Sentiment Analysis using Rapid Miner

Aravindasamy. R ,C. Nalini, Sangeetha.S, S. Theivasigamani

Abstract: ***Now a day the data grows day by day so data mining replaced by big data. Under data mining, Text mining is one of the processes of deriving structured or quality information or data from text document. It helps to business for finding valuable knowledge. Sentiment analysis is one of the applications in text mining. In sentiment analysis, determine the emotional tone under the text. It is the major task of natural language processing. The objective of this paper to categorize the document in sentence level and review level, and classification techniques applied on the dataset (electronic product data). There is an ensemble number of classification techniques applied on the dataset. Then compare each techniques, based on various parameters and find out which one is best. According to that give better suggestions to the company for improving the product.***

I. INTRODUCTION

Web mining is another technique that handles the web data. It derives the quality data from web document. It improves the efficiency of web search engine by identifying the web pages and classifying the web documents. The main application of web mining is e commerce sites. It mainly divided into three types: content mining, structure mining, and usage mining. Web content mining is used to extract the text data from web document. Web structure mining is used to mine the link structure of hyperlink and it identifies the pages are either linked by information or direct link. Web usage mining mainly focuses on web log records or history records[1],[3],[5]

Text mining is the process of extracting or deriving high quality information from unstructured content. The applications of text mining are, information extraction(IE), natural language processing(NLP), data mining(DM), information retrieval(IR). The main advantages of text mining are, the companies use this technique for knowing better customer relationship. [2],[4],[6]

Sentiment analysis is one of the applications in text mining[7],[9] ,[11]

. The amount of data that is generated from this website will increases. Based on opinions the organizations will take decisions. Peoples also take recommendations from the others before purchasing a product. Every compan y get the benefit from the customer review. They analyse the reviews and decide what are the steps will take for increasing profit. Reviews are collection of sentences and each sentence has aspects and sentiments associated with it. [8],[10] ,[12]

The objective of this paper mainly focuses on sentiment analysis of a document. The sentences are categorized into sentence level and review level and applied classification techniques like KNN, naïve bayes, decision tree on the dataset. The sentence level categorization means classify the sentences into positive sentence, negative sentence and neutral sentence based on the sentiments conveyed. The review level categorization means classify the sentences based on rating of the reviews. The star rating used to classify the products. Based on that, company can improve the product. The 1 star, 2 star used for negative sentence, 3 star create neutral effect and 4 star and 5 star used for positive sentence. After analysing the result, compare various parameters and find out which one is giving the best result. According to that given better suggestions to the company for improving the product. [13], [15] ,[17]

II. LITERATURE SURVEY

Kim Schouten and Flavius Frasincar [2018]: This survey paper focuses on aspect-level sentiment analysis, where the objective of this paper is to find and aggregate sentiment on entities within documents or aspects of them. The paper explains how to aspect level sentiment analysis is done on the document. And it uses machine learning to model the languages because it is non-random and very complex. [14],[16], [18]

Xing Fang and Justin Zhan [2015]: This paper managing the problem of sentiment polarity categorization. They used the online product review dataset collected from Amazon.com. Experiment will do on both sentence-level and review-level. The experiment done on the scikit, is an open source tool written in python. They compare the techniques naïve bayes, random forest and

Revised Manuscript Received on July 22, 2019.

Aravindasamy R,Student,Department of CSE, Bharath Institute of Higher Education and Research, Chennai, Tamilnadu, India
C Nalini Department of CSE, Bharath Institute of Higher Education and Research, Chennai, Tamilnadu, India.
Sangeetha.S Department of CSE, Bharath Institute of Higher Education and Research, Chennai, Tamilnadu, India.
S. Theivasigamani Department of CSE, Bharath Institute of HigherEducation and Research, Chennai, Tamilnadu, India.

SVM. [19],[21],[23]Deb Dutta Das, Sharan Sharma, Shubham Natani, Neelu Khare and Brijendra Singh [2017]: The paper uses twitter data for sentiment analysis. They did the sentiment analysis using R and Rapid Miner. The tweets are derivers and classified into neutral, negative and positive sentiments. The naïve bayes algorithm is used by the two software. Both result is analysed and finally conclude which tool give better result. [20],[22], [24]

Md Shoeb, Jawed Ahmed [2017]: This paper, done sentiment analysis on twitter dataset. They collect the dataset and applied text mining techniques and perform sentiment analysis. This paper used the rapid miner tool and applied classification techniques. This compared three classification techniques and find which one give the better accuracy. [25],[27],[29]

Sindhu C Dyawanapally Veda Vyas Kommareddy Pradyoth [2017]: The paper uses sentiment analysis then pre-processed the data and the reviews are classified according to their polarity confidence. They used the rapid miner tool for review processing and compare the algorithms SVM and Naïve Bayes. In future they planned to use decision tree and KNN. [26],[28],[30]

III. METHODOLOGY

A. Module Description

Dataset: Electronic product data

The dataset is downloaded from Data.world.com. The list involves 7000 online reviews of 50 electronic products from websites like Amazon and Best Buy provided by Datafiniti's product database. The dataset include 26 attributes like id, brand, colours, review rating, review text, review title etc. The paper uses review text and the sentiment analysis are done. Based on that, it is classified into positive, negative sentences.

