Human Behavior Prediction based on Opinions using Machine Learning Techniques

Sanjay K S, Ajit Danti

Abstract: ***Prediction is the way of identifying the behavior of a person towards online shopping by analyzing the reviews publicly available on the web. In the present study, machine learning approaches are used to extract reviews from the web and segregate and classify them in to five categories, namely, strongly positive, positive, neutral, negative, and strongly negative, for the prediction of human behavior. Several pre-processing methods (including stop-word removal) are applied and web crawler is used to gather the data. This is followed by the application of Stanford POS tagger for tagging the reviews, which is done after stemming by using the porter stemmer algorithm. Analysis of a person's behavior is performed and experimental results are compared with machine learning approaches.***

Keywords :Behavior, Prediction, Porter Stemmer, POS tagging, Classification.

I. INTRODUCTION

Prediction is a technique used for identifying the pattern and extracting data in the field of statistics. Any kind of unknown past, present, and future facts can be subjected to predictive analysis. The association between the variables predicted from the past and the explanatory variables is determined by using predictive analysis, resulting in the prediction of future. The quality of input or assumptions and the degree of data analysis determines the accuracy of the outcomes. Various kinds of datasets are used by data mining for experimentation. Homogeneous data are collectively termed as a dataset. A variety of data, which is collected in the survey stage and which needs to be examined, constitutes the datasets. A researcher can choose any dataset to conduct the study. Some of the datasets include vehicles, products, movies, etc. Comparisons can be done by adopting different methods with different datasets. The real-valued functions are estimated by using a regression tree, which is a variant of decision tree. Input variables are allowed to be a mixture of categorical and continuous variables by the regression tree building methodology. A decision tree is generated after a test on the value of some input variable is contained by each decision node in the tree. Predicted values of the predicted output variable are contained by the tree's terminal nodes. Various machine-learning approaches are applied on our dataset to predict the behavior of a person, based on the online reviews.

Manuscript received on February 10, 2020.
Revised Manuscript received on February 20, 2020.
Manuscript published on March 30, 2020.
* Correspondence Author

Sanjay K S, Benedictine Academy,Bangalore,Karnataka, India, sanjukumsi@gmail.com
Ajit Danti, Christ(Deemed to be University),Bangalore,Karnataka, India, ajit.danti@christuniversity.in

II. LITERATURE SURVEY

Several studies can be found in the field of prediction, classification, and sentiment analysis. A classical model behavior for the perception of online customers was proposed by Singh and Sailo (2013).

Ahmed and Danti (2015) used machine learning algorithms (Multilayer Perception, SVM, and Naive Bayes) to evaluatetheir proposed method. The suggested method was based on SentiWordNet, which generated score words in seven categories for the opinion mining task. The seven categories of score words werestrong-negative, negative, weak-negative, neutral, weak-positive, positive, and strong-positive.The performance was examined for classification in terms of accuracy, precision, and recall; and the feature selection approaches were applied for sentiment analysis. Sharma and Dey (2012) analyzed 2000 reviews of a movie by using three popular sentiment feature lexicons (Opinion, GI, and HM) and five feature selection methods (Relief-F, Chi-Squared, Gain Ratio, Information Gain, and Document Frequency). Vaidya and Rafi (2014) suggested an improved technique of SentiWordNet for sentimental analysis. Gamallo and Garcia (2014) identified the polarity of English tweets through a naive-bayes classifer. The experiments evidenced that using a binary classifier between two sharp polarity categories (positive and negative) resulted in optimal functioning. Hess,Abbruzzese, Lenzi, Raber, and Abbruzzese (1999) utilized recursive partitioning to perform a multivariate analysis of survival. Turney (2002) suggested an unsupervised and simple learning algorithm for the classification of reviews as either recommended or not recommended. Saa (2016) presented a qualitative model that classified and predicted students' performance determined by the associated individual and social factors. Thakar (2015) conducted a study on students' behavior and academic performance by examining the factors hidden in the past and the present information that affected their learning process and functioning. Priyadarshini and Lakshmi (2017) conducted a survey associated with several data mining methods for the prediction of diabetes.

