International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169
Volume: 3 Issue: 8 5167 - 5172

Sentiment Analysis using Rapid Miner for Polarity Dataset

Dr.Siddhartha Ghosh
HOD of CSE Department, KMIT
KMIT
Hyderabad, India
siddhartha@kmit.in

Sujata M.Thamke
CSE Department, KMIT
KMIT
Hyderabad, India
sujata.thamke@gmail.com

U.R.S Kalyani
IT Department, KMIT
KMIT
Hyderabad, India
upadhyayula.kalyani@gmail.com

***Abstract*—** Usage of social media like whatsapp, facebook, twitter, blogs etc is increasing day by day which makes every people to feel free to comment and share their views, opinions and suggestions which can be either positive, negative or neutral comments on various topics like politics, business, advertisement, entertainment etc. This may contain likes, dislikes, good, bad or Emotions etc which are nothing but some type of sentiments. Judging these sentiments helps to find out whether the given sentiment is positive, negative or neutral by using sentiment analysis. In this paper we are discussing about the concept of polarity in sentiment analysis by using polarity movie review dataset from Bo Pang and Lillian Lee.

Keywords*-Sentiment analysis; Polarity; Natural Language Processing*

*****

I. INTRODUCTION

Sentiment analysis is a process of identifying opinions, attitudes, emotions and feelings of a particular topic or product with respect to writer's view to judge whether the topic or product is positive, negative or neutral. Sentiment analysis refers to NLP, where the analysis of the text helps in extracting different information from the source data. Sentiment analysis is often known as opinion mining, opinion extraction, sentiment mining, and subjective analysis. The main task of sentiment analysis is to identify the polarity or subjective/objective of a given text.

A basic task in sentiment analysis is to classify the polarity of a given text document or a sentence for identifying whether the document or a sentence is positive, negative, or neutral Fig.1. Beyond polarity the classification of sentiment looks, for instance, at emotional states such as angry, sad, happy, pleasant, unpleasant etc.

Scaling System is another method for determining sentiments by the words which are commonly combined having negative, positive or neutral sentiments with a scale of -5 to +5.
Another classification of sentiment analysis is subjectivity/objectivity identification. This identification of a given text is classified into one of two classes i.e. either objective or subjective. Compared to polarity classification, subjective/objective classification is more complex when the subjective sentence occurs in objective document and also the words and phrases of subjectivity depend on their context.

The feature/aspect-based sentiment analysis is more fine grained analysis model in which the opinions or sentiments are determined on different features or aspects of entities such as cell phone , digital camera etc.

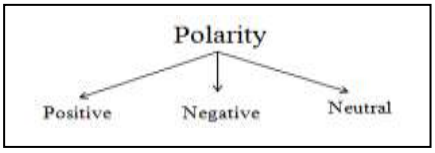

Figure 1. Classification of Polarity

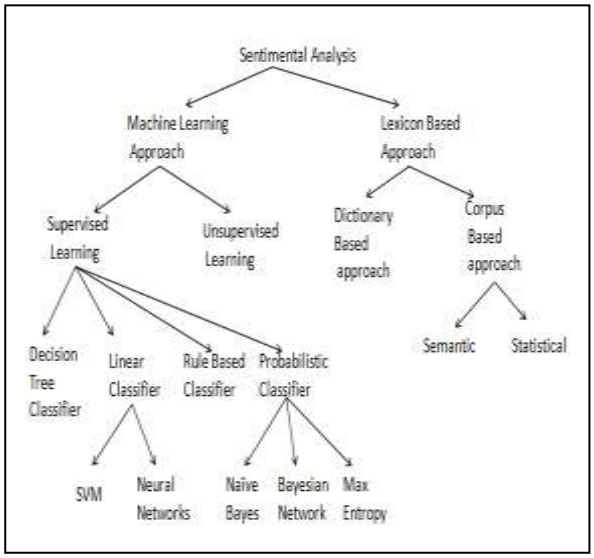

Figure 2. Classification of Sentiment Analysis

A. *Naïve Bayes*

Naïve Bayes is a simple probabilistic classifier which is a Supervised Machine Learning approach. Naïve Bayes works on Bayes theorem by strong independence assumptions. It is taken from Bayesian Statistics. Naïve Bayes classifier requires small amount of training data to calculate means and variances of the necessary variables for classification. In this we will assume the independent variables and only the variances of the

 ISSN: 2321-8169
5167 - 5172

variables need to be determined. In this Rapid Miner we will consider the data table as training data set for Naïve Bayes operator. Laplace Correction parameter is to prevent high influence of zero probabilities and the range is Boolean, which is the main advantage of Naïve Bayes classifier.