Data Pre-processing:

Data preprocessing means cleans the data.The steps during data pre-processing are,

Data cleaning: The unwanted data is cleaned by replacing missing values or removing the data.

Data Integration: Different types of data are combined together and problems are solved.

Data Transformation: Data is converted to another format. Data reduction: The size of data is reduced and used in data warehouse. [31],[33],[35]

Sentiment sentence extraction:

The subjective content is extracted and done sentiment analysis on that. The subjective content consist all type of sentiment sentences. The sentiment sentences must include positive or negative word. [32],[34],[36]

Part of speech tagging:

Each sentence has its own meaning by using the words. In English, there are 8 part of speech like the noun, verb, pronoun, adjective, adverb, preposition, conjunction and interjection. The POS tagger filtered out noun and pronoun.

Removal of implicit statements:

The statements consists neutral words. So such types of words are deleted from the sentences. The identification of neutral words improves overall accuracy.

Sentiment phrase identification:

Each phrase has different meaning in different sentence. The sentiment phrase depends on the situation of the sentence. There are two types of phrases, verbal phrase and adjective phrase. After removing the neutral words, the phrases should be identified. [37],[39],[41]

Sentiment score computing:

The token is a word or phrase. In sentiment score computing calculate, "how many times the word will occur in a sentence".

Feature vector generation:

The tokens are derived from the product dataset. These tokens are known as a feature, which is used for sentiment categorization. Each data converted to vector and known as feature vector generation. In sentence level categorization, vector is created based on sentence. [38],[40],[42]

Document categorization:

The document is categorized into two levels, sentence and review levels. The sentence level means the document classified into positive sentence, negative sentence, and neutral. The review level categorization done based on ratings of the product. The 1 star, 2 star used for negative impact of the product. The 3 star creates neutral effects and 4 stars, 5 stars are for positive effect of the product.

Classification techniques:

The classification is a data mining technique, which assign the data into predefined class. The commonly used classification algorithms are decision tree, rule based techniques, neural networks and Bayesian classification. The various parameters in classification technique are,

Accuracy: Accuracy is defined as the amount of assumptions is correct.

(Number of data correct)/(Total number of data)

Precision: Precision the percentage of the correct result.

(True Positive)/(Total corrct data)

Recall: Recall is defined as the quantity of the dataset.

B. Classification Techniques

Naïve bayes classifier:

The naïve bayes classification follows Bayes' theorem with strong (naive) independent assumptions. It is the descriptive term under probability model. The advantage of the algorithm is small amount of dataset needed. The entire matrix not determined; only the some variables are considered.

K nearest neighbour:

This is another classification algorithm. The data is classified by the nearest neighbours and it is assigned to the class. The value of K is one and data assigned to nearest neighbour. K is typically small and positive integer. The distance functions include,

Euclidean Distance: √ (∑_(i=1)^n▒ 〖(X〗 i)-Yi)2

Manhattan Distance: ∑_(k=0)^n▒ 〖| X k-Yk|2〗

Decision Tree:

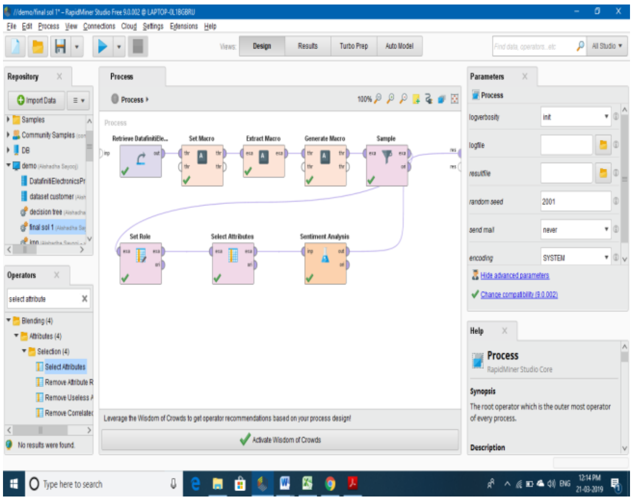
This algorithm is used for machine learning. This technique includes classification and regression methods. Decision tree is visually represented the decisions. It uses tree like model of decisions where it consist root node and child node. The decision rule is:

