# Mahalanobis Distance-the Ultimate Measure for Sentiment Analysis

Valarmathi Balasubramanian[1], Srinivasa Gupta Nagarajan[2], and Palanisamy Veerappagoundar[3]
[1]Faculty of Soft Computing Division, VIT University, India
[2]Faculty of Manufacturing Division, VIT University, India
[3]Faculty of Electronics and Communication Engineering, Anna University, India

**Abstract**: *In this paper, Mahalanobis Distance (MD) has been proposed as a measure to classify the sentiment expressed in a review document as either positive or negative. A new method for representing the text documents using Representative Terms (RT) has been used. The new way of representing text documents using few representative dimensions is relatively a new concept, which is successfully demonstrated in this paper. The MD based classifier performed with 70.8% of accuracy for the experiments carried out using the benchmark dataset containing 25000 movie reviews. The hybrid of MD based Classifier (MDC) and Multi Layer Perceptron (MLP) resulted in a 98.8% of classification accuracy, which is the highest ever reported accuracy for a dataset containing 25000 reviews.*

**Keywords**: *Sentiment analysis, MD, opinion mining, machine learning algorithms, hybrid classifier.*

*Received August 20, 2012; accepted September 26, 2013; published online April 1, 2015*

## 1. Introduction

Sentiment analysis is fast becoming the focus area of research for both academic and business community. The present generation is increasingly depending on the internet to exchange their views through the online forums. Hence, it becomes very essential for the business community to make use of this opportunity to understand the voice of customers expressed through the online forums. The volume of data available for analysis is really huge and it is impossible for an analyst to read all the reviews and understand the sentiments of the customers. Currently, the research is in full swing to detect sentiment expressed in a review document automatically.

In the document level sentiment analysis, a review document is classified as either positive or negative based on the sentiment expressed in it [6]. This kind of sentiment classification is carried out at the document level, without discovering about what people liked or did not like.

## 2. Existing Works in Sentiment Analysis

Research in the field of sentiment analysis is carried out either using Machine Learning (ML) algorithms or by Natural Language Processing (NLP) techniques. Recently the focus has been shifted to combine the best of both ML and NLP techniques in the form of hybrid approaches.

### 2.1. The Data Mining Approach

In this approach, text documents are expressed as term-document matrix containing numerical scores. After this, ML algorithms are applied for the purpose of classification.

A simple unsupervised learning algorithm for the classification of reviews [14] relating to movies, banks and automobiles achieved 66%, 80% and 84% classification accuracy respectively for each category.

Support Vector Machines (SVM) based classifier achieved 88.9% of accuracy using unigrams, bigrams and trigrams model [2]. NB and SVM based classifier [9] achieved 87.2% accuracy on movie reviews using unigrams alone. Linear SVM trained with large feature vectors in combination with feature reduction [3] has been used for automatic sentiment classification in the very noisy domain of customer data.

Document level sentiments assigned on a three-point/four-point scale [10] with SVM and regression tools classified the movie reviews with an accuracy of 66.3%.

Cosine distance, Euclidean distance and Manhattan distance [1] have been used for finding the best algorithm for text mining and found that the Self-Organizing Map (SOM) based clustering algorithm with cosine distance is the best for text mining.

Mahalanobis Distance (MD) has been used as a measure to classify the movie review documents [15] as either positive or negative. The reference set or the Mahalanobis Space (MS) was constructed using 93% of the positive reviews. In the present work, the given reviews are expressed as a reduced dimension matrix using the Representative Terms (RT) and MD is used for the purpose of classification.

### 2.2. The NLP Approach

NLP approach to sentiment analysis uses POS tagging, lexicon development and pattern analysis.

NLP based sentiment analyzer [16] assigns sentiments to each of the references corresponding to a given subject. A sentiment lexicon and sentiment pattern database is used for this purpose. The feature extraction algorithm used in this method identifies the topic related feature terms, which results in a finer level of sentiment analysis.

The pattern based approach based on NLP techniques [8] for assigning sentiments at topic level reported a classification accuracy of 94.5% on a dataset consisting of 255 camera reviews.