Let us consider an example, a red fruit whose diameter is 4 inches and round in shape. Naive Bayes classifiers use these individual properties where the probability which is provided considered that this fruit is an apple

1) Working of Naive Bayes

The naïve bayes classifier uses the following equations to calculate the probability of the class.

In the above mentioned Fig. 3,

- C = Class
- w = words
- argmax = Higher maximum value
- P (c_j) = Probability of Class C,
- C_{NB} = Likely Class
- $\pi_{i \epsilon positions}$ = Every Position of the document,
- $P(w_i|c_j)$ = Probability of word and class like w_i is positive and c_j is negative
- count(w,c)+1 = Laplace, count(c)=Total no of words, |V|=Total number of words

a) Training Document

- P(Chinese) = Total No of Chinese Class/Total Documents in Training
  =3/4
- P(Japan) = Total No of Japan Class/Total Documents in Training
  =1/4

b) Conditional Probabilites

- P(Chinese|c) = (No of Chinese word occur in Chinese training Class +1) / (Total No of words in Chinese Class) + (No of Probability Smoothing in Total Training)
  = (5+1) / (8+6)
  = 3/7
- Here No of Chinese word occur in Chinese training Class = 5

$$c_{NB} = \underset{c_j \in C}{\operatorname{argmax}} P(c_j) \prod_{i \in positions} P(w_i \mid c_j)$$

$$\hat{P}(w \mid c) = \frac{count(w,c)+1}{count(c)+|V|}$$

Figure 3. Naïve Bayes Formula

Table I Example of Naïve Bayes Classifier

	Document	Words					Class
Train	1	Chinese	Bejing	Chinese			Chinese
	2	Chinese	Chinese	Shilong			Chinese
	3	Chinese	Mizoram				Chinese
	4	Tokyo	Japan	Chinese			Japan
Test	5	Chinese	Chinese	Chinese	Tokyo	Japan	?

- Total No of words in Chinese Class = 8
  (i.e., Chinese, Beijing, Chinese, Chinese, Chinese, Shilong, Chinese, Mizoram)
- No of Probability Smoothing in Total Training = 6
  (i.e. Chinese, Beijing, Shilong, Mizoram, Tokyo, Japan)
- P(Tokyo|c) = (No of Tokyo word occur in Chinese training Class +1) / (Total No of words in Chinese Class) + (No of Probability Smoothing in Total Training)
  = (0+1) / (8+6)
  =1/14
- P(Japan|c) = (No of Japan word occur in Chinese training Class +1) / (Total No of words in Chinese Class) + (No of Probability Smoothing in Total Training)
  = (0+1) / (8+6)
  = 1/14
- P(Chinese|j) = (No of Chinese word occur in Japan training Class +1) / (Total No of words in Japan Class) + (No of Probability Smoothing in Total Training)
  = (1+1) / (3+6)
  = 2/9
- P(Tokyo|j) = (No of Tokyo word occur in Japan training Class +1) / (Total No of words in Japan Class) + (No of Probability Smoothing in Total Training)
  = (1+1) / (3+6)
  = 2/9
- P(Japan|j) = (No of Japan word occur in Japan training Class +1) / (Total No of words in Japan Class) + (No of Probability Smoothing in Total Training)
  = (1+1) / (3+6)
  = 2/9

 ISSN: 2321-8169
5167 - 5172

c) Testing Document

Choosing a Class

- P (w|d5) α P(c)*C*T*J
- P(Chinese|d5) = 3/4*3/7*3/7*3/7*1/14*1/14 =0.0003
- P (Japan|d5) = 1/4*2/9*2/9*2/9*2/9*2/9 = 0.0001
- Document 5(d5) is common for both training and testing.