III. PROPOSED METHODOLOGY

Even though sentimental analysis gives scores and words through SentiWordNet, the major work is to predict human behavior through online reviews. A regression tree method and a machine learning method are applied in this work, wherein SentiWordNet is used to gather different reviews and segregate them as strong-negative, negative, neutral, positive, and strong-positive. The five score words can be used to investigate reviews on any web domain and better outcomes can be obtained by applying various machine learning methods. Web users can employ this method to identify the behavior analysis on products.

The present work has used web crawler for the extraction of online web reviews of android applications, products, and movies. Data reviews must be used for opinion mining task only after being subjected to pre-processing. The overall architecture of the proposed prediction model is shown in Figure 1.

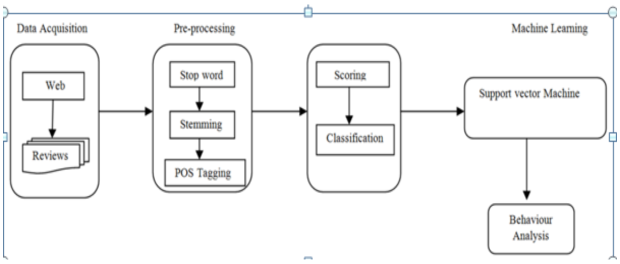

Figure 1: Architecture of the proposed Prediction Model

A.Data Acquisition

The quality of datasets is crucial and it plays a vital role in prediction analysis, wherein data acquisition is the first step. Web crawler is used to gather big data from the web for obtaining online reviews and storing them as text documents.

B. Pre-Processing

Stop words removal

Stop words (which, at, is, the, etc.) are not required for behavior prediction. Such words may create problems when they are used as phrases and some simple steps are used to filter them for the natural processing of data. More than 60 irrelevant stop words were removed for the present study.

Stemming

Informal language that include mis spelt words and internet slang is used to frame sentences in online reviews. Hence, stemming is used to remove such words for a proper retrieval of data. Stemming algorithms vary according to their accuracy and functioning. The porter stemming algorithm is used for the appropriate retrieval of data in the present study.

Parts of Speech Tagging

Parts of Speech (POS) tagging is used to tag online reviews in the form of parts of speech. Equations (1) and (2) are used to tag each term with parts of speech and the strings of words are parsed by a POS tagger.

$$T\left(w_{i,n}\right) = \arg\max max_{t1,n}\, xe^{-x^2} \prod_{i=1}^{n} p\left(w_i \middle| w_{i,i-1}\right)(1)$$

$$T\left(w_{i,n}\right) = \arg\max max_{t1,n}\, xe^{-x^2} \prod_{i=1}^{n} p(t_i|w_i)(2)$$

where ‘w’ = the word, ‘t_i’ = the word tag. The tagging issue is defined as the sequence of tags (‘t_i, n’). Stanford POS Tagger is adopted in the present study and Table 1 depicts its conversion to SentiWordNet Tags.

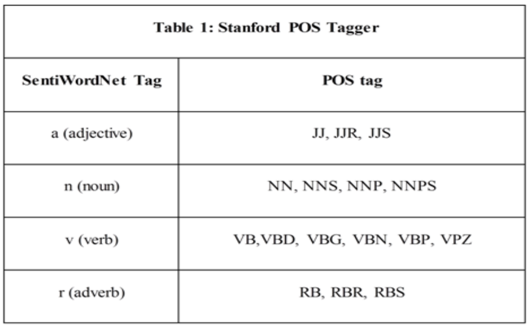

Table 1: Stanford POS Tagger

SentiWordNet Tag	POS tag
a (adjective)	JJ, JJR, JJS
n (noun)	NN, NNS, NNP, NNPS
v (verb)	VB,VBD, VBG, VBN, VBP, VPZ
r (adverb)	RB, RBR, RBS

C. Scoringand Classification

SentiWordNet

SentiWordNet is a lexical resource that describes how synsets contain positive or negative terms because two different numerical scores Pos(s) and Neg(s) are associated with WordNet synset ‘s’. If any pre-defined keyword is found by the analyzer in a product’s blog, then it receives its score from SentiWordNet for further processing .This is done after the analyzer searches the modifier associated with the keyword.

A new method is suggested for prediction in opinion mining wherein SentiWordNet is used for generating the count of scored words classified into five various categories, namely, strong-positive, positive, neutral, negative, and strong-negative. Prediction analysis is conducted by utilizing these score counts.