A pattern based approach using full parsing and top-down tree matching [4] has been used for extracting the sentiment units with high precision. In this method, the task of sentiment analysis is equated to the task of language translation and based on which, a system was proposed for extracting the sentiment units from the text. Transfer based machine translation engine has been used. The translation patterns and bilingual patterns were replaced with sentiment patterns and sentiment polarity lexicon.

### 2.3. The Hybrid Approach

A hybrid of pattern based classifier and SVM [5] achieved 91% of accuracy with 90% training data using unigrams, bigrams and trigrams. Sequential Minimal Optimization (SMO) algorithm was used for training the SVM classifier.

Rule-based classifiers supervised learning based classifiers and machine learning based classifiers were used in a sequence [11] to create a hybrid classifier and it was proved that using multiple classifiers sequentially in a hybrid manner can result in a better effectiveness than any individual classifier.

## 3. Mahalanobis Distance

In 1930, PC Mahalanobis, the founder of Indian Statistical Institute, introduced a statistical measure called MD. MD is a superior statistical measure than the other statistical measures like Euclidean distance and manhattan distance used for classification and clustering because it is based on the correlation among the various dimensions of the given problem [13].

Mahalanobis-Taguchi System (MTS) uses the MD for solving the pattern recognition problems. If a reference set can be created using the characteristic dimensions of the given problem, then using the reference set, the test set can be classified whether it belongs to the family of reference set or not by calculating the MD between the test set and the reference set [13]. The reference set is called as MS.

For example, in a medical diagnosis system, the improvement in health after a medication can be assessed using the MD between the MS and the test data. MS for this purpose is created using the data of a group of healthy people. If the medication has resulted in a health improvement, then the MD between the patient's test data and MS will be less and MD will be more if the health improvement has not happened.

In this paper, a relatively new method based on RT [12] is used. Each document is represented by a feature vector containing just eight dimensions viz. Good, very good, excellent, recommended, bad, very bad, disgusting and never recommended. These dimensions are named as RT. The matrix representation of text documents using these RT as columns and each document as its row is named as Representative Term-Document Matrix (RTDM).

## 4. RTDM Creation from the Text Documents

First, the documents are structured as RTDM using a PERL program. Each review document is represented as a row in the RTDM with eight features representing it. The eight representative features are good, very good, excellent, recommended, bad, very bad, disgusting and never recommended. Based on the RT occurrence (number of times the corresponding category of phrase/word occurs in a document), the RTDM is constructed. A portion of RTDM is shown in the Table 1, in which, the rows represent the reviews and columns represent the features mentioned above.

The numbers in Table 1 represents the number of times a corresponding category of RT appeared in that review document. The rules to capture the RT from the review documents were written based on 200 positive reviews and 200 negative reviews from the large movie review dataset containing 25000 reviews.

Table 1. A portion of RTDM.

| **Document No.** | **RT** | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | **Good** | **Very Good** | **Excellent** | **Recommended** | **Bad** | **Very Bad** | **Disgusting** | **Never Recommended** |
| **1.** | 0 | 4 | 2 | 0 | 2 | 3 | 2 | 0 |
| **2.** | 5 | 2 | 8 | 2 | 3 | 2 | 0 | 0 |
| **3.** | 0 | 3 | 2 | 2 | 6 | 0 | 1 | 1 |
| **4.** | 3 | 2 | 1 | 1 | 2 | 0 | 0 | 0 |
| **5.** | 4 | 3 | 3 | 1 | 11 | 4 | 5 | 0 |
| **6.** | 5 | 0 | 1 | 1 | 4 | 6 | 2 | 4 |
| **7.** | 2 | 1 | 1 | 0 | 4 | 1 | 2 | 1 |
| **8.** | 4 | 3 | 3 | 0 | 2 | 0 | 5 | 1 |

