CHAPTER 4

DESIGN AND DEVELOPMENT OF MAHALANOBIS DISTANCE BASED CLASSIFIER (MDC)

4.1 INTRODUCTION

In 1930, P.C. Mahalanobis, founder of the Indian Statistical Institute, introduced a statistical measure called the Mahalanobis distance (MD). MD is a superior statistical measure than the other statistical measures like Euclidean distance and Manhattan distance used for clustering and classification because it is based on the correlation among the various dimensions of the given problem (Genichi Taguchi and Rajesh Jugulum 2002).

4.2 MD FOR PATTERN RECOGNITION

Genichi Taguchi and Rajesh Jugulum (2002) popularized the Mahalanobis-Taguchi system (MTS) and proposed a method for using the MD for the pattern recognition problems. They established that for a pattern recognition problem, if a reference set can be created using the characteristic dimensions of the problem, then using the reference set, the test set can be classified either it belongs to the family of reference set or not by calculating the MD between the test set and the reference set. The reference set is called as Mahalanobis space (MS).

For example, in a medical diagnosis system, the improvement after a medication can be checked by creating the MS using the data of a group of healthy people and then the MD between the MS and test data will depict any

improvement has happened. It means, if the medication has resulted in a health improvement, then the MD between the patient's test data and MS will be less and more otherwise.

4.2.1 Calculation of MD

This section is included just to explain how MD can be calculated without using any software. For the analysis in this thesis, MATLAB software has been used.

MD can be calculated using the Equations (4.1) and (4.2)

$$MD = 1/K \, (Z_{ij} \, C^{-1} Z_{ij}^{T}) \qquad (4.1)$$

Where K = Number of features or variables

Z_{ij} = Standardized matrix

Z_{ij}^{T} = Transpose of Z_{ij}

C^{-1} = Inverse of the correlation matrix of Z_{ij}

In the event of high correlations, inverse of the correlation matrix cannot be calculated and hence the MD. In such cases, Equation (4.2) should be used for calculating the MD, which is based on the Gram-Schmidt process (Genichi Taguchi and Rajesh Jugulum (2002)).

$$MD_j = 1/k \, (u_{1j}^{2}/s_1^{2} + u_{2j}^{2}/s_2^{2} + \ldots\ldots.. + u_{kj}^{2}/s_k^{2}) \qquad (4.2)$$

Where 'j' denotes the document number, 1 to n

'k' denotes the number of features, 1 to k

's' denotes the standard deviation of the column of the orthogonal matrix found using GSP

Note: MATLAB software has built-in function "mahal (Y, X)" for calculating the MD.

4.3 SENTIMENT DETECTION USING MD

In this section, how MD is used to decide whether a review document expresses positive sentiment or negative sentiment is discussed. The steps involved in carrying out the document-level sentiment analysis using MD are as explained in the subsequent sections.

4.3.1 RTDM Creation from the Review Documents

RTDM of the given set of review documents is created by following the procedure given in section 3.2. The RTDM of 403 movie reviews (LDS403) is given in Appendix 1. The entries in the RTDM represent the frequency of RT in a review document. The RTDM created for the other sizes of LDS dataset viz. LDS2000, LDS11000, and LDS25000 are not included in the appendix due to the space constraint.

4.3.2 Selection of Mahalanobis Space (MS)

The MS must be carefully chosen from the RTDM. MS can be either from the positive side or from the negative side i.e. if RTDM is created for a review dataset consisting of 403 reviews (202 negative and 201 positive), then the MS can be chosen from the first 202 rows representing the negative reviews or from the later 201 rows representing the positive reviews. The size of MS will be many times smaller than the RTDM.

Selecting a set of rows that can act as MS for calculating MD is really the critical task, as it decides the classification performance. The MS chosen for our analysis of the movie reviews is shown in Table 4.1 and the abridged version of RTDM of LDS403 is shown in Table 4.2. At present, though there is no standard method available for selecting the MS from the RTDM, care should be taken to ensure that the MS chosen has a good discriminatory power. Hence it is decided based on the classification performance with various sizes of MS from both the positive and the negative side.