III. WORK CONTRIBUTION

In this paper we are discussing about sentiment analysis for polarity data. The polarity dataset is downloaded from the data used by Bo Pang and Lillian Lee on Movie Reviews. The total data contains 2000 text files in which 1000 are Positive text files which are placed in pos folder whereas and 1000 are Negative text files are placed in neg folder. In this paper we have used the Rapid Miner which is a data mining tool used to test sentiment analysis for text data. The text data is been processed in this tool with the help of text mining operator. List of Operators used to perform sentiment analysis for Movie Reviews are listed below.

A. Process Document from files

1) Tokenize

This operator separates the content of a document into an arrangement of tokens. Each non-letter character is utilized as separator. Thus, every word in the content is denoted by a single token.

2) Filter Tokens (by length)

This filter tokens will filter the words by its length (i.e. the quantity of characters they contain).

3) Stem (Porter)

This operator stems English words utilizing the Porter stemming calculation by applying an iterative, standard based substitution of word additions proposing to decrease the length of the words until it reaches its minimum length.

4) Filter Stopwords (English)

This operator filters the English stopwords from a text data by separating each and every token till it matches the already described stopword list.

B. X-Validation

1) Naïve Bayes in Training Module

It helps to generate a Naive Bayes classification model.

2) Apply Model and Performance in Testing Module

a) Apply Model

This model applies an already trained model on an ExampleSet. The required parameters are used within this model. It is necessary that both ExampleSets should have precisely the same number, order, type and role of attributes.

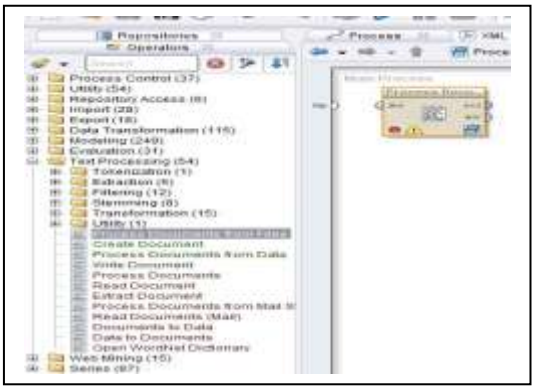

Figure 4. Process Document from files Operator

b) Performance

It is used to evaluate the performance. It conveys a list of performance criteria values. These execution criteria are naturally decided keeping in mind the end goal to fit the learning task type.

C. Store Operators

It helps to store an IO Object in the data repository at a specific location with the help of repository entry.

IV. STEPS TO WORK WITH SENTIMENT ANALYSIS IN RAPID MINER

a. In this paper we have taken Polarity movie review dataset which contains 1000 positive reviews and 1000 negative reviews.

b. Install Rapid Miner; working with sentiment analysis for text data we need Text Processing operator in Rapid Miner.

c. To get Text processing Operator in Rapid Miner, Go to help tab and select updates and extensions (marketplace) option.

d. A Rapid Miner Marketplace window will appear in which select Top downloads where list of operators are available.

e. Select Text Mining Extension operator and check on Select for Installation & then click Install packages button which will Install Text Mining Extension.

f. After completing the Installation, Text Processing operator is visible in operators tab of Rapid Miner.

 ISSN: 2321-8169
5167 - 5172

g. Expand text processing operator and Drag and Drop Process Document from Files on to the Main Process as per shown in Fig. 4.

h. Click on Edit List option from Process Documents from Files tab at the right side of the window.

i. Browse the Positive data of the polarity movie data.

Figure 5. Importing Data in Process Document from Files Operator

j. Click on Add Entry Button to browse negative data.

k. Browse the Negative data of the polarity movie data.

l. Click on Apply button as shown in Fig. 5.