A sample set of rules used for capturing the relevant RT in the document are given below:

```
$_ =~ s/\s+memorable\s+/ good /g.
$_ =~ s/\s+fine performance\s+/ good /g.
$_ =~ s/\s+emotional roller coaster\s+/ good /g.
$_ =~ s/\s+worked so well\s+/ very_good /g.
$_ =~ s/\s+brillantly acted\s+/ very_good /g.
$_ =~ s/\s+great surprise\s+/ very_good /g.
$_ =~ s/\s+hugely entertained\s+/ very_good /g.
$_ =~ s/\s+high throughput\s+/ excellent /g.
$_ =~ s/\s+very touching\s+/ excellent /g.
$_ =~ s/\s+worth watching\s+/ recommended /g;
$_ =~ s/\s+most memorable\s+/ recommended /g.
$_ =~ s/\s+outstanding\s+/ recommended /g.
$_ =~ s/\s+poorly\s+/ bad /g.
$_ =~ s/\s+outdated\s+/ bad /g.
$_ =~ s/\s+slow\s+/ bad /g.
$_ =~ s/\s+stilted acting\s+/ bad /g.
$_ =~ s/\s+starting confuse\s+/ very_bad /g.
$_ =~ s/\s+totally disconnected\s+/ very_bad /g.
$_ =~ s/\s+under looked\s+/ very_bad /g.
$_ =~ s/\s+very slow\s+/ very_bad /g.
$_ =~ s/\s+really annoying\s+/ disgusting/g.
$_ =~ s/\s+severely annoying\s+/ disgusting /g.
$_ =~ s/\s+stopped watching\s+/ disgusting /g.
$_ =~ s/\s+worst joke\s+/ disgusting /g.
$_ =~ s/\s+avoid\s+/ never_recommended /g.
```

The rules for assigning words and phrases to a particular RT were written based on how the human mind would understand while reading a review. The rules were written considering all the individual words, phrases and negative patterns. A Perl program has been developed for capturing the opinion words and phrases and to assign the appropriate RT for them.

## 5. MS and Threshold Value

MS can be constructed from either positive reviews or negative reviews. In this paper, 12 positive reviews were considered for the creation of MS. Table 2 shows the MS selected for this analysis. MDs for all the documents in the MS were calculated using the Equation 1. MATLAB software has been used for performing the matrix analysis.

$$MD=1/K(Z_{ij}\,C^{-1}Z_{ij}^{T}) \qquad (1)$$

Where $K$=Number of RT, $Z_{ij}$ =RTDM, $Z_{ij}^{T}$=Transpose of the RTDM, and $C^{-1}$=Inverse of the correlation matrix of $Z_{ij}$.

Figure 1 shows the MD of 25000 reviews with respect to the MS. In the dataset, the first 12500 documents are negative reviews and the remaining reviews are positive. The threshold value to classify the documents as either positive or negative has been calculated based on the least misclassification point. As per the MTS theory, the threshold value of MD for the purpose of classification should minimize the loss due to misclassification in both cases. In this analysis, it was found that, for a threshold value of 56, the classification accuracy was 70.8%. Since the MS is chosen from the positive reviews, the positive reviews should have smaller MD with respect to the MS compared to the negative reviews. In the Figure 1, the first 12500 documents belong to the negative reviews and the remaining documents are positive reviews. It can be observed that the MD of the positive reviews is less compared to the negative reviews.

Table 2. Mahalanobis Space (MS).

| Document No. | RT | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Good | Very Good | Excellent | Recommended | Bad | Very Bad | Disgusting | Never Recommended |
| **351** | 1 | 1 | 4 | 0 | 1 | 0 | 0 | 0 |
| **352** | 2 | 1 | 2 | 2 | 2 | 0 | 1 | 0 |
| **353** | 9 | 13 | 7 | 3 | 13 | 2 | 3 | 1 |
| **354** | 0 | 0 | 3 | 2 | 1 | 0 | 1 | 0 |
| **355** | 5 | 4 | 5 | 1 | 2 | 2 | 0 | 1 |
| **356** | 3 | 1 | 3 | 1 | 3 | 0 | 0 | 0 |
| **357** | 4 | 1 | 7 | 0 | 3 | 0 | 0 | 1 |
| **358** | 1 | 1 | 4 | 4 | 0 | 0 | 0 | 0 |
| **359** | 2 | 1 | 6 | 2 | 1 | 0 | 1 | 0 |
| **360** | 4 | 2 | 3 | 0 | 3 | 2 | 0 | 0 |
| **361** | 5 | 2 | 10 | 3 | 11 | 1 | 0 | 1 |
| **362** | 1 | 1 | 2 | 0 | 1 | 0 | 0 | 0 |

