Advances and Applications in Mathematical Sciences
Volume 18, Issue 11, September 2019, Pages 1399-1405

COMPARISON OF CLASSIFIERS FOR SENTIMENT ANALYSIS IN UNSTRUCTURED DATA

M. RAMESH RAJA, J. ARUNADEVI and R. BOWMIYA BARZITH

Ph.D. Scholar (Part-Time)
Department of Computer Science
Raja Doraisingam Govt. Arts College
Sivaganga, Affiliated to Alagappa University

Assistant Professor, Department of Computer Science
Raja Doraisingam Govt. Arts College
Sivaganga, Affiliated to Alagappa University

M.Phil. Scholar, Department of Computer Science
Raja Doraisingam Govt. Arts College
Sivaganga, Affiliated to Alagappa University

Abstract

Aim: The aim of this research is to compare the performance of the various types of classifiers for the purpose of sentiment analysis. This analysis is done through the text mining methods to analyze the subjectivity of the reviews in unstructured data.

Background: Unstructured data is a big challenge to the data science community. The abundant nature of this type of data is opening an arena for research. The knowledge hidden in this type of data is a hidden treasure for the decision makers. In this paper we consider sentiment analysis, which is a topic to be concentrated for the promotion of the business, customer feedback, examine the brands, market research, product analysis, etc.

Methodology: For this study the researchers employed four classifiers for the comparison. They are Support Vector Machine, Decision Tree, Niave Bayes, *K*-Nearest neighbor classifiers. These classifiers are tested against three datasets. The datasets used for the experiment is review datasets from Amazon, IMDB and Yelp. They contain the unstructured data. This unstructured data is transformed in to the structured one through the text mining tasks. After this the structured data is used for classification.

Contribution: This paper identifies the better classifier in the given environment. This classification is done based on the polarity of the data. This work is done on the sentence level

2010 Mathematics Subject Classification: 68U15.

Keywords: sentiment analysis, unstructured data, text mining, support vector machine, decision tree, Niave Bayes, *K*-nearest neighbor.

Received July 11, 2019; Accepted September 19, 2019

sentiment analysis for the customer reviews. This will be helpful for the producers to understand the customer's expectation and to comprehend pros and cons of their product. This makes the usage of the unstructured data into a vital component of the business.

1. Introduction

Unstructured data are the data which doesn't follow any format. In the real world most of the data are unstructured. The need for the processing of this type of data is challenge for any data scientist. The type of data provides the rich source of information about everything in the globe [1]. Unstructured data contributes nearly 70 % of the data that is being procured and it doubles in every two months [2].

Sentiment analysis is the important tool for the business enterprises [3]. It is an open problem in the research. This analysis will help the business to improve based on the customer's reviews. But the fact we have to consider is that the most of the data for this analysis will be unstructured in nature. The research challenge in this problem is to seek information in the unstructured data.

The problem addressed in this paper is to classify the sentiments given by the reviewers. The challenge in the problem is to get information from the unstructured data. The methodology we have employed is that to apply the text mining based classifiers to accomplish the task.

2. Background Study

The study requires the concepts related to unstructured data, sentiment analysis and the text mining procedures.

2.1. Unstructured Data

We are all living in the era of data. That is data is more important in deciding everything. Data play mightier role in the deciding the behavior of a person. The data could be mostly unstructured. Comments, reviews are all the common things that everyone will come across. It doesn't have any predefined model or it doesn't fit into the model. It doesn't follow any semantics that normally a structured on has. The sources are social media data, web pages, images, reports etc.

2.2. Sentiment Analysis

Sentiment analysis is the contextual information extraction and the analysis of the same. Subjective information is to be gathered and the analysis on the polarity could be gained from this [4-5]. It could be positive, negative or neutral. The application domain varies from problem to problem.

2.3. Text Mining

This is the mining process which could extract the information from the textual data. The text is the common form that the human communicate, document and speak. The text is the basic bed for the sharing of feelings in terms of speech, writing and the hearing. Thus the text mining should be given importance over that normal data mining. The process in the text mining is to pre process the text and then it is used for the information extraction [6].

3. Problem Statement

The problem discussed here is the sentiment analysis from the unstructured data. The unstructured data should be processed and to be presented for the analysis. The analysis process must be reflected for the business values to make decisions out of it.

4. Proposed Methodology

The proposed methodology in the paper is to apply text mining techniques to the unstructured data and then the application of classifiers to the structure data out of it.

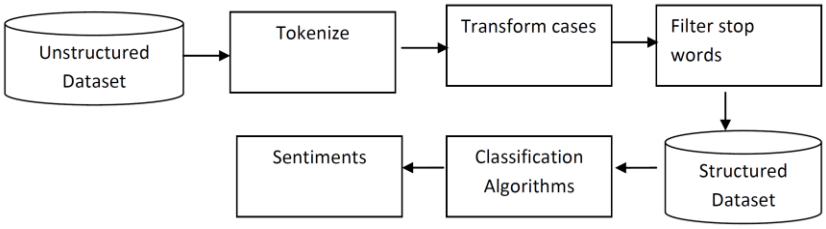

Figure 1. Workflow of the proposed methodology.

5. Experimental Setup

The experiment is carried out with three datasets, which is tested on four classifiers for this research. The classifiers used are naïve bayes, *K*-NN, Decision tree and SVM.