IF
{condition 1 and condition 2
THEN
The outcome
}

IV. EXPERIMENTAL RESULT

Rapid miner Workflow:

A. Sentiment analysis


//demo/final sol 1* – RapidMiner Studio Free 9.0.002 @ LAPTOP-0L1BGBRU
File Edit Process View Connections Cloud Settings Extensions Help
Views: Design Results Turbo Prep Auto Model
Repository
Process
Retrieve DatafiniteEle...
Set Macro
Extract Macro
Generate Macro
Sample
Set Role
Select Attributes
Sentiment Analysis
Operators
Help

Parameters: Process

logverbosity: init
logfile:
resultfile:
random seed: 2001
send mail: never
encoding: SYSTEM

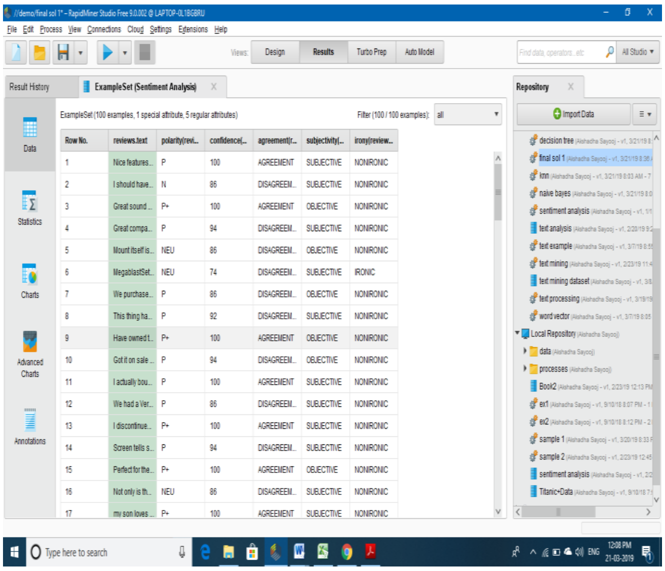
Hide advanced parameters
Change compatibility (9.0.002)

Process
RapidMiner Studio Core

Synopsis
The root operator which is the outer most operator of every process.

Description

Leverage the Wisdom of Crowds to get operator recommendations based on your process design!
Activate Wisdom of Crowds

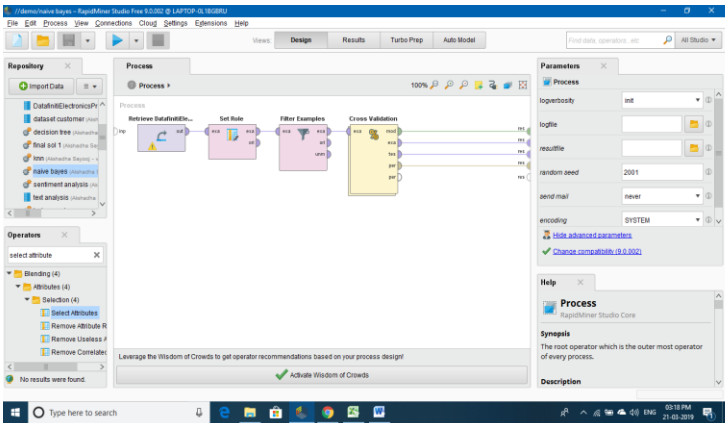

//demo/final sol 1* – RapidMiner Studio Free 9.0.002 @ LAPTOP-0L1BGBRU
File Edit Process View Connections Cloud Settings Extensions Help
Views: Design Results Turbo Prep Auto Model
Result History
ExampleSet (Sentiment Analysis)
Data
Statistics
Charts
Advanced Charts
Annotations
Repository
Import Data
Type here to search

ExampleSet (100 examples, 1 special attribute, 5 regular attributes)