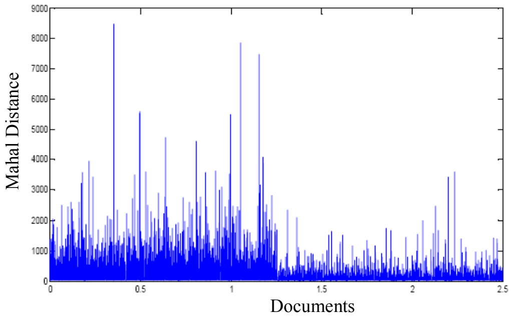


Figure 1. MD of documents with respect to the MS.

## 6. Experimental Results

Table 3 shows the performance of MD based Classifier (MDC) on the benchmark dataset [7]. This benchmark data set is referred as Large movie review Data Set (LDS) in this paper. LDS contains 25000 reviews for training and 25000 reviews for testing.

To evaluate the performance of MDC on various sizes of data set, small subsets of LDS like LDS400, LDS2000 etc. have been used. In each subset, equal number of positive and negative reviews was used.

The performance of MDC showed in Table 3 is based on the rules developed using 400 reviews from the total set of 25000 movie reviews. Performance of this classifier can be improved by adding more rules captured from the misclassified reviews. To prove this fact, an attempt has been made to understand the performance of MDC with other popular ML algorithms as a hybrid classifier.

Table 3. Performance of MDC.

| S.No | Dataset | Precision | Recall | F-Measure | Accuracy |
|---|---|---|---|---|---|
| 1 | LDS400 | 0.86 | 0.72 | 0.78 | 0.803 |
| 2 | LDS2000 | 0.75 | 0.73 | 0.74 | 0.74 |
| 3 | LDS11000 | 0.72 | 0.7 | 0.71 | 0.715 |
| 4 | LDS25000 | 0.72 | 0.68 | 0.70 | 0.708 |
| Average | | **0.76** | **0.71** | **0.73** | **0.74** |

The reviews that have been wrongly classified by the MDC have been classified using ML algorithms like Naïve Bayes (NB), Bayesian Logistic Regression (BLR), Multi Layer Perceptron (MLP), SMO and Classification and Regression Tree (CART). The results are amazing with a performance of above 95% of accuracy for MLP and CART. Even the NB classifier performed with 81.4 % of accuracy.

Table 4 shows the classification accuracy of various ML classifiers available in WEKA 3.6.3 for the 25000 review dataset. The MDC has wrongly classified 7277 documents of LDS25000, in which, 3366 are negative documents misclassified as positive and 3911 are positive documents misclassified as negative. These misclassified documents were then classified using the ML classifiers and the results are shown in Table 5. All the results are based on a ten-fold cross validation.

Table 4. Performance of various classifiers for 25000 reviews (12500 positive and 12500 negative).

| Classifier | Precision | Recall | F-Measure | Accuracy |
|---|---|---|---|---|
| NB | 0.71 | 0.85 | 0.77 | 0.75 |
| BLR | 0.8 | 0.8 | 0.8 | 0.8 |
| MLP | 0.78 | 0.81 | 0.8 | 0.8 |
| SMO | 0.79 | 0.8 | 0.8 | 0.8 |
| CART | 0.78 | 0.77 | 0.77 | 0.77 |

Table 5. Performance of various ML classifiers on the documents misclassified by MDC total reviews -7277 (3911 positive and 3366 negative).

| Classifier | Precision | Recall | F-Measure | Accuracy |
|---|---|---|---|---|
| NB | 0.942 | 0.733 | 0.824 | 0.814 |
| BLR | 0.886 | 0.854 | 0.870 | 0.877 |
| MLP | 0.947 | 0.964 | 0.955 | 0.959 |
| SMO | 0.899 | 0.842 | 0.869 | 0.875 |
| CART | 0.948 | 0.944 | 0.946 | 0.950 |

For the further discussion on the performance of MDC as a hybrid classifier, only the performance of MLP has been considered as it performed with 95.9% of accuracy. Of the 3366 negative documents 3189 have been correctly classified and of the 3911 positive documents 3791 have been correctly classified. Thus a total of 24699 documents have been correctly classified by the hybrid of MDC and MLP together. This amounts to an accuracy of 98.8%.