- Double click on Process Document from files operator, vector creation tab will display.
- Expand Text Processing Operator then drag and drop the following operators on to Vector Creation.
- Search for the operators in the operators tab.
-

    a. Tokenization -> Tokenize
    b. Filtering -> Filter Tokens (by Length)
    c. Stemming -> Stem (Porter)
    d. Filtering -> Filter Stop words (by English)

Connect Left side ***doc*** as input for Tokenize.
Provide the output of Tokenize as input for Filter Tokens (by Length) operator.

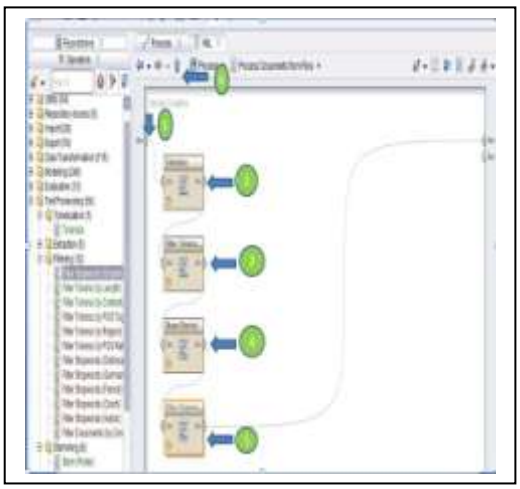

Figure 6. Vector Creation Tab

Provide the output of Filter Tokens (by Length) as input for Stem (Porter) operator.

Provide the output of Stem (Porter) as input for Filter Stopwords (by English) operator.

Provide the output of Filter Stopwords (by English) to doc as shown in Fig.6. Click on up arrow button and go back to main process.

Drag and drop X-Validation Operator.

Double click on X-Validation Operator which will display Training and Testing Modules.

Drag and Drop Naive Bayes Operator in Training Module from ***Modeling -> Classification and Regression -> Bayesian Modeling -> Naive Bayes.***

Connect ***tra*** as input to Naive Bayes operator and provide ***mod*** output to ***mod*** as input shown Fig. 7.

In testing module drag and drop Apply model operator from ***Modeling ->Model Application -> Apply Model***.

Drag and Drop Performance Operator in testing module from ***Evaluation-> Performance Measurement-> Performance.***

Give ***Mod*** and ***tes*** as input for ***mod*** and ***uni*** for Apply model operator.

The ***lab*** output of apply model is given as ***lab*** input for performance operator and the ***per*** output of performance operator is connected to ***ave.***

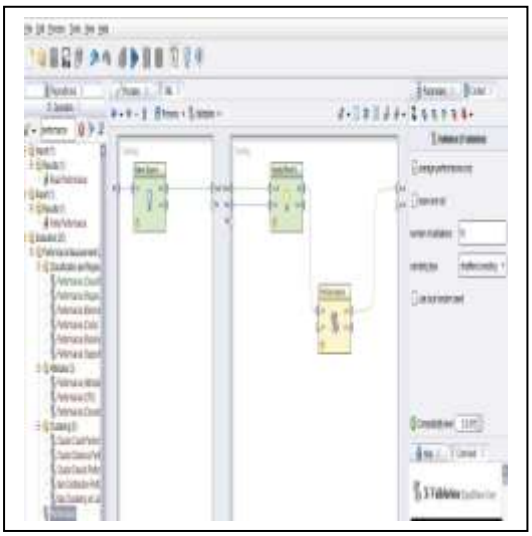

Figure 7. Training and Testing Module Tab

Click on up arrow button and go back to main process.
Drag and Drop 2 store operators in the main process ***Repository Access -> Store***.

Connect the ouput ***wor*** of Process Document from Files operator as an input to ***inp*** of Store (1) operator. Store (1)

 ISSN: 2321-8169
5167 - 5172

operator is used to store the data present in process document from files.

Connect the ouput ***mod*** of Validation operator as an input to ***inp*** of Store (2) operator.

Connect the ouput ***thr*** of store operator as to **res**. Store (2) Operator is used to store the result. Here we have taken number of validations=10

Click on Store (1), in parameters tab repository entry option is available.

Click on browse button, a repository entry window will open in which select the data folder and give the name (here we have given name as data1) to the store data.

Click on Store (2), in parameters tab repository entry option is available.

