THINK INDIA JOURNAL

ISSN: 0971-1260
Vol-22-Issue-3-July-2019

An Innovative Approach To Analyze Customer Purchase Behavior Through Mining

Virrat Devaser

School of Computer Science and Engineering, Lovely Professional University

DrPriyanka Chawla

School of Computer Science and Engineering, Lovely Professional University

Abstract:Customer purchasing ability refers to a customer's capacity to buy the quantities of goods. High buying ability means a customer have good appetite to purchase and have a capacity to purchase the high rate things. Low buying power means a customer cannot afford to buy high rate products having not a good income.From a sentimental analysis we will try to figure out purchase power of customer from available data sources. Sentence level sentimental analysis checks whether a sentence has negative, positive or neutral opinion. Aspect level sentimental analysis do analysis on basis of sentiments present in a file and sentiment required for analysis. In our scenario, documental level sentimental analysis has been taken into account. In machine learning mostly two techniques are used for sentimental analysis are supervised learning and unsupervised learning. In first technique, datasets are labeled and trained data is used to show reasonable output that is used for decision making. In second technique, no labeled dataset is needed and they cannot be processed easily. To solve above issue, a new approach is used in current study analyzing based on parameters.

Introduction

Sentiment analysis is used to analyzes an individual opinion towards an entities such as products, companies or their attributes Scholars and researchers applied different machine learning techniques to get the meaningful information from different opinion mining resources like twitter, Facebook and other platforms. Aspect level determines sentiment or expressions within a document and document to which it implies.In current research work, sentiment analysis is done at document level. There are generally two kinds of machine learning techniques, first one based on supervised learning and other is basedon unsupervised learning which are used in sentiment analysis. In supervised learning technique, the dataset is labeled and thus, training is done to obtain an output which help in decision making. Unlike supervised learning, unsupervised learning processes do not need any label data; therefore they cannot be handled and processedas such. To solve the issue processing of unlabeled data, clustering techniquescan be used.

Background and Terminology

Sentiment is an opinion, an attitude, thought or judgment of any by feeling. Opinion mining finds the consumers sentiments towards products or any other entities. Social networking, online

ISSN: 0971-1260
Vol-22-Issue-3-July-2019

shopping sites are the places where you can find sentiments of clients .This sentiments may be analyzed by the researchers for their specific purpose. Application programming interfaces are provided by many sites which help to collect and analyze the data and use for decision making. For example, Twitter has three API are: REST API, Streaming API and Search API.REST API is used for collecting data and information about users and consumers where Search API helps developers to access specific Twitter content for analysis. The Streaming API helps to developer for collecting real time content. So, we can relate various applications with theses API's to solve our problems. So, Sentimental analysis can be done with Social media and online shopping data from Facebook, Twitter, Amazon, Flipkart and Ebay etc. But there are some problems in data to be analyzed by sentimental analysis because opinion may not be desired to your choice because everyone has different opinion to a single thing. There is possibility of fake opinion from different fake accounts and other problems like spam data may also be present in sentimental analysis. Text data is basically words that are used for identifying columns of data such as headings, labels and names etc. Text data can contain numbers, special characters such as # or @ and letters. Numbers can be used for computations.

These days the majority of information is stored into database in the form of content. This content comes from enterprise data, manufacturing companies and differentorganizations at very high rate and in different forms. Mostly semi–organized information stored into these databases. The report may have a few to a great extent unstructured content parts for example summary of contents also some organized attributes like heading, entity name, date of birth etc.

Text mining is method of getting information out from a large content of a report, feedbacks etc. It is not easy to extract information from large content of data from different sites, PDF's, documents etc.First of all, Data is collected by different tools and technologies. This data may be structured, unstructured, semi-structured and quasi-structured. So, this data must be pre-processed by a tool to convert a suitable format which is according to tool or technologies you are using. Next, we have to use content investigation method to getting out great values from large content data. By doing, so quality data is extracted and used as a database for further analysis.It is beginning stage in which we investigate unstructured content by discovering key expressions and linkswithin content. Pattern matching is used to do this task, search for predefined sequences in content. It contains segmentation, recognizable proof of elements, sentence division, and grammatical feature task. Firstly expressions and sentences are parsed and semantically translated, afterward useful bits of data are stored in databases. This technology can be extremely helpful when managing huge quantity of content.