D. Machine Learning Approaches

A machine learning approach called support vector machine is described in this section for predicting behavior based on online reviews.

Support Vector Machine

Support Vector Machine is a statistical learning method, which can be perceived as a novel technique of using splines, neural networks, radial basis functions, polynomial functions, or other functions as the basis for training classifiers. A hyper-linear separating plane is used by support vector machines for creating a classifier.

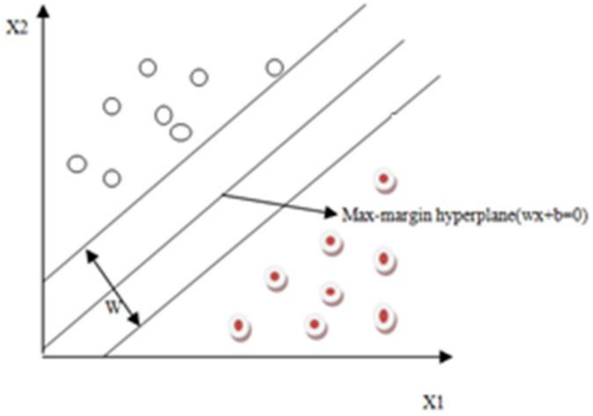

Figure 2: SVM Classifier

Max margin hyper plane is given by Equation (3).

$$g(x) = \vec{w}\vec{x} + \mathrm{b} \qquad (3)$$

where $\vec{x}$= Feature vector for particular rule
$\vec{w}$= Weight Vector
b = constant
g(x)= hyper plane value

IV. EXPERIMENTAL RESULTS

Table 2 shows the success rates of theSVMmachine learning approaches. The comparison of all machine learning approaches is shown in Table 2, which shows prediction behavior as happy, neutral, and unhappy for various aspects, based on online reviews.

Table 2: success rates of theSVMmachine learning approaches.

Machine learning Approaches	Movie			Product(Haier,AC)			Application(Android application)		
	Happy	Neutral	Un-Happy	Happy	Neutral	Un-Happy	Happy	Neutral	Un-Happy
Support Vector Machine	80.42%	8.20%	11.38%	7.79%	10%	82.21%	84.02%	12.21%	3.77%

In the present study, reviews of 2,800 users posted on a website (www.mouthshut.com) for a movie (Dangal) were used for gathering the dataset and for experimentation. Likewise, reviews of 1,900 users of an air conditioner (Haier) posted on a website (www.amazon.com) were used to gather the dataset and for experimentation. Similarly, reviews of 2,350 users of an Android application posted on a website (www.appreviews.com) were used to gather the dataset.

The efficiency of the suggested system is evidenced by the experimental outcomes. The count of scored words for online reviews of application, movie, and product datasets is depicted in Table 3.

Table 3: Count of scored words for online reviews on movie, product, and application datasets

Reviews Type(Ri)	Movie	Product(Haier,AC)	Application(Android application)
Positive(P)	50%	8%	20%
Strong_Positive(SP)	24%	4%	64%
Neutral(Neu)	16%	20%	10%
Negative(N)	8%	58%	4%
Strong_Negative(SN)	2%	10%	2%

Table 4: Behavior prediction for different reviews

Behaviour Type (Bi)	Movie	Product	Application
Happy	74%	12%	84%
Natural	16%	20%	10%
Unhappy	10%	68%	6%

According to Table 4, behavior prediction can be determined using Equation (7).

$$B_i = \text{Max}\{R_i\} \quad (7)$$

where B_i = Behavior of review type i.
R_i = Number of reviews of type i.