Filter (100 / 100 examples): all

Row No.	reviews.text	polarity(revi...	confidence(...	agreement(r...	subjectivity(...	irony(review...
1	Nice features...	P	100	AGREEMENT	SUBJECTIVE	NONIRONIC
2	I should have...	N	86	DISAGREEM...	SUBJECTIVE	NONIRONIC
3	Great sound ...	P+	100	AGREEMENT	OBJECTIVE	NONIRONIC
4	Great compa...	P	94	DISAGREEM...	SUBJECTIVE	NONIRONIC
5	Mount itself is...	NEU	86	DISAGREEM...	OBJECTIVE	NONIRONIC
6	MegablastSet...	NEU	74	DISAGREEM...	SUBJECTIVE	IRONIC
7	We purchase...	P	86	DISAGREEM...	OBJECTIVE	NONIRONIC
8	This thing ha...	P	92	DISAGREEM...	SUBJECTIVE	NONIRONIC
9	Have owned t...	P+	100	AGREEMENT	OBJECTIVE	NONIRONIC
10	Got it on sale ...	P	94	DISAGREEM...	OBJECTIVE	NONIRONIC
11	I actually bou...	P	100	AGREEMENT	SUBJECTIVE	NONIRONIC
12	We had a Ver...	P	86	DISAGREEM...	SUBJECTIVE	NONIRONIC
13	I discontinue...	P+	100	AGREEMENT	SUBJECTIVE	NONIRONIC
14	Screen tells s...	P	94	DISAGREEM...	SUBJECTIVE	NONIRONIC
15	Perfect for the...	P+	100	AGREEMENT	OBJECTIVE	NONIRONIC
16	Not only is th...	NEU	86	DISAGREEM...	SUBJECTIVE	NONIRONIC
17	my son loves ...	P+	100	AGREEMENT	SUBJECTIVE	NONIRONIC

B. CLASSIFICATION TECHNIQUES:



Parameters: Process

logverbosity: init
logfile:
resultfile:
random seed: 2001
send mail: never
encoding: SYSTEM

Hide advanced parameters
Change compatibility (9.0.002)

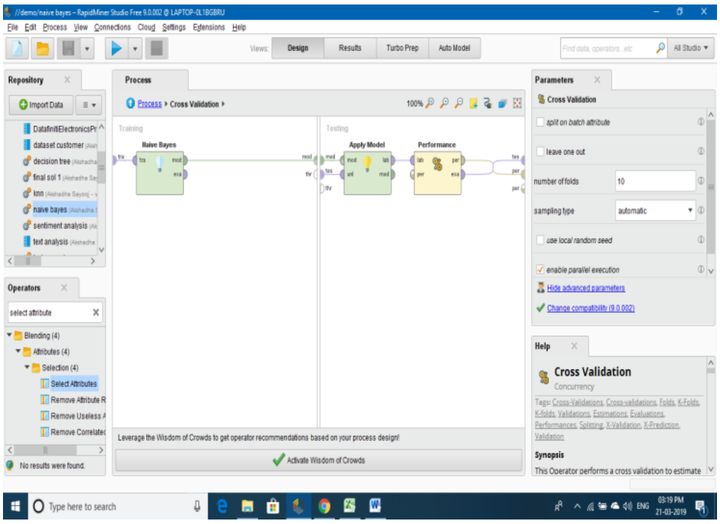
Process
RapidMiner Studio Core

Synopsis
The root operator which is the outer most operator of every process.

Description

Leverage the Wisdom of Crowds to get operator recommendations based on your process design!
Activate Wisdom of Crowds

Naïve bayes


//demo/naive bayes – RapidMiner Studio Free 9.0.002 @ LAPTOP-0L1BGBRU
File Edit Process View Connections Cloud Settings Extensions Help
Views: Design Results Turbo Prep Auto Model
Repository
Process ▸ Cross Validation ▸
Training
Naive Bayes
Testing
Apply Model
Performance
Operators
select attribute
No results were found.
Type here to search

Parameters: Cross Validation

□ split on batch attribute
□ leave one out
number of folds: 10
sampling type: automatic
□ use local random seed
☑ enable parallel execution

Hide advanced parameters
Change compatibility (9.0.002)