Classification is a method in which the content documents are arranged according to their observed similarities. Classification techniques are k-nearest neighbor classifier, Decision Tree, Naive Bayes classifier, and Support Vector Machine.Naive Bayes classification technique is mostly used because it is very easy to train and classify the data. In general, it is known as probabilistic classifier and is capable of learning documents to determine the pattern of

ISSN: 0971-1260
Vol-22-Issue-3-July-2019

document so that they have been categorized. It is comparison of contents with the list of words so that classify the documents to their correct category.For example for the given comment “He types fast” it provides Positive polarity.

Maximum entropy is based on the conditional probability distribution which maximizes the entropy. It follows the processes of naïve Bayes, providing the polarity of the data.

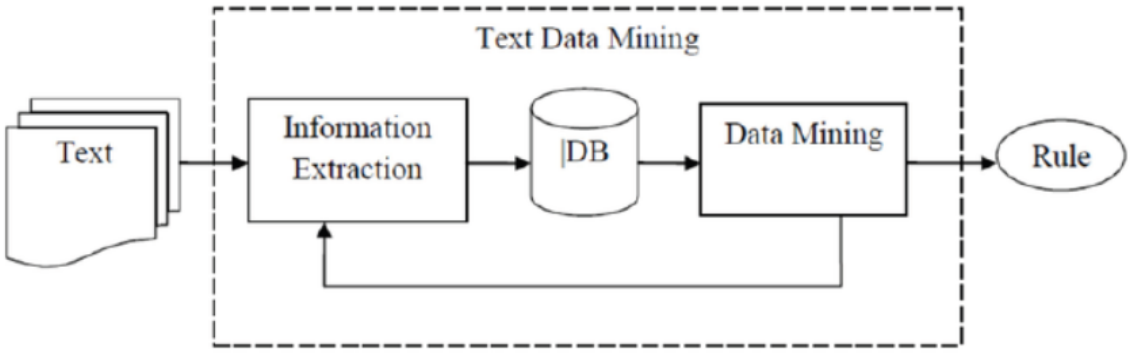

Figure 1: Information Extraction overview

Support vector machine is utilized to examine the information and characterize the choice limits and utilizes the parts for calculation which are created in info space. At that point each information spoken to as a vector is grouped in a specific class. Presently the assignment is to discover an edge between two classes that is a long way from any report. The separation characterizes the edge of the classifier, amplifying the edge lessens uncertain choices. SVM additionally underpins arrangement and relapse which are valuable for factual learning hypothesis and it helps perceiving the components correctly, that should be considered, to comprehend it effectively.

Semantic analysis is followed by training and classification of data. If two words are close to each other, they are semantically same. We have a capacity to find out the similar words. The mapping is done to examine their relationship.Clustering is a method which we make groups of similar content of documents. These groups are called clusters. Number of documents is in every cluster. The qualities of cluster view better, if the one cluster contains the more similar content rather than others.

Anshul et.al has worked on sentiment analysis to discover relationshipamong community emotion and bazaar emotion. Author proposed Self Organizing FuzzyNeural Networks (SOFNN) algorithm for classification of community emotion into 4categories that is cool, happy, attentive and mercy. Author developed his ownmethodology which is word list generation based on questionnaire. In this paper, authorasked 65 different questions to find out the current mood of user. The answers of 65questions are feed into the pre processor. They made a standard of 6

