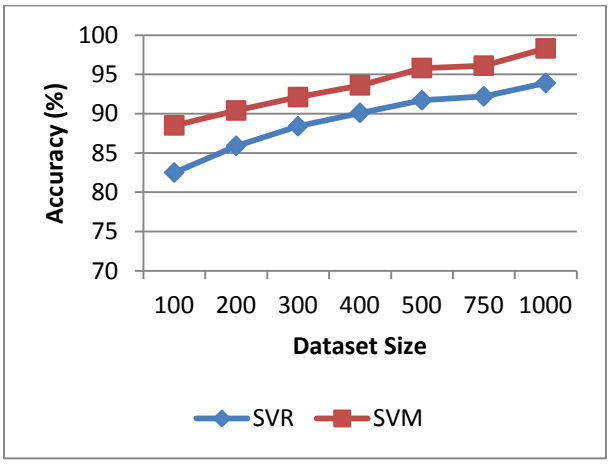

Fig.2 Accuracy (%)

The proposed data model is evaluated with the two different classifiers with the same data set and data model. The obtained performance in terms of accuracy for both the models is given in figure 2. The X axis of this line graph shows the dataset size in terms of instances and the Y axis contains the accuracy of both the classifiers in terms of percentage (%). According to the line graph we can easily identify the SVM (support vector machine) model perform better then the SVM (support vector regression). Thus the proposed model improves the classification accuracy of the existing model.

3.2 Error Rate

The error rate of an algorithm demonstrates the misclassification rate of the algorithm as a performance parameter. That can be calculated using the following equation.

$$ErrorRate = 100 - Accuracy$$

Or

$$Errorrate = \frac{totalmisclassifiedsamples}{totalsamplestoclassify} X100$$

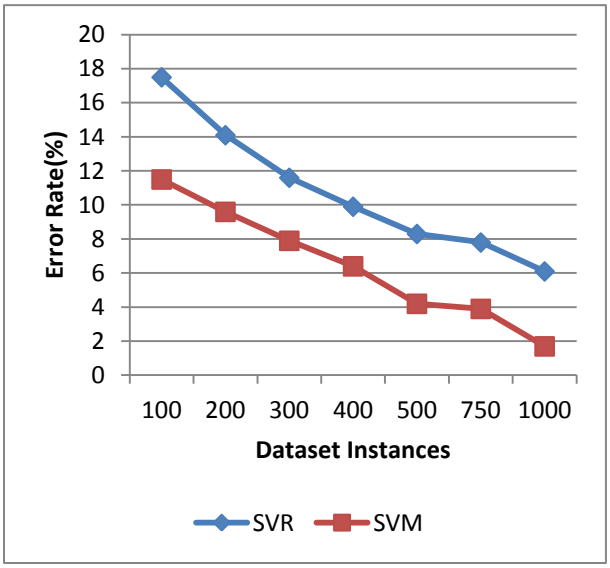

Fig.3 Error Rate

Table.4 Error Rate (%)

Data instances	SVR	SVM
100	17.5	11.5
200	14.1	9.6
300	11.6	7.9
400	9.9	6.4
500	8.3	4.2
750	7.8	3.9
1000	6.1	1.7

Reducing error rate is a good indicator of learning algorithm. In this experiment the SVM and SVR algorithm is compared for the error rate. Both the algorithm demonstrates the reducing error rate with increasing amount of data. The line graph for the performance of the system is given in figure 3 and the observation values are reported in table 4. The X axis of the diagram shows the experimental data instances and Y axis shows the percentage error rate produced. According to the experimental analysis the SVM algorithm produces less error rate as compared to SVR algorithm.

3.3 Memory Usage

The memory usages are also an essential parameter for

performance evaluation of a data mining algorithm. The amount of total memory utilized for execution of an algorithm is measured here as the memory consumption or usages. The memory usages of the algorithm are computed using the following equation.

$$memoryusage = totalmemory - freememory$$

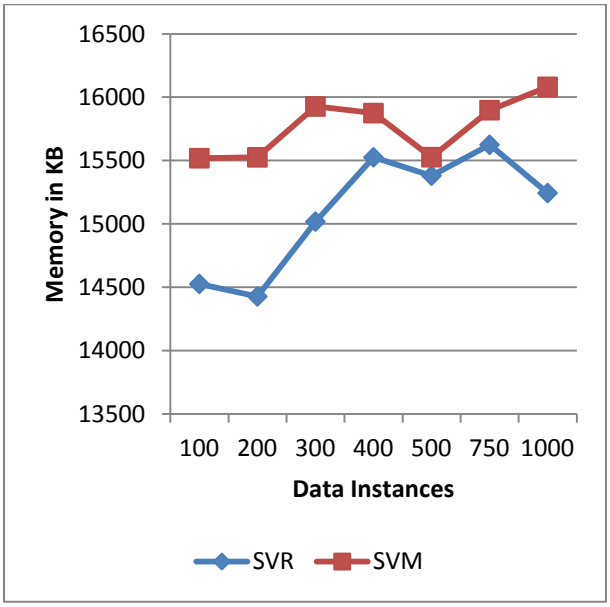

Fig.4 Memory Usage

Table.5 Memory Usage in KB

Data instances	SVR	SVM
100	14527	15519
200	14428	15526
300	15019	15927
400	15526	15875
500	15381	15527
750	15625	15898
1000	15244	16082