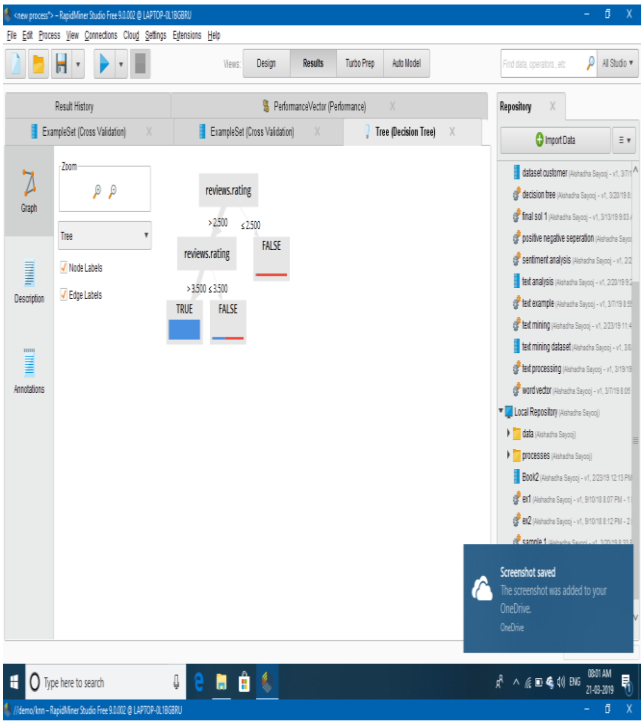
Cross Validation
Concurrency

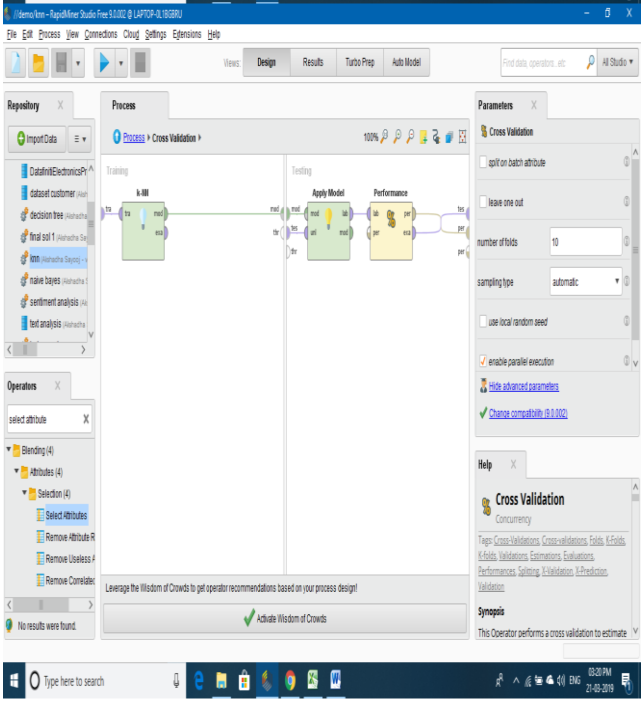
Tags: Cross-Validations, Cross-validations, Folds, K-Folds, K-folds, Validations, Estimations, Evaluations, Performances, Splitting, X-Validation, X-Prediction, Validation

Synopsis
This Operator performs a cross validation to estimate

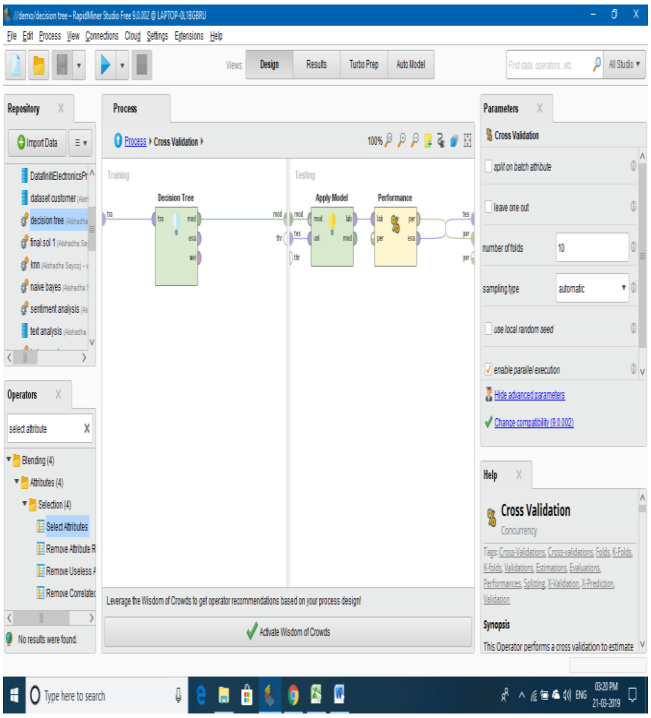
Leverage the Wisdom of Crowds to get operator recommendations based on your process design!
Activate Wisdom of Crowds

K nearest neighbour





Decision tree



Naïve Bayes:

	trueTRUE	trueFALSE	Class precision
Pred.TRUE	761	29	96.33%
Pred.FALSE	208	111	34.80%
Class recall	78.53%	79.29%	

K Nearest neighbour:

	trueTRUE	trueFALSE	Class precision
Pred.TRUE	938	16	98.38%
Pred.FALSE	31	124	80.00%
Class recall	96.80%	88.57%	

Decision Tree:

	trueTRUE	trueFALSE	Class precision
Pred.TRUE	933	8	99.15%
Pred.FALSE	36	132	78.57%
Class recall	96.28%	94.29%	