In general, in order to understand the performance of any proposed classifier, the datasets with class label are used. In the real-time environment, the proposed classifier should classify the document and assign a class label. We can depend on the results of any classifier, if it performs with very high-accuracy on the datasets with class label. In this paper, the proposed MDC performed with an average accuracy of 0.74 and we wanted to understand how it can be enhanced.

Separating out the types of reviews within a set of reviews [2] helps to treat them separately. A close observation of the proposed hybrid method reveals that, those documents that were wrongly classified by the MDC form a unique family. This is the reason for the very high classification accuracy of MLP. It should be noted that, MLP performed with 79.2% of accuracy while classifying the entire 25000 reviews. But for the 7277 reviews misclassified by MDC, which is actually a subset of 25000 reviews, it performed with 95.9% of accuracy. Also, MLP performed with 98% of accuracy on the documents that were correctly classified by MDC. This confirms the fact that the misclassified documents of MDC form a unique family.

The hybrid experiment was carried out to demonstrate the fact that the misclassified reviews form a unique family i.e., the sentence patterns in these reviews are distinct from the sentence patterns found in the correctly classified reviews. Also, this hybrid experiment helped us to understand the fact that the number of rules that were framed from a total of 400 reviews consisting of 200 positive and 200 negative reviews for the creation of RTDM is not sufficient and hence the average accuracy of the proposed MDC is 0.74. Accuracy of the proposed MDC can be enhanced by framing more rules by reading the misclassified documents for capturing additional rule patterns and adding them into the rule lexicon. If the rule lexicon is exhaustive with all the possible patterns of review writing, the proposed MDC's performance will be enhanced and hence the proposed MDC can be used as a stand-alone classifier for classifying the real reviews. This forms the scope for future research in this topic.

## 7. Scope for Future Research

In this research, the trial and error method was used for identifying the best MS, which is a time consuming task. Further research can be carried out on the methodology for identifying the MS. In the present work, sentiment analysis is done at the document level but a separate examination of the reviews that were misclassified by the MDC revealed that a sentence level analysis would result in a still better accuracy. Whenever in a sentence both positive and negative opinion words are present, it leads to the misclassification. For example, consider the following review:

"…You can quibble about its clichés, predictability, and rare moments of overcooked sappiness, but none of that takes away from the entertainment value."

This review would certainly be interpreted as positive by humans, but it is very challenging to get the same output from a computer due to the presence of strong negative words. Special kinds of rules must be developed to overcome this issue and the analysis must be carried out at the sentence level using MD. This could be a separate research in itself.

## 8. Conclusions

In this research paper, an attempt has been made to combine the best of rule based classification and machine learning approaches to achieve a better accuracy, precision, recall and F-measure. The important aspect of this research is the creation of RTDM from the text data and application of MD as a measure to classify the reviews. In any conventional approach using data mining techniques, after the extraction of words and phrases from all the review documents, Singular Value Decomposition (SVD) technique is used to obtain a reduced dimension matrix. In this research, we have succeeded in creating a reduced dimension matrix with just eight representative dimensions. The proposed MDC performed with classification accuracy of 70.8% and as a hybrid, MDC along with MLP performed with an amazing classification accuracy of 98.8%, which is certainly the highest ever reported classification accuracy for a dataset containing 25000 reviews.

## References

[1] Amine A., Elberrichi Z., and Simonet M., "Evaluation of Text Clustering Methods using WordNet," *the International Arab Journal of Information Technology*, vol. 7, no. 4, pp. 349-357, 2010.

[2] Dave K., Lawrence S., and Pennock D., "Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews," *in Proceedings of the 12th International WWW Conference*, Budapest, Hungary, pp. 519-528, 2003.

[3] Gamon M., "Sentiment Classification on Customer Feedback Data: Noisy Data, Large Feature Vectors and the Role of Linguistic Analysis," *in Proceedings of the 20th International Conference on Computational Linguistics*, Geneva, Switzerland, pp. 841-847, *2004*.