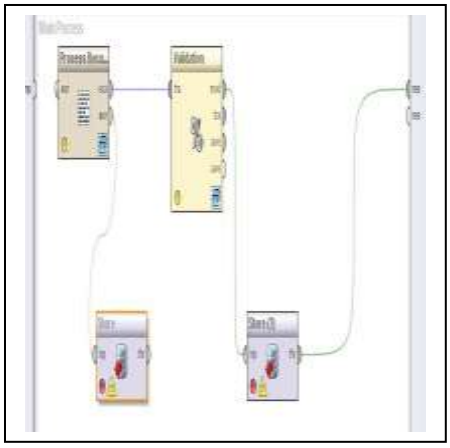

Figure 8. Connection of Operators in Main Process Tab

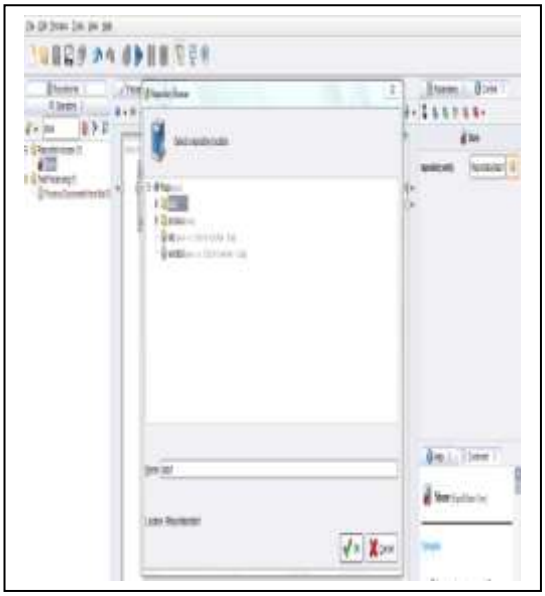

Figure 9. Storing of Data in Repository

Click on browse button, a repository entry window will open in which select the process folder and give the name (here we have given name as res2) to the store result.

V. RESULTS

A. Calculation of Accuracy for 10 validations

Confusion Matrix

	True Positive (TP)	True Negative (TN)
Predicted Positive (PP)	72	30
Predicted Negative (PN)	28	70

Sensitivity = PPTP / (PPTP + PPTN)
= 72 / (72+30)
= 0.7059

Specificity = PNTN / (PNTP + PNTN)
= 70 / (28+70)
= 0.71428

Accuracy = (Sensitivity+Specificity)/2
= (0.7059+0.71428)/2
= 0.71

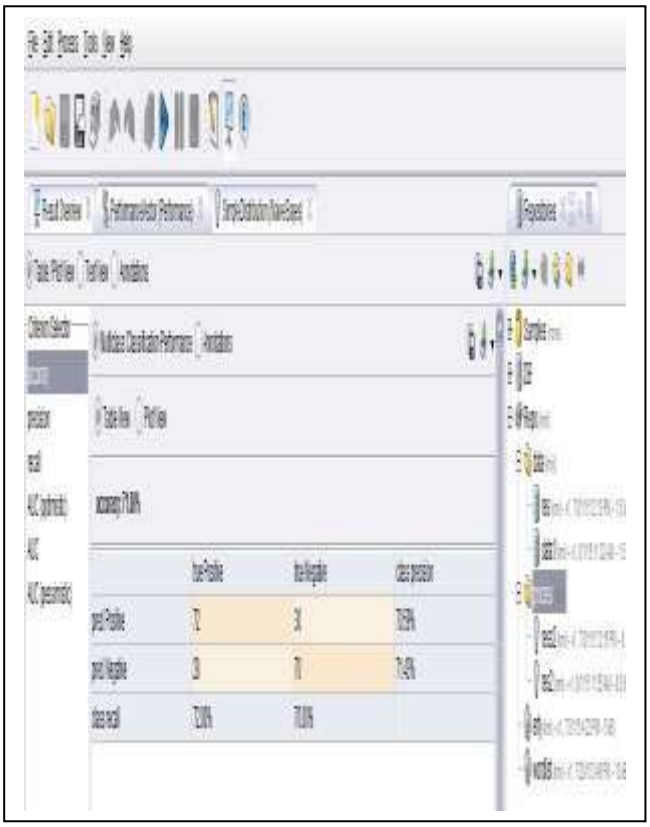

Figure 10. Naïve Bayes Performance Vector Output for 10 Validations

 ISSN: 2321-8169
5167 - 5172

B. Calculation of Accuracy for 50 validations

Confusion Matrix

	True Positive (TP)	True Negative (TN)
Predicted Positive (PP)	12	4
Predicted Negative (PN)	8	16