Accuracy:

Classification Techniques	Accuracy
Naïve bayes	78.63%
K nearest neighbour	95.76%
Decision Tree	96.03%

Retrieval Number: I33320789S319/2019©BEIESP
DOI: 10.35940/ijitee.I3332.0789S319

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

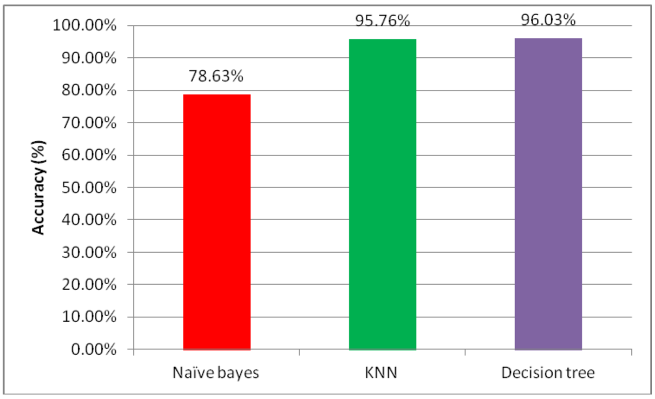


Fig 1: Accuracy chart

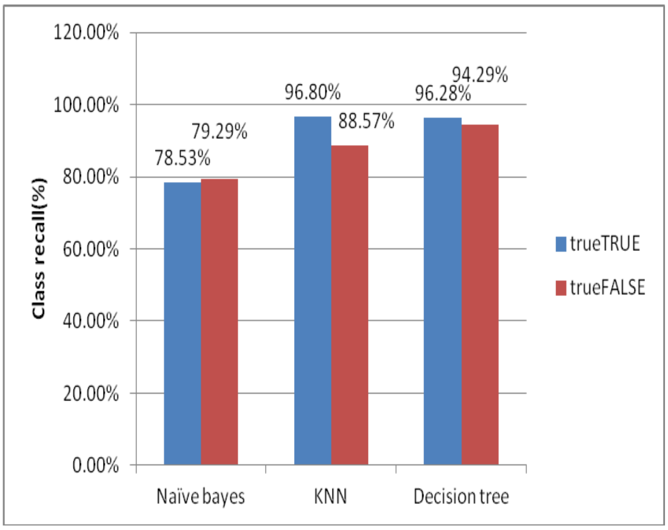


Fig 2: Class Recall

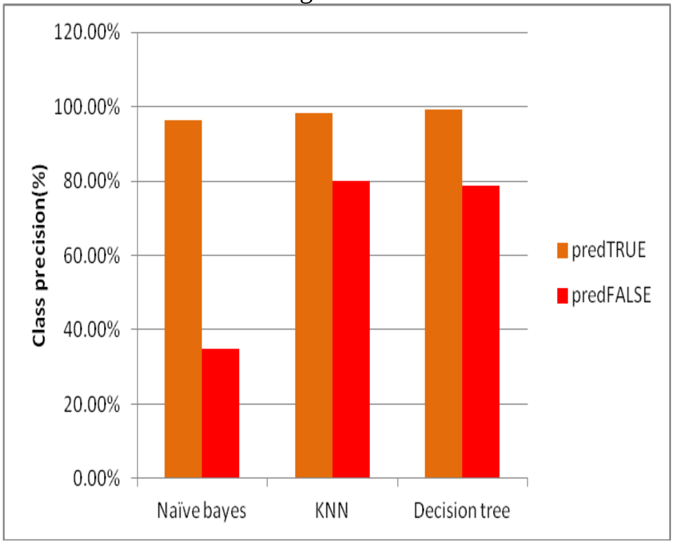


Fig 3: Class Precision

From the experimental result, the decision tree have highest accuracy rate than naïve bayes and k nearest neighbour. So this algorithm has the highest quality and handles the dataset. Therefore it is considered as best classification algorithm.

V. CONCLUSION

The paper deals with the sentiment analysis for different level categorization using rapid miner tool. The text document is categorized into sentence level and review level, and classification techniques applied on the dataset (electronic product data). There is an ensemble number of classification techniques applied on the dataset. Then compare each techniques, based on various parameters and find out which one is best. According to that give better suggestions to the company for improving the product. From the experimental result it concluded that decision tree give highest accuracy than others. So decision tree is the best classification algorithm.