The memory usages are the indicator of the system efficiency. The measured memory usages for both the models are demonstrated in figure 4. It is measured here in terms of KB (kilobytes). The memory usages of the SVM algorithm are consistent as compared to SVR algorithm. But the SVM algorithm requires higher amount of memory as compared to SVM. To demonstrate the performance line graph 3.3 displays the values of table 5. Here the X axis shows the amount of data instances used for experiments and Y axis shows the used memory.

3.4 Time Consumption

The time consumption is also termed as the time complexity. The amount of time consumed for classification is also calculated in this section using the following formula.

$$timeconsumed = endtime - starttime$$

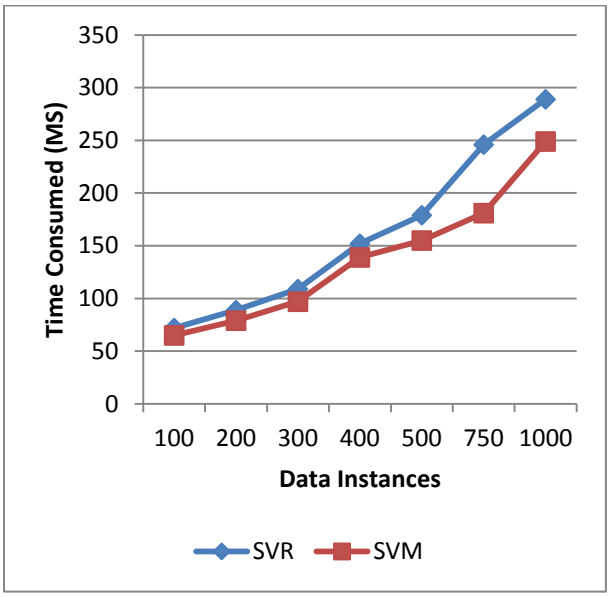

Fig.5 Time Consumption

Table.6 Time Consumption in MS

Data instances	SVR	SVM
100	72	65
200	89	79
300	109	97
400	152	139
500	179	155
750	246	181
1000	289	249

The time consumption of the proposed Amazon text review analysis using two different data mining algorithms namely SVM and SVR. The figure 5 shows the line graph for demonstrating time requirements of algorithms in terms milliseconds (MS). The X axis of this line graph shows the data instances used during the experiment and Y axis shows the time consumption. According to the line graph the SVR consumes higher time as compared to SVM. Thus the SVM shows efficiency as compared to SVR.

4. CONCLUSION & FUTURE WORK

The proposed work is aimed to find an efficient and accurate method for analyzing the Amazon product reviews. Additionally aim to classify the reviews to recognize which of the reviews are potentially negative or positive. This chapter offers the summary of the efforts made and the promising future extensions.

4.1 Conclusion

The proposed work is an investigation of the sentiment based text classification. The classification of text is a problem of text data mining but the classification of the text according to the text sentiments or emotions are known as the NLP (Natural Language Processing). In this work the application of NLP is demonstrated for identifying the reviews about a product offered by some e-commerce. Now in these days the NLP is being applied in various real world problems such as social media text analysis, student help and support system, branding of product and various other business and research subjects. The key application of the sentiment based text analysis is to understand the feeling of an author who posted a block of text using machine learning and data mining algorithms. Therefore a model using the techniques of text mining and NLP is presented.

That model first refines the contents of posted reviews on the target platform. Thus review extraction and preprocessing of review text is the initial step of the proposed data model. In next process the features are computed. In this work we considered two kinds of feature set namely the TF-IDF based weights and the POS (part of speech) tags. Further we utilized with two supervised learning techniques namely SVM and SVR. Among the SVM is modified for accepting the multi-class classification problem using the One-Vs-All technique. The trained model us to classify the raw test datasets. The classification outcomes of the test data is used for the performance assessment. The proposed model is implemented using the JAVA technology and using the WEKA machine learning library. The evaluated performance of the implemented review classification model is summarized in table 7. The table consists of the mean performance values of different experimental observations.

Table.7 Performance Summery

S. No.	Parameters	SVR	SVM
1	Accuracy	89.24 %	93.54 %
2	Error rate	10.76 %	6.46%
3	Time consumption	162.28 MS	137.85 MS
4	Memory usages	15107.14 KB	15764.85 KB

The given performance summary demonstrate the proposed working model which works with SVM and one vs. all scenario is producing higher accurate results as compared to SVR based method. Additionally SVM and one vs. all method are efficient in terms of less time consumption but costly for the memory resource consumption. Therefore according to the overall performance the proposed SVM based Amazon product review classification system is an accurate and efficient technique.