Sensitivity=PPTP / (PPTP + PPTN)
= 12 / (12+4)
= 0.75

Specificity=PNTN / (PNTP + PNTN)
= 16 / (8+16)
= 0.66

Accuracy = (Sensitivity+Specificity)/2
= (0.75+0.66)/2
=0.705

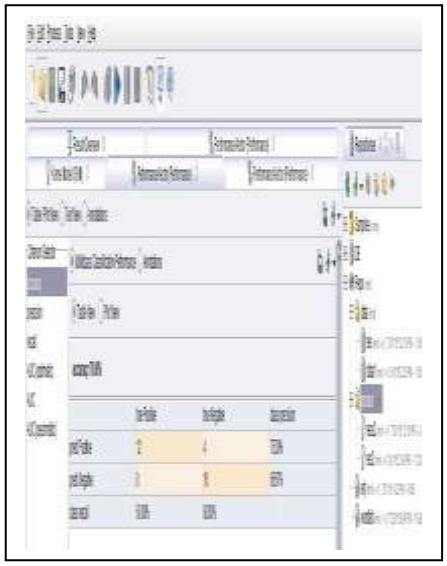

Figure 11. Naïve Bayes Performance Vector Output for 50 Validations

VI. CONCLUSION

In this paper we have worked with sentiment analysis on Movie Review data which is taken from Bo Pang and Lillian Lee, we have passed Polarity text dataset as an input in Rapid Miner Tool, where we have shown step by step process of working with sentiment analysis using Naïve bayes Classifier for 10 validations in Fig.10 and for 50 validations in Fig.11 and calculated the accuracy for both the validations i.e. 71% for 10 Validations and 70.5% for 50 Validations. Future Enhancement can be done for Sentiment analysis for Dravidian Languages; we can also work with different types of modeling techniques like Support Vector Machine, Maximum Entropy etc.

REFERENCES

[1] Aljaz Blatnik ˇ , Kaja Jarm , Marko Mezaˇ, “Movie sentiment analysis based on public tweets”, http://ev.fe.uni-lj.si/4-2014/Blatnik.pdf.

[2] Almas Y., and Ahmad K., "A note on extracting sentiments in financial news in English, Arabic & Urdu." The Second Workshop on Computational Approaches to Arabic Script-based Languages. 2007.

[3] Bo Pang and Lillian Lee “Movie Review Data”, https://www.cs.cornell.edu/people/pabo/movie-review-data/

[4] Dan Jurafsky and Christopher Manning, “Natural Language Processing”, https://class.coursera.org/nlp/lecture/145

[5] Dr.Sved Saifur Rahman, “http://applieddatamining.blogspot.in/2013/09/naive-bayes-classification-using.html”

[6] MullenSentimentCourseSlides, “Introduction to Sentiment Analysis”
http://www.lct-master.org/files/MullenSentimentCourseSlides.pdf.

[7] Rapid Miner Documentation, http://docs.rapidminer.com/studio/operators/modeling/classification_and_regression/bayesian_modeling/naive_bayes.html

[8] Ray Chen, Marius Lazer, "Sentiment Analysis of Twitter Feeds for the Prediction of Stock Market Movement".

[9] Sheamus McGovern**, “**Sentiment Analysis in Rapid Miner Part-I” Posted on October 4, 2012

[10] Tanu Verma, Renu and Deepti Gaur, “Tokenization and Filtering Process in RapidMiner”
http://research.ijais.org/volume7/number2/ijais14-451139.pdf , April 2014