For the movie review analysis (reviews of LDS), the MS shown in Table 4.1 resulted in a better classification.

Table 4.1 The MS chosen for analysis of movie reviews

Document No.	RT								
	Good	Very Good	Excellent	Recommended	Bad	Very Bad	Disgusting	Never Recommended	MD
350	1	1	4	0	1	0	0	0	10.291
351	2	1	2	2	2	0	1	0	4.3326
352	9	13	7	3	13	2	3	1	135.95
353	0	0	3	2	1	0	1	0	3.7665
354	5	4	5	1	2	2	0	1	30.868
355	3	1	3	1	3	0	0	0	10.509
356	4	1	7	0	3	0	0	1	27.449
357	1	1	4	4	0	0	0	0	13.09
358	2	1	6	2	1	0	1	0	21.789
359	4	2	3	0	3	2	0	0	13.162
360	5	2	10	3	11	1	0	1	48.759
361	1	1	2	0	1	0	0	0	3.3445
								Average	26.943

Table 4.2 Abridged RTDM of LDS403

Document No.	RT							
	Good	Very Good	Excellent	Recommended	Bad	Very Bad	Disgusting	Never Recommended
1	2	0	0	0	3	5	2	2
2	1	3	2	0	3	4	4	1
3	2	2	3	1	6	2	2	0
4	4	2	0	0	7	1	1	1
5	6	0	1	0	1	2	3	2
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
399	0	3	4	0	3	0	0	0
400	10	2	3	1	6	2	2	0
401	0	0	1	0	1	0	0	0
402	0	1	6	1	1	0	1	1
403	6	2	6	3	3	0	0	1

4.3.3 Calculation of MD

MD can be calculated using the "mahal (Y, X)" function available in MATLAB. This function will calculate the MD for each row of matrix Y from the sample in matrix X. In the sentiment classification problem, the RTDM is the matrix Y and the MS is the matrix X. The MD values of the 403 documents are shown in Appendix 1. For the other sizes of RTDM, the MD values are not given as they would occupy more space. Table 4.3 shows the abridged RTDM of LDS403 with MD of each review with respect to the MS.

Table 4.3 Abridged RTDM of LDS403 with MD of each review

Document No.	RT								
	Good	Very Good	Excellent	Recommended	Bad	Very Bad	Disgusting	Never Recommended	MD
1	2	0	0	0	3	5	2	2	304.39
2	1	3	2	0	3	4	4	1	222.23
3	2	2	3	1	6	2	2	0	37.861
4	4	2	0	0	7	1	1	1	59.108
5	6	0	1	0	1	2	3	2	212.69
.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.
399	0	3	4	0	3	0	0	0	28.691
400	10	2	3	1	6	2	2	0	128.1
401	0	0	1	0	1	0	0	0	9.9422
402	0	1	6	1	1	0	1	1	32.118
403	6	2	6	3	3	0	0	1	29.474

4.3.4 Determination of "Threshold MD" for Sentiment Classification

The threshold MD value is decided based on the least misclassification point (or the maximum accuracy point). This is also done by trial and error. The average MD value of documents in MS is the starting

point and from that value, using the trial and error method the least misclassification point is found. For example, the average MD of MS shown in Table 4.1 is 26.943. First using this as the threshold MD, the classification is done, and then by increasing the value of threshold the classification accuracy is checked. If the accuracy has improved compared to the initial accuracy, then the threshold value is increased until the maximum classification accuracy is reached. In the case of movie reviews, for a threshold value of 56, the maximum accuracy was reached. After this value, the accuracy started decreasing.

4.3.5 Sentiment Classification

Classification is done based on the fact that if two reviews are similar or near similar the MD between them will be less and vice versa. For the movie review analysis, the MS is chosen from the positive reviews. So if a test document is positive, its MD will be less than or equal to the threshold MD. If the test document is negative, its MD will be more than the threshold value.