Table 1. Dataset description.

Dataset	No. of Documents in training dataset	No. of Documents in test dataset	No. of attributes generated after preprocessing
Amazon	700	300	1253
IMDB	520	220	1797
Yelp	700	300	1390

5.1. Performance Metrics used

Table 2. Confusion Matrix for a binary classifier.

	Class-I Predicted	Class-II Predicted
Class I - Actual	TP	FN
Class II - Actual	FP	TN

Where,

True Positive (TP). Observation is positive, and is predicted to be positive.

False Negative (FN). Observation is positive, but is predicted negative.

True Negative (TN). Observation is negative, and is predicted to be negative.

False Positive (FP). Observation is negative, but is predicted positive.

The performance metrics used for the experiment is given below

$$\text{accuracy} = (\text{Correct predictions})/(\text{Number of Examples})$$

$$= (TP + TN)/(TP + FP + FN + TN)$$

$$\text{classification error} = (\text{Incorrect predictions})/(\text{Number of Examples})$$

$$-(FP + FN)/(TP + FP + FN + TN).$$

$$\text{Cohen's kappa} = (po - pe)/(1 - pe)$$

where

$$po = \text{observed accuracy} = (TP + TN)/(TP + FP + FN + TN)$$

$$pe = \text{expected accuracy} = [(TP + FP)(TP + FN) + (FN + TN)(FP + TN)]/$$

$$[(TP + FP + FN + TN)^2]$$

precision = (True positive predictions)/(All positive predictions)

$$= TP/(TP + FP)$$

recall = True positive rate = (True positive predictions)/(Number of positive Examples) $= TP/(TP + FN)$

f measure $F1 = 2$ (precision $*$ recall)/(precision + recall)

$$= 2TP/(2TP + FP + FN)$$

specificity = True negative rate = (True negative predictions)/(Number of negative Examples) $= TN/(TN + FP)$.

6. Experimental Results

The following tables gives the details of the consolidated results obtained by the experiments conducted, which is discussed above.

Table 3. Results obtained from Amazon database.

Classifier/ Measure	Accuracy	Error	Kappa	Precision	Recall	*F*-Measure	Specificity
Naïve Bayes	70.71	29.29	0.411	67.45	82.25	73.99	58.70
KNN	67	33	0.332	61.82	95.19	74.74	37.67
Decision Tree	65.71	34.29	0.321	90.74	36.20	50.32	96.22
SVM	75	25	0.498	71.64	84.53	77.41	65.12

Table 4. Results obtained from IMDB database.

Classifier/ Measure	Accuracy	Error	Kappa	Precision	Recall	F-Measure	Specificity
Naïve Bayes	68.27	31.73	0.373	61.48	76.52	68.04	61.72
KNN	69.23	30.77	0.399	61.25	83.48	70.45	57.93
Decision Tree	61.15	38.85	0.138	86.84	14.35	24.63	98.28
SVM	66.73	33.27	0.273	92.07	27.83	41.96	97.59

Table 5. Results obtained from Yelp database.

Classifier/ Measure	Accuracy	Error	Kappa	Precision	Recall	*F*-Measure	Specificity
Naïve Bayes	69.86	30.14	0.401	64.34	74.96	68.72	65.77
KNN	54.29	45.17	0.156	49.91	92.81	64.33	23.92
Decision Tree	60.29	39.71	0.133	84.45	22.39	27.99	90.01
SVM	69.43	30.57	0.351	77.46	45.11	56.02	88.49

6. Conclusion

This research is carried out for the sentiment analysis. We have used the text mining procedures to convert the unstructured data to the structured data format and then the classifiers are applied for the given dataset for the sentiment analysis. We have used four binary classifiers, which classify as positive or negative. The results are tabulated. The future research could be concentrated on multi label classifiers for the sentiment analysis.

Dataset Acknowledgment

The dataset used for this paper is obtained from the research work by D. Kotzias, M. Denil, N. De Freitas, and P. Smyth described in the KDD 2015 paper 'From Group to Individual Labels using Deep Features'. We are grateful to these authors for making the dataset available.

References

[1] Orobor, Ise, Integration and Analysis of Unstructured Data for Decision Making: Text Analytics Approach, International Journal of Open Information Technologies 4 (2016), 82-88.

[2] S. Chitra, N. Shunmuga Karpagam and K. Venkataramanan, Unstructured Data into Intelligent Information Analysis and Evaluation, International Conference on Global Innovations in Computing Technology, 2014.

[3] V. Mika Mäntylä, Daniel Graziotin, Miikka Kuutila, The evolution of sentiment analysis-A review of research topics, venues, and top cited papers, Computer Science Review, Volume 27, February 2018.

[4] Rui Xia, Feng Xu, Chengqing Zong, Qianmu Li, Yong Qi and Tao Li, Dual Sentiment Analysis: Considering Two Sides of One Review, IEEE Trans. on Knowledge and Data Engineering, 2015.

[5] Walaa Medhat, Ahmed Hassan and Hoda Korashy, Sentiment Analysis Algorithms and Applications: A survey, Ain Shams Engineering Journal, 2014.

[6] C. C. Aggarwal and C. Zhai, A survey of text classification algorithms, Springer (2012), 163-222.