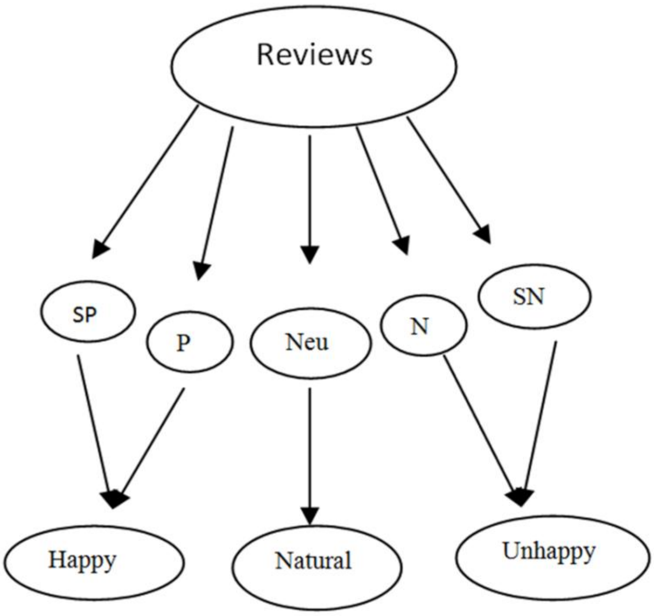

Figure 5: Classification of review behavior

It can be observed from Table 4 that the behavior predicted from online reviews is happy for movie, unhappy for product, and happy for application. Figure5 shows that review type strong positive(SP) and positive(P) is classified as Happy Behavior. Similarly, review type negative (N) and strong negative (SN) is classified as Unhappy Behavior.Review type Neutral (Neu) is classified as Natural Behavior.

V. CONCLUSION

In this paper, the prediction of behavior based on online reviews is done using machine learning approaches. Online reviews are classified into five categories,which are further used for the prediction of behavior of a person by using machine learning algorithm calleds upport vector machine (SVM). Experimental results show that SVM outperforms all other approaches.

REFERENCES

1. Singh, A. K., & Sailo, M. (2013). Consumer Behavior in Online Shopping: A Study of Aizawl. *International Journal of Business & Management Research*, *1*(3) 45-49, ISSN: 2347-4696.
2. Sharma, A., & Dey, S. (2012). Performance investigation of feature selection methods and sentiment lexicons for sentiment analysis. Special Issue of International Journal of Computer Applications (0975–8887) on Advanced Computing and Communication Technologies for HPC Applications - ACCTHPCA, 3, 15-20.
3. Saa, A. A. (2016). Educational data mining & students' performance prediction. International Journal of Advanced Computer Science and Applications, 7(5), 212-220.
4. Gamallo, P., & Garcia, M. (2014). Citius: A NaiveBayes Strategy for Sentiment Analysis on English Tweets. In Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval), Dublin, Ireland, pp. 171–175.
5. Hess, K. R., Abbruzzese, M. C., Lenzi, R., Raber, M. N., & Abbruzzese, J. L. (1999). Classification and regression tree analysis of 1000 consecutive patients with unknown primary carcinoma. Clinical Cancer Research, 5(11), 3403-3410.
6. Priyadarshini, K., & Lakshmi, I. (2017). A survey on prediction of Diabetes using data mining technique. International Journal of Innovative Research in Science, Engineering and Technology, 6(11), 369-373. ISSN(Online): 2319-8753,ISSN(Print): 2347-6710.

7. Thakar, P. (2015). Performance analysis and prediction in educational data mining: A research travelogue. arXiv preprint arXiv:1509.05176. International Journal of Computer Applications (0975 – 8887), 110(15), 60-68.
8. Ahmed, S., & Danti, A. (2015). A novel approach for Sentimental Analysis and Opinion Mining based on SentiWordNet using web data. In 2015 International Conference on Trends in Automation, Communications and Computing Technology (I-TACT-15) (pp. 1-5). IEEE.ISBN: 978-1-4673-6667-0.
9. Vaidya, S., & Rafi, M. (2014). An improved SentiWordNet for opinion mining and sentiment analysis. Journal of Advanced Database Management & Systems, 1(2), 1-7.
10. Turney P. D. (2002).Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews.In proceedings of the 40th annual meeting on association for computational linguistics, pp-417–424. Association for Computational Linguistics, Philadelphia.

AUTHORS PROFILE

Sanjay K S, MCA, (Ph.D) Having 11 Years Of Experience In Teaching And 5 Years Of Research Experience, Working As A Principal In Benedictine Academy, Bangalore.

Dr. Ajit Danti, is working as a Professor in Computer Science & Engineering Dept. at Christ University, Bangalore, India. His research areas include Image Processing, Pattern Recognition & Computer Vision. He has 29 years of experience in Academic, Research & Administration in India as well as abroad. He has published more than 150 research papers in peer reviewed International Journals & conferences. He has guided 10 PhD candidates. He has worked as a Chairman of Board of Studies & Board of Examiners under Visvesvaraya Technological University (VTU) and other Indian Universities.