The 1st International Conference on Innovations in Computer Science & Software Engineering, (ICONICS 2016), December 15-16, 2016

Comparative study on sentiment analysis using WEKA classifiers

Rabica Shamim
Dept. Computer Science and Information Technology, NED University of Engineering and Technology
Karachi, Pakistan
Rabica_shamim@yahoo.com

Syeda Sadaf Shahab
Dept. Computer Science and Information Technology, NED University of Engineering and Technology
Karachi, Pakistan
shahabsadaf@hotmail.com

Sohail Abdul Sattar
Dept. Computer Science and Information Technology, NED University of Engineering and Technology
Karachi, Pakistan
sattar@neduet.edu.pk

Abstract*— Classification of text is an important aspect of sentiment analysis and helpful in making decision about any product, article, book, movie etc available in market. There are different algorithms available for classifying text data. We have tested a dataset comprising of 2000 text files containing movie reviews[1] using Neural Network (NN), Support Vector Machine (SVM), Naïve Bayes (NB), Logistic regression and J48 algorithm in supervised learning environment using WEKA on training sets and testing test to proposed the best suitable algorithm for text classification in positive and negative sentiment. In this paper we will define the relevancy of above mentioned algorithms, test results which we have performed, comparative study of different results we observed and conclusion.

Keywords*— text mining; algorithm for text mining; decision making;

I. INTRODUCTION

Contextual base analysis classifies the data with respect to polarity and the opinion given in text. It is widely used in online shopping cart usually for product reviews; provide a buyer to make their mind to purchase or not about particular item.

There are multiple classification algorithm, they are only useful if we use right algorithm for right category of data which provides which provides optimal and efficient results.

In next paragraph we will provide a brief description about the algorithm we have used to conclude best option for text mining.

A. Neural Network

Neural Networks is a widely accepted paradigm to analyze trends and pattern. The best part of NN is that it results are unbiased and doesn't worry the distribution of the data.

B. Support Vector Machine

SVMs are learning method introduced by V. Vapnik[3]. It is best suitable supervised machine learning algorithm which is suitable for classification of text based problem.

C. Naïve Bayes

This is based on Bayes' algorithm using conditional probability for analyzing each result of observation.

D. Logistic Regression

Logistic regression is best suitable for mine the binary data, its output is discrete.

E. J48

J48 is a branch from decision tree; it is a powerful technique in decision making and data extraction. It uses training data set or extracting information benefits and it examines the results from choosing an attribute for splitting the data. The splitting procedure stops if all instances in a subset belong to the same class. Then the leaf node is created in a decision tree telling to choose that class

II. CLASSIFICATION

WEKA classification classifies the data in to training set which allow choosing classifiers. For data classification we need to provide the following:

- Data set
- Division of data into training set
- Identify the suitable classier
- Understanding of training data result

III. EXPERIMENT

We have performed the experiment using test data to predict the optimal data mining algorithm using WEKA. It is working with different format; Data can be loaded from various sources, including files, URLs and databases. Supported file formats include WEKA‟s own ARFF format, CSV, Lib SVM‟s format, and C4.5‟s format. It is also possible to generate data using an artificial data source and edit data manually using a dataset editor. We had data in text files which we had converted into .arrf format using WEKA SimpleCLI using following commands:

- Open SimpleCLI from Weka GUI chooser
- write:
- java weka.core.converters.TextDirectoryLoader –dir Location_of_Text_Data_files_folder\ > Location_to_save_arff_format_data\Name_of_arff_file.arff
- press enter

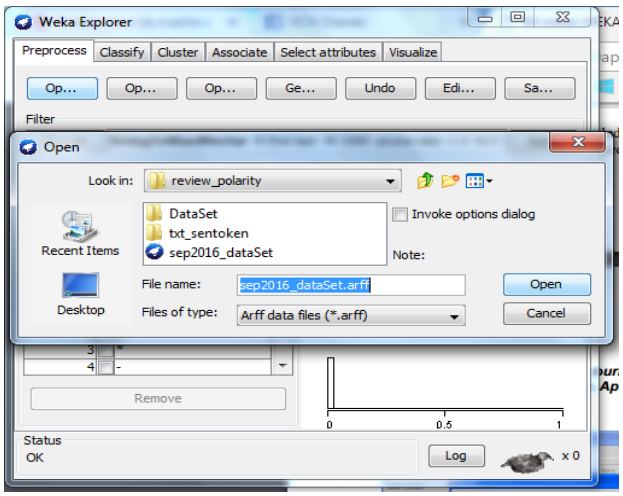

Figure 1: Loading Data File in Weka

This procedure generates the .arff files of data set provided. Load the .arff file using WEKA Explorer see figure 1.

Filter option allow to filter data as per requirement, here had chosen *StringtoWordVector* which divides the text into positive and negative classes showing in figure 2.