4.4 RESULTS

Table 4.4, Table 4.5, Table 4.6, Table4.7, Table 4.8 and Table 4.9 show the confusion matrix of MDC for Camera reviews, cell phone reviews, LDS403, LDS2000, LDS11000 and LDS25000 respectively. Table 4.10 shows the classification performance of MDC on various datasets. The performance of MDC for Camera and Cell phone reviews is comparatively better than the performance for movie reviews.

Table 4.4 Confusion matrix for camera reviews (MDC)

	Classified Negative	Classified Positive
Actual Negative	74	48
Actual Positive	30	386

Table 4.5 Confusion matrix for cell phone reviews (MDC)

	Classified Negative	Classified Positive
Actual Negative	70	52
Actual Positive	62	354

Table 4.6 Confusion matrix for LDS403 (MDC)

	Classified Negative	Classified Positive
Actual Negative	179	23
Actual Positive	56	145

Table 4.7 Confusion matrix for LDS2000 (MDC)

	Classified Negative	Classified Positive
Actual Negative	749	251
Actual Positive	268	732

Table 4.8 Confusion matrix for LDS11000 (MDC)

	Classified Negative	Classified Positive
Actual Negative	3998	1502
Actual Positive	1633	3867

Table 4.9 Confusion matrix for LDS25000 (MDC)

	Classified Negative	**Classified Positive**
Actual Negative	9130	3370
Actual Positive	3911	8589

Table 4.10 Classification performance of MDC on various datasets

S.No	**Dataset**	**P**	**R**	**F-Measure**	**A**
1	Camera	0.89	0.93	0.91	0.855
2	Cell phone	0.88	0.89	0.89	0.788
3	LDS403	0.86	0.72	0.78	0.803
4	LDS2000	0.75	0.73	0.74	0.74
5	LDS11000	0.72	0.7	0.71	0.715
6	LDS25000	0.72	0.68	0.70	0.708

P-Precision; R-Recall; A-Accuracy

Figure 4.1 shows the graph plotted for the Camera reviews. The X axis represents the number of documents and the Y axis represents the corresponding MD value. In the Camera reviews the first 122 reviews belong to the negative category and the remaining 416 reviews belong to the positive category. The MS for the analysis was chosen from the negative reviews and the threshold MD was 4.5. From the figure it can be clearly seen that, the negative reviews have small MD and the positive with larger MD, thus clearly paving the way for classification of review documents.

Figures 4.2, 4.3, 4.4, 4.5 and 4.6 shows the MD plot of cell phone reviews, LDS403, LDS2000, LDS11000 and LDS25000 respectively. Table 4.11 shows the details of MS for each dataset and its polarity, threshold MD identified for sentiment classification for each dataset.