ISSN: 0971-1260
Vol-22-Issue-3-July-2019

words is calledProfile Of Mood State (POMS). These 65 words were mapped with POMS. Author used ascale rate from 1 to 5. Author proposed SentWordnet approach to extend their list byconsiders all relevant synonyms. Author used only some words to filter the tweets.Total no. of matches of all words in tweets varies every day. This algorithm worked formaximum only 140 characters. Author proposed static correlation mapping techniques tomap the score of each word and also cross validated their results by comparing the valuesreturned by algorithm and sentiment analysis technique. He compared the events likeMichael Jackson's death which was on 25th June of 2009 and thanks giving day whichwas on 26th of November 2009 and found that the value of line in graph suddenlydecreased after 25th June 2009. A naive greedy strategy to make the decision about buyand sell decision was created. Buy decision have to be taken if predicted stock value isgreater than standard deviation for next day. Sell decision are required to be taken ifpredicted stock value and standard deviation are more than the actual value at that time.[1]Hiroshi Sugimura et.al has proposes a framework in which feature patterns wereextracted. This data is extracted from fund, medical research, modern sensors, and so on.The framework extracted the features that describe similar information in database. Theyconcentrate on two parts of the characteristic design: universal and limited occurrence.The framework cut out sub successions from this data. Through utilizing grouping,numerous representatives' sequence was separated from these sub successions. Theymade a strategy that applies TF*IDF weight method to time series data, which was usedas a part of content mining. The time series information was grouped by using the gainedcharacteristic design. The characteristic designs get to be characteristics in the machinelearning and were enhanced with use of hereditary calculation. They settled on a choicetree that decided future practices. They clarified how these two instruments werecombinatory connected in the whole information disclosure process. [2]Xing Huang et.al has discovered a number of micro-blog business-related miningalgorithms which consider the connection among the expressions and the how many timesit appears in the text, and ignores the expression which is divided in a specific category,that decreased correct levels of micro-blog business-related word mining. When they compared with conventionalalgorithms, this enhanced algorithm has very much better accuracy. The speed of dataprocessing has also really enhanced by using Hadoop. The micro-blog called mini blog.They faced a hard problem in micro-blog data processing that how todiscover the precious information exactly and fast among such huge text information inmicro-blog. An expression plays a vital role on community view investigation, but also ithas a very high commercial value. So in the field of enormous social network data study,it was very necessary to pull out the expensive commercial phrase fast and precisely indistributed computing environment. [3]Xiuzhen Zhang et.al has discussed Internet business applications used reputationbased trust models to figure vendors reputation trust scores. The reputation scores wereall around high for merchants and it was troublesome for purchasers to choose reliablevendors. Purchasers could express their sentiment straightforwardly in free contentcriticism remarks. Author proposed Comment-based Multi-dimensional trust display forpurchasers to assess reputation score from client input remarks. They proposed analgorithm for mining input remarks for measurement ratings and weights, joiningstrategies of natural language processing, conclusion mining and subject displaying. eBayand Amazon information demonstrated that CommTrust could efficiently deal with "allgreat reputation" matter and level vendors adequately.eBay and Amazon reputation framework figure reputation scores for merchants wereprocessed by totaling feedback ratings. The fundamental issue with the eBay reputationadministration framework was "all great reputation" issue, where input feedback is morethan 99% positive by and large. eBay detailed

ISSN: 0971-1260
Vol-22-Issue-3-July-2019

seller rating for merchants (DSRs) on fourparts of exchanges, to be specific thing, communication, delivery time, delivery andpaying were likewise examined. The clients who leave the negative criticism theirremarks could draw in the negative ratings, in this way they harm their own reputation.They used vocabulary opinion mining strategy to separate the feeling from criticismremarks. Author proposed Latent Dirichlet Allocation (LDA) cluster strategy to processthe collected measurement ratings and weights. The entire algorithm was called Lexical-LDA. They worked fundamental in three zones was trust assessment, examination ofcriticism remarks and the motion picture review and item survey. The EigenTrustalgorithm was used to process the trust evaluations. [4]Thin Nguyen et.al has examined that online people group were used by expansivenumber of individuals to talk about emotional wellness issues. The principle thoughtprocess to concentrate the attributes of online sadness groups (CLINICAL) in correlationwith those joining other online groups (CONTROL). They used machine learning andstatistical techniques to segregate online mail amongst depressions and control groups viastate of mind, psycholinguistic procedures and substance subjects pulled out from postscreated via individuals from groups. Each and every one viewpoints counting frame ofmind, composed substance and composing manner are observed to be essentially uniqueamong two groups. Opinion examination demonstrated the clinical gathering have beensmall valence than individuals in the group. They demonstrated great analytical strengthin depression order utilizing points and psycholinguistic pieces of information likecomponents. Unbiased understanding between composing styles and substance, greatvalidity influence was a vital stride in comprehension web-based community networkingand their utilization in emotional well-being. They have explored online depressiongroups and concentrated their differential essentials to other online groups. Underlying themes were bringing into to have more noteworthyextrapolative influence than language characteristics for expectation of despair groups.[5]