[4] Hiroshi K., Tetsuya N., and Hideo W., "Deeper Sentiment Analysis using Machine Translation Technology," *in Proceedings of the 20th International Conference on Computational Linguistics*, Geneva, Switzerland, pp. 494-500, 2004.

[5] Konig A. and Brill E., "Reducing the Human Overhead in Text Categorization," *in Proceedings of the 12th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, Pennsylvania, USA, pp. 598-603, 2006.

[6] Liu B., *Web Data Mining Exploring Hyperlinks, Contents and Usage Data*, *Springer*, 2008.

[7] Mass A., Daly R., Pham P., Huang D., Ng A., and Potts C., "Learning Word Vectors For Sentiment Analysis," *in Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, Oregon, USA, pp. 142-150, 2011.

[8] Nasukawa T. and Yi J., "Sentiment Analysis: Capturing Favorability using Natural Language Processing," *in Proceedings of the 2nd International Conference on Knowledge Capture*, Florida, USA, pp. 70-77, 2003.

[9] Pang B. and Lee L., "A Sentiment Education: Sentiment Analysis using Subjectivity Summarization based on Minimum Cuts," *in Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, Barcelona, Spain, pp. 271-278, 2004.

[10] Pang B. and Lee L., "Seeing Stars: Exploiting Class Relationships for Sentiment Categorization with Respect to Rating Scales," *in Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, University of Michigan, USA, pp. 115-124, 2005.

[11] Prabowo R. and Thelwall M., "Sentiment Analysis: A Combined Approach," *the Journal of Informetrics*, vol. 3, no. 2, pp. 143-157, 2009.

[12] Srinivasagupta N., Valarmathi B., and Joseph S., "Sentiment Analysis using Representative Terms-a Grouping Approach for Binary Classification of Documents," *the Journal of Theoretical and Applied Information Technology*, vol. 44, no. 2, pp. 161-165, 2012.

[13] Taguchi G. and Jugulum R., *The Mahalanobis-Taguchi Strategy: A Pattern Technology System*, John Wiley and Sons, 2002.

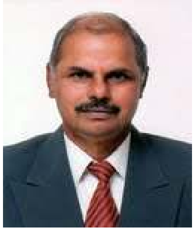
[14] Turney P., "Thumbs up or Thumbs Down? Sentiment Orientation Applied to Unsupervised Classification of Reviews," *in Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, Philadelphia, USA, pp. 417-424, 2002.

[15] Valarmathi B. and Palanisamy V., "Opinion Mining of Customer Reviews using Mahalanobis-Taguchi System," *the European Journal of Scientific Research*, vol. 62, no. 1, pp. 95-100, 2011.

[16] Yi J., Nasukawa T., Niblack W., and Bunescu R., "Sentiment Analyzer: Extracting Sentiments about a Given Topic using Natural Language Processing Techniques," *in Proceedings of the 3rd IEEE International Conference on Data Mining*, Florida, USA, pp. 427-434, 2003.

**Valarmathi Balasubramanian** is Associate Professor in the school of Information Technology at VIT University, India. She holds PhD degree in computer science from Anna University, India. Valarmathi has more than 20 years of experience in teaching and research. Currently, she supervises five PhD students in big-data analysis, sentiment mining and pattern recognition. She has coauthored a text book on total quality management.

**Srinivasa Gupta Nagarajan** is Assistant Professor in the school of Mechanical and Building Sciences at VIT University, Vellore, India. He holds MTech degree in Industrial Management from IIT Madras, India, and currently pursuing research in the area of application of clustering tools for manufacturing management. Srinivasa has more than 17 years of experience in teaching and industrial consultancy. He is a certified six sigma Black Belt and a certified a Master Trainer. He has written a text book on Total Quality Management.



**Palanisamy Veerappagoundar** is Principal, Info Institute of Engineering, Coimbatore, India. Palanisamy has more than 40 years of experience in teaching and research. He holds a PhD in communication-antenna theory from Indian Institute of Technology, India. Currently, he supervises eight PhD students in big-data analysis, sentiment mining, and pattern recognition.