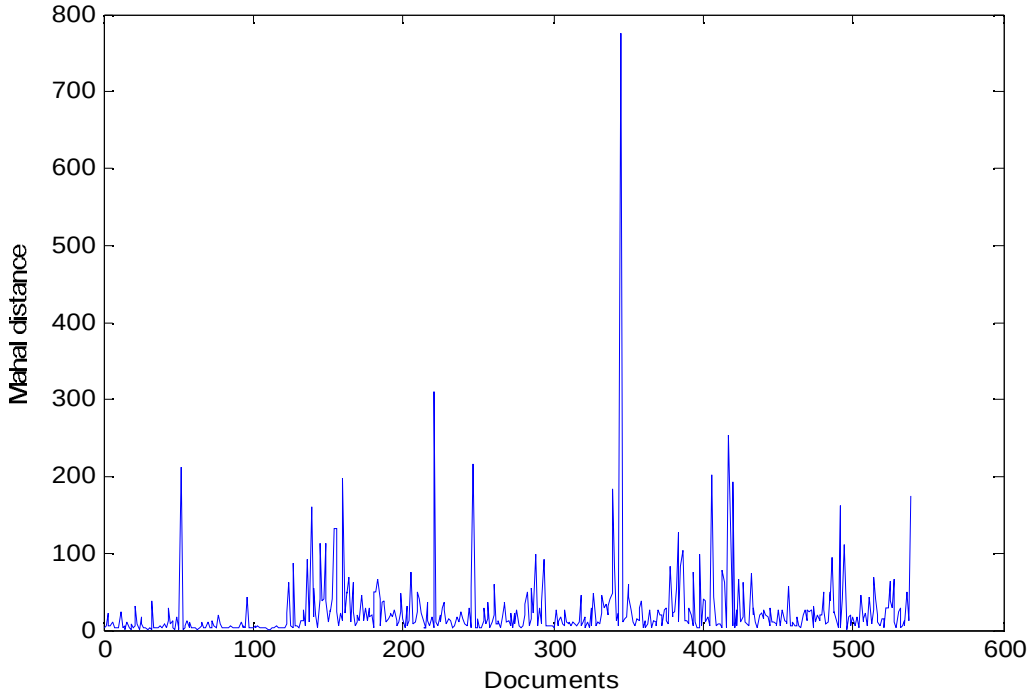

Figure 4.1 MD plot for camera reviews

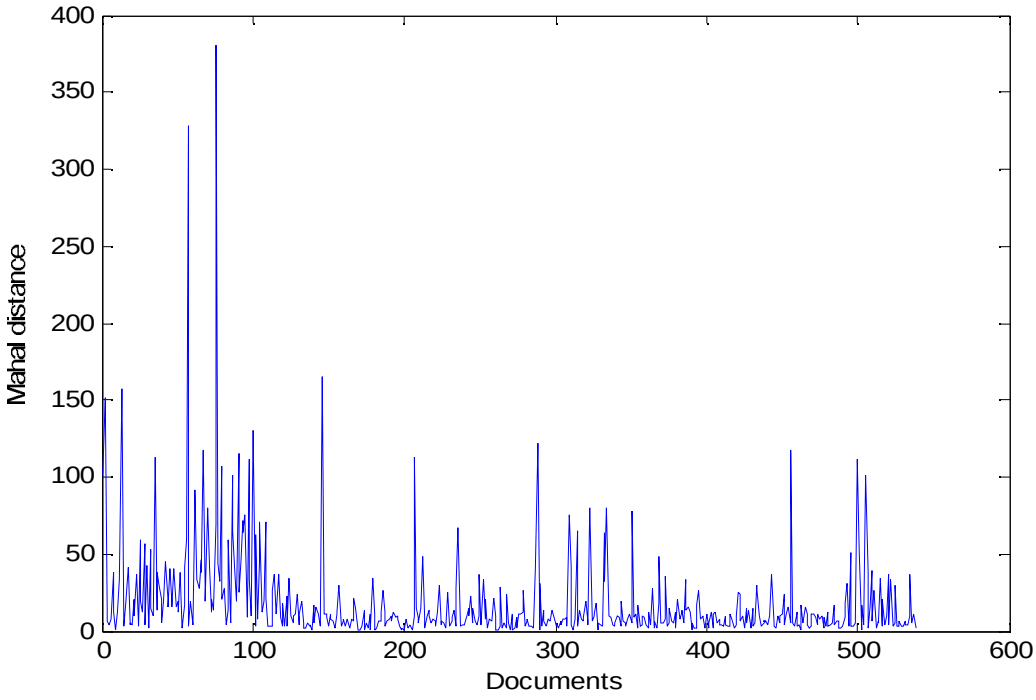

Figure 4.2 MD plot for cell phone reviews