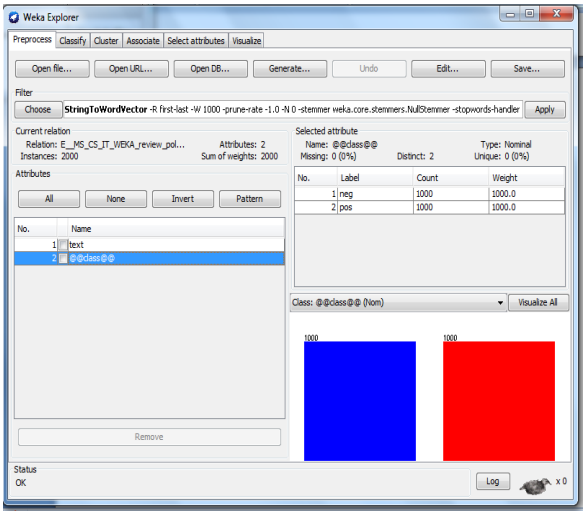

Figure 2: Filtered Data in Positive and Negative classes

IV. COMPARATIVE ANALYSIS

Java provides open source Opinion mining to analyze evolutionary algorithms for big data. Main issues in mining include techniques of data classification. To evaluate data we used two modes: 1. K fold and 2 Percentage spilt mode. In K fold database is randomly divided into K disjoint blocks of objects , then the algorithm is trained using k- 1 blocks and the remaining blocks is used to test the performance of the algorithm. This process is repeated k times. In last all records averaged. In percentage split, the database is randomly split in to two disjoint datasets. The first set, called training set where as second set is called test set; it is common to randomly split a data set under the mining task in to 2 parts. We divided the dataset into 75% training set & 25% test set. Once the tests is carried out using the selected datasets, then using the available classification and test modes ,results are collected and an overall comparison is conducted.

We performed the testing using multiple classifiers their correctly classification of instances vary. In our experiment we observed that J48 accuracy level for Correctly Classified Instances is lowest where as Naïve Bayes accuracy level was highest among five algorithms used.

Not only correct instance recognition Naïve Bayes has lowest Incorrectly Classified Instances count, Table 1.

=== Summary of Naïve Bayes ===

Correctly Classified Instances	1688	84.4%
Incorrectly Classified Instances	312	15.6 %
Kappa statistic	0.688	
Mean absolute error	0.1989	

Table 1: Comparative results of Algorithms

Coverage of cases (0.95 level)	98.1 %
Total Number of Instances	2000

Using Table 1 as reference we can nominate the best suitable algorithm for text polarity classification, i.e.

1. Naïve Bayes

Data Type (Test Data)	**Naive Bayes %**	**J48 %**	**Logistic R %**	**SVM %**	**NN%**
Accuracy	80.6	71.8	80.2	80	78.2
Non Accuracy	19.4	21.2	19.8	20	21.8

2. Logistic regression
3. SVM
4. Neural Network
5. J48

V. CONCLUSION

In this paper we analyzed text dataset to review the best optimal Text extraction tool. Weka helped to analyze the large text files using various mining algorithms. Test result suggests that '*Naïve Bayes'* is optimal solution for text based data with low rate of incorrect instance finding and high rate of correct one. We can use naïve bayes algorithms in online stores (for product review), statement analysis, Personality analysis questioner etc.

VI. REFERENCES

[1] http://www.cs.cornell.edu/People/pabo/movie-review-data/

[2] M. Govindarajan, and R. M. Chandrasekaran, "Classifier Based Text Mining for Neural Network", World Academy of Science, Engineering and Technology International Journal of Computer, Electrical, Automation, Control and Information Engineering Vol:1, No:3, 2007

[3] Vladimir N. Vapnik. The Nature of Statistical Learning Theory. Springer, New York, 1995.

[4] Dr. Neeraj Bhargava, Girja Sharma, Dr. Ritu Bhargava and Mansih Mathuri, Decision Tree Analysis on J48 Algorithm for Data Mining", International Journal of Advanced Research inComputer Science and Software Engineering, Volume 3, Issue 6, June 2013.

[5] Trilok Chand Sharma and Manoj Jain, "WEKA Approach for Comparative Study of Classification Algorithm", International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 4, April 2013

[6] Neelamadhab Padhy , Dr. Pragnyaban Mishra , and Rasmita Panigrahi, "The Survey of Data Mining Applications And Feature Scope", International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol.2, No.3, June 2012

[7] Swasti Singhal, and Monika Jena, "A Study on WEKA Tool for Data Preprocessing, Classification and Clustering", International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-2, Issue-6, May 2013.

[8] Nur Hafieza Ismail, Fadhilah Ahmad , and Abdul Azwa, "Implementing WEKA as a Data Mining Tool to Analyze Students' Academic Performances Using Naïve Bayes Classifier", UniSZA Postgraduate Research Conference 2013, At UniSZA