REFERENCES

[1] Kumaravel A., Rangarajan K.,Algorithm for automaton specification for exploring dynamic labyrinths,Indian Journal of Science and Technology,V-6,I-SUPPL5,PP-4554-4559,Y-2013

[2] P. Kavitha, S. Prabakaran “A Novel Hybrid Segmentation Method with Particle Swarm Optimization and Fuzzy C-Mean Based On Partitioning the Image for Detecting Lung Cancer” International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249-8958, Volume-8 Issue-5, June 2019

[3] Kumaravel A., Meetei O.N.,An application of non-uniform cellular automata for efficient cryptography,2013 IEEE Conference on Information and Communication Technologies, ICT 2013,V-,I-,PP-1200-1205,Y-2013

[4] Kumarave A., Rangarajan K.,Routing alogrithm over semi-regular tessellations,2013 IEEE Conference on Information and Communication Technologies, ICT 2013,V-,I-,PP-1180-1184,Y-2013

[5] P. Kavitha, S. Prabakaran “Designing a Feature Vector for Statistical Texture Analysis of Brain Tumor” International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249-8958, Volume-8 Issue-5, June 2019

[6] Dutta P., Kumaravel A.,A novel approach to trust based identification of leaders in social networks,Indian Journal of Science and Technology,V-9,I-10,PP--,Y-2016

[7] Kumaravel A., Dutta P.,Application of Pca for context selection for collaborative filtering,Middle - East Journal of Scientific Research,V-20,I-1,PP-88-93,Y-2014

[8] Kumaravel A., Rangarajan K.,Constructing an automaton for exploring dynamic labyrinths,2012 International Conference on Radar, Communication and Computing, ICRCC 2012,V-,I-,PP-161-165,Y-2012

[9] P. Kavitha, S. Prabakaran “Adaptive Bilateral Filter for Multi-Resolution in Brain Tumor Recognition” International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8 Issue-8 June, 2019

[10] Kumaravel A.,Comparison of two multi-classification approaches for detecting network attacks,World Applied Sciences Journal,V-27,I-11,PP-1461-1465,Y-2013

[11] Tariq J., Kumaravel A.,Construction of cellular automata over hexagonal and triangular tessellations for path planning of multi-robots,2016 IEEE International Conference on Computational Intelligence and Computing Research, ICCIC 2016,V-,I-,PP--,Y-2017

[12] Sudha M., Kumaravel A.,Analysis and measurement of wave guides using poisson method,Indonesian Journal of Electrical Engineering and Computer Science,V-8,I-2,PP-546-548,Y-2017

[13] Ayyappan G., Nalini C., Kumaravel A.,Various approaches of knowledge transfer in academic social network,International Journal of Engineering and

Technology,V-,I-,PP-2791-2794,Y-2017

[14] Kaliyamurthie, K.P., Sivaraman, K., Ramesh, S. Imposing patient data privacy in wireless medical sensor networks through homomorphic cryptosystems 2016, Journal of Chemical and Pharmaceutical Sciences 9 2.

[15] Kaliyamurthie, K.P., Balasubramanian, P.C. An approach to multi secure to historical malformed documents using integer ripple transfiguration 2016 Journal of Chemical and Pharmaceutical Sciences 9 2.

[16] A.Sangeetha,C.Nalini,"Semantic Ranking based on keywords extractions in the web", International Journal of Engineering & Technology, 7 (2.6) (2018) 290-292

[17] S.V.GayathiriDevi,C.Nalini,N.Kumar,"An efficient software verification using multi-layered software verification tool "International Journal of Engineering & Technology, 7(2.21)2018 454-457

[18] C.Nalini,ShwtambariKharabe,"A Comparative Study On Different Techniques Used For Finger – Vein Authentication", International Journal Of Pure And Applied Mathematics, Volume 116 No. 8 2017, 327-333, Issn: 1314-3395

[19] M.S. Vivekanandan and Dr. C. Rajabhushanam, "Enabling Privacy Protection and Content Assurance in Geo-Social Networks", International Journal of Innovative Research in Management, Engineering and Technology, Vol 3, Issue 4, pp. 49-55, April 2018.

[20] Dr. C. Rajabhushanam, V. Karthik, and G. Vivek, "Elasticity in Cloud Computing", International Journal of Innovative Research in Management, Engineering and Technology, Vol 3, Issue 4, pp. 104-111, April 2018.

[21] K. Rangaswamy and Dr. C. Rajabhushanamc, "CCN-Based Congestion Control Mechanism In Dynamic Networks", International Journal of Innovative Research in Management, Engineering and Technology, Vol 3, Issue 4, pp. 117-119, April 2018.