Research Methodology

In the paper we have designed a RFM technique as well as naive Bayes classification that compute the purchasing power of customers using sentiment analysis. This study will help to improve the sales for a business entity. Further this study will help to maintain the prime customers as well as finding the customers having high purchasing power.Main objectives of this research work are to predict customer purchasing power depending upon various factors like recent purchases, frequency of visit, monetary strength and predicting upcoming sales.The proposed system uses the RFM method for the prediction of customer purchasing power. The inputs to the system are customer ID, start date, end date, amount, recency, monetary, frequency and output is total purchasing power score. In RFM method, recency means how a customer recently visited, frequency means how many times the customer visited and monetary means how much a customer spends the money per deal in RFM method some ranges are set to calculate the total purchasing power of each customers.

ISSN: 0971-1260
Vol-22-Issue-3-July-2019

No.	Amount	Classification label
1.	250-500	High
2.	0-249	Low

Table1: Classification of Customer Purchasing Power

The purposed system also used the naïve Bayes classification method to calculate the positive and negative reviews of customers.

Naive Bayes (NB) Method:

It is used for both classifications as well as training. This is a probabilistic classifier method based on Bayes' theorem.

Following steps were followed:

Step1: Collect the data for customers as well as reviews of customers.

Step2: After insert the data set into Rstudio, it is pre-processed for better decisions. Pre-processing remove the duplicate entries from data as well as fill the missing value in customer purchasing data. In the review data set it is pre-processed by removing the stop word, numerical data and special characters etc.

Step3: RFM technique is applied to calculate the purchasing power of customers. Vectorization is applied on the pre-processed review data to get the required data.

Step4: In the parallel process the naive Bayes classification technique is also applied for classification of reviews.

Step5: classification gives the result of positive and negative polarity according to the purchasing power of customers.

Results and Discussion

Using the Proposed technique named as RFM results were obtained and compared with standard procedures like Naïve Bayes technique.

ISSN: 0971-1260
Vol-22-Issue-3-July-2019

```
> head(df)
    ID       Date Amount Recency Frequency Monetary
4    4 1997-12-12  26.48     201         4   25.125
6   21 1997-01-13  11.77     534         2   37.555
7   50 1997-01-01   6.79     546         1    6.790
8   71 1997-01-01  13.97     546         1   13.970
9   86 1997-01-01  23.94     546         1   23.940
25 111 1998-06-20  55.47      11        16   69.190
```

Figure 2: Snapshot of sample data set

This snapshot shows the result of attributes which required for analysis of customer purchasing power.

```
> # Use the NB classifier we built to make predictions on the test set.
> system.time( pred <- predict(classifier, newdata=testNB) )
   user  system elapsed
 359.25    0.14  362.11
```

PURPOSED APPROACH	
1.Target	To predict the purchasing power of customers
2.Parameters	Recency, frequency ,monetary
3.Technique Used	RFM
4.Result	94.117

Table 2: Purposed Work Details

CONCLUSION

In this research, a system is purposed to predict the purchasing power of the customers by considering different attributes .These factors can affect the purchasing power of the customers. This prediction system tries to estimate the purchasing power of customers by considering a threshold value which denotes the purchasing power of customers. If the customer purchasing

ISSN: 0971-1260
Vol-22-Issue-3-July-2019

power value is above than threshold value, then the customer has higher purchasing power otherwise the customer has lower purchasing power.