4.2 Future Work

The key aim of the proposed investigation is achieved successfully. The implemented data model is a promising model which can be extended further other kinds of text analysis by using small modifications. Some essential extension ideas are provided for work in future is given as:

1. The proposed work can also be used for analyzing the consumer's sentiments in other real world applications

2. The proposed work just utilizing the NLP parser thus in near future need to explore other suitable and effective NLP tools for improving the proposed model

3. The proposed work includes the experimental datasets, in near future we are tried to experiment using the real world data.

5. REFERENCES

[1] A. Usai, M. Pironti, M. Mital, C. A. Mejri. "Knowledge discovery out of text data: a systematic review via text mining". Journal of Knowledge Management 2018.

[2] P. Mikalef, J. Krogstie, I. O. Pappas, P. Pavlou, "Exploring the relationship between big data analytics capability andcompetitive performance: The mediating roles of dynamic and operationalcapabilities". Elsevire 2020.

[3] F. J. B. Stoffi, J. Niederreiter, M. Riccaboni, "Supervised Learning for the Prediction of Firm Dynamics". arXiv 2020

[4] V. Palanisamy, R. Thirunavukarasu, "Implications of big data analytics in developing healthcare frameworks –A review", Science Direct 2019.

[5] D. W. Otter, J. R. Medina, J. K. Kalita, "A Survey of the Usages of Deep Learning for Natural Language Processing", IEEE Transactions on Neural Networks and Learning Systems, Vol. XX, NO. X, July 2019.

[6] E. J. Topol. "High-performance medicine: the convergence of human and artificial intelligence". Nature Medicine, Vol 25, January 2019.

[7] B. Liu, "Sentiment Analysis and Opinion Mining", Sentiment Analysis Symposium 2012.

[8] B. Guo, Y. Liu , Y. Ouyang, V. W. Zheng, D. Zhang, Z. Yu, "Harnessing the Power of the General Public for Crowdsourced Business Intelligence: A Survey". IEEE Access 2019.

[9] S. Gupta, A. Leszkiewicz, V. Kumar, T. Bijmolt, D. Potapov, "Digital Analytics: Modeling for Insights and New Methods", Journal of Interactive Marketing 2020.

[10] W. N. Tun, "Sentiment Classification of Movie Review Comments Using Naive Bayesian Model", University of Computer Studies, Yangon 2016.

[11] A. S. Rathor, A. Agarwal, P. Dimri, "Comparative Study of Machine Learning Approaches for Amazon Reviews", Procedia Computer Science 132 (2018) 1552–1561.

[12] T. U. Haque, N. N. Saber, F. M. Shah, "Sentiment Analysis on Large Scale Amazon Product Reviews", 2018 IEEE International Conference on Innovative Research and Development (ICIRD), 978-1-5386-5283-1/18/$31.00 ©2018 IEEE.

[13] R. S. Jagdale, V. S. Shirsat, S. N. Deshmukh, "Sentiment Analysis on Product Reviews Using Machine Learning Techniques", Cognitive Informatics and Soft Computing, Advances in Intelligent Systems and Computing 768, https://doi.org/10.1007/978-981-13-0617-4_61.

[14] K. S. Srujan, S. S. Nikhil, H. RaghavRao, K. Karthik, B. S. Harish, and H. M. Keerthi Kumar, "Classification of Amazon Book Reviews Based on Sentiment Analysis", Information Systems Design and Intelligent Applications, Advances in Intelligent Systems and

Computing 672, https://doi.org/10.1007/978-981-10-7512-4_40.

[15] J. Singh, G. Singh, R. Singh, "Optimization of sentiment analysis using machine learning classifers", Hum. Cent. Comput. Inf. Sci. (2017) 7:32 DOI 10.1186/s13673-017-0116-3.

[16] K. Dashtipour, S. Poria, A. Hussain, E. Cambria, A. Y. A. Hawalah, A. Gelbukh, Q. Zhou, "Multilingual Sentiment Analysis: State of the Art and Independent Comparison of Techniques", CognComput (2016) 8:757–771 DOI 10.1007/s12559-016-9415-7.

[17] B. Zhao, Y. He, C. Yuan, Y. Huang, "Stock Market Prediction Exploiting Microblog Sentiment Analysis", 978-1-5090-0620-5/16/$31.00 c 2016 IEEE.

[18] A. Potdar, P. Patil, R. Bagla, R. Pandey, Prof. N. Jadhav, "SAMIKSHA - Sentiment Based Product Review Analysis System", Procedia Computer Science 78 (2016) 513 – 520, 2016 The Authors. Published by Elsevier B.V.

[19] S. Poria, E. Cambria, N. Howard, G. B. Huang, A. Hussain, "Fusing audio, visual and textual clues for sentiment analysis from multimodal content", Neurocomputing 174 (2016) 50–59, & 2015 Elsevier B.V. All rights reserved.

[20] K. Santos, E. Loures, F. Piechnicki, O. Canciglieri, "Opportunities Assessment of Product Development Process in Industry 4.0", Procedia Manufacturing 11 (2017) 1358 – 1365, © 2017 Published by Elsevier B.V.