[22] Kavitha, R., Nedunchelian, R., "Domain-specific Search engine optimization using healthcare ontology and a neural network backpropagation approach", 2017, Research Journal of Biotechnology, Special Issue 2:157-166

[23] Kavitha, G., Kavitha, R., "An analysis to improve throughput of high-power hubs in mobile ad hoc network" , 2016, Journal of Chemical and Pharmaceutical Sciences, Vol-9, Issue-2: 361-363

[24] Kavitha, G., Kavitha, R., "Dipping interference to supplement throughput in MANET" , 2016, Journal of Chemical and Pharmaceutical Sciences, Vol-9, Issue-2: 357-360

[25] Michael, G., Chandrasekar, A.,"Leader election based malicious detection and response system in MANET using mechanism design approach", Journal of Chemical and Pharmaceutical Sciences(JCPS) Volume 9 Issue 2, April - June 2016 .

[26] Michael, G., Chandrasekar, A.,"Modeling of detection of camouflaging worm using epidemic dynamic model and power spectral density", Journal of Chemical and Pharmaceutical Sciences(JCPS) Volume 9 Issue 2, April - June 2016 .

[27] Pothumani, S., Sriram, M., Sridhar, J., Arul Selvan, G., Secure mobile agents communication on intranet,Journal of Chemical and Pharmaceutical Sciences, volume 9, Issue 3, Pg No S32-S35, 2016

[28] Pothumani, S., Sriram, M., Sridhar , Various schemes for database encryption-a survey, Journal of Chemical and Pharmaceutical Sciences, volume 9, Issue 3, Pg NoS103-S106, 2016

[29] Pothumani, S., Sriram, M., Sridhar, A novel economic framework for cloud and grid computing, Journal of Chemical and Pharmaceutical Sciences, volume 9, Issue 3, Pg No S29-S31, 2016

[30] Priya, N., Sridhar, J., Sriram, M. "Ecommerce Transaction Security Challenges and Prevention Methods- New Approach" 2016 ,Journal of Chemical and Pharmaceutical Sciences, JCPS Volume 9 Issue 3.page no:S66-S68 .

[31] Priya, N.,Sridhar,J.,Sriram, M."Vehicular cloud computing security issues and solutions" Journal of Chemical and Pharmaceutical Sciences(JCPS) Volume 9 Issue 2, April - June 2016

[32] .

[33] Priya, N., Sridhar, J., Sriram, M. "Mobile large data storage security in cloud computing environment-a new approach" JCPS Volume 9 Issue 2. April - June 2016

[34] Anuradha.C, Khanna.V, "Improving network performance and security in WSN using decentralized hypothesis testing "Journal of Chemical and Pharmaceutical Sciences(JCPS) Volume 9 Issue 2, April - June 2016 .

[35] Anuradha.C, Khanna.V, "A novel gsm based control for e-devices" Journal of Chemical and Pharmaceutical Sciences(JCPS) Volume 9 Issue 2, April - June 2016 .

[36] Anuradha.C, Khanna.V, "Secured privacy preserving sharing and data integration in mobile web environments " Journal of Chemical and Pharmaceutical Sciences(JCPS) Volume 9 Issue 2, April - June 2016 .

[37] Sundarraj, B., Kaliyamurthie, K.P. Social network analysis for decisive the ultimate classification from the ensemble to boost accuracy rates 2016 International Journal of Pharmacy and Technology 8

[38] Sundarraj, B., Kaliyamurthie, K.P. A content-based spam filtering approach victimisation artificial neural networks 2016 International Journal of Pharmacy and Technology 8 3.

[39] Sundarraj, B., Kaliyamurthie, K.P. Remote sensing imaging for satellite image segmentation 2016 International Journal of Pharmacy and Technology 8 3.

[40] Sivaraman, K., Senthil, M. Intuitive driver proxy control using artificial intelligence 2016 International Journal of Pharmacy and Technology 8 4.

[41] Sivaraman, K., Kaliyamurthie, K.P. Cloud computing in mobile technology 2016 Journal of Chemical and Pharmaceutical Sciences 9 2.

[42] Sivaraman, K., Khanna, V. Implementation of an extension for browser to detect vulnerable elements on web pages and avoid click jacking 2016 Journal of Chemical and Pharmaceutical Sciences 9 2.

AUTHORS PROFILE

Aravindasamy R, Student, Department of Computer Science & Engineering, Bharath Institute of Higher Education and Research, Chennai, India

C.Nalani, Assistant Professor, Department of Computer Science & Engineering, Bharath Institute of Higher Education and Research, Chennai, India

Sangeetha.S, Assistant Professor, Department of Computer Science & Engineering, Bharath Institute of Higher Education and Research, Chennai, India

S. Theivasigamani Assistant Professor, Department of Computer Science & Engineering, Bharath Institute of Higher Education and Research, Chennai, India