Opinion mining is also used to find the customers opinion about various bought products. Sentiment classification is used to classify various customers who have positive and negative opinion about a bought product. Sentiment classification is under taken by gathering the reviews of customers. Reviews are unstructured in nature that cannot provide the important information. Opinion mining is used to extract the important information from reviews that can help the customer management relationship in future.

The purposed system using Naive Bayes and RFM technique for data mining and the accuracy of the system by using these techniques is 94.11.

References

[1] Mittal, Anshul, and ArpitGoel. "Stock prediction using twitter sentimentanalysis."15 (2012).

[2] Sugimura, Hiroshi, and Kazunori Matsumoto. "Classification system for timeseries data based on feature pattern extraction." Systems, Man, and Cybernetics(SMC), 2011 IEEE International Conference on.IEEE, 2011.

[3] Huang, Xing, and Qing Wu. "Micro-blog commercial word extraction based onimproved tf-idf algorithm." TENCON 2013-2013 IEEE Region 10 Conference(31194).IEEE, 2013.

[4] Zhang, Xiuzhen, Lishan Cui, and Yan Wang. "Commtrust: Computing multidimensionaltrust by mining e-commerce feedback comments." IEEE Transactionson Knowledge and Data Engineering 26.7 (2014): 1631-1643.

[5] Nguyen, Thin, et al. "Affective and content analysis of online depressioncommunities." IEEE Transactions on Affective Computing 5.3 (2014): 217-226.

[6] Rosa, Renata L., Demsteneso Z. Rodriguez, and GraçaBressan. "Musicrecommendation system based on user's sentiments extracted from socialnetworks." IEEE Transactions on Consumer Electronics 61.3 (2015): 359-367.

[7] Li, Yuefeng, et al. "Relevance feature discovery for text mining." IEEETransactions on Knowledge and Data Engineering 27.6 (2015): 1656-1669.

[8] Gatti, Lorenzo, Marco Guerini, and Marco Turchi. "SentiWords: Deriving a HighPrecision and High Coverage Lexicon for Sentiment Analysis." IEEETransactions on Affective Computing 7.4 (2016): 409-421.

ISSN: 0971-1260
Vol-22-Issue-3-July-2019

[9] Bouazizi, Mondher, and TomoakiOhtsuki. "Sentiment analysis: From binary tomulti-class classification: A pattern-based approach for multi-class sentimentanalysis in Twitter." Communications (ICC), 2016 IEEE International Conferenceon.IEEE, 2016.

[10] Pang, Bo, and Lillian Lee. "A sentimental education: Sentiment analysis usingsubjectivity summarization based on minimum cuts." Proceedings of the 42nd annual meeting on Association for Computational Linguistics.Association forComputational Linguistics, 2004.

[11] Salvetti, Franco, Stephen Lewis, and ChristophReichenbach. "Automatic opinionpolarity classification of movie." Colorado research in linguistics 17.1 (2004):2.

[12] Beineke, Philip, Trevor Hastie, and ShivakumarVaithyanathan. "The sentimentalfactor: Improving review classification via human-providedinformation." Proceedings of the 42nd annual meeting on association forcomputational linguistics.Association for Computational Linguistics, 2004.

[13] Mullen, Tony, and Nigel Collier. "Sentiment Analysis using Support VectorMachines with Diverse Information Sources."EMNLP.Vol. 4. 2004.

[14] Dave, K. , Lawrence, S. , &Pennock, D. M. (2003). Mining the peanut gallery:opinion extraction and semantic classification of product reviews. In Proceedingsof the 12th nternational conference on World Wide Web (pp. 519–528). ACM .

[15] Matsumoto, Shotaro, HiroyaTakamura, and Manabu Okumura. "Sentimentclassification using word sub-sequences and dependency sub-trees." Advances inKnowledge Discovery and Data Mining (2005): 21-32.

[16] Zhang, Dongwen, et al. "Chinese comments sentiment classification based onword2vec and SVM perf." Expert Systems with Applications 42.4 (2015): 1857-