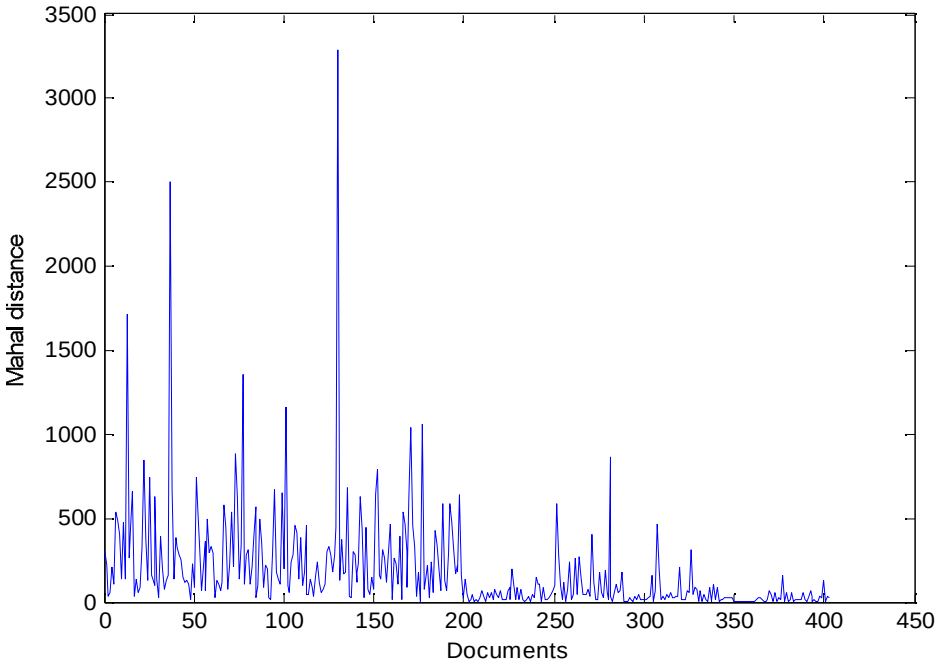

Figure 4.3 MD plot for LDS 403

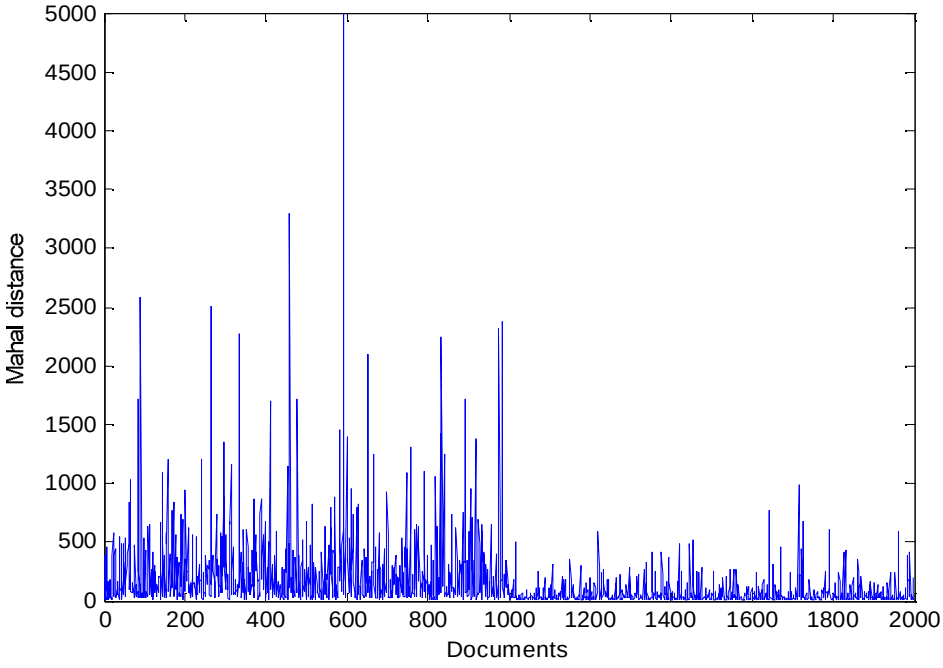

Figure 4.4 MD plot for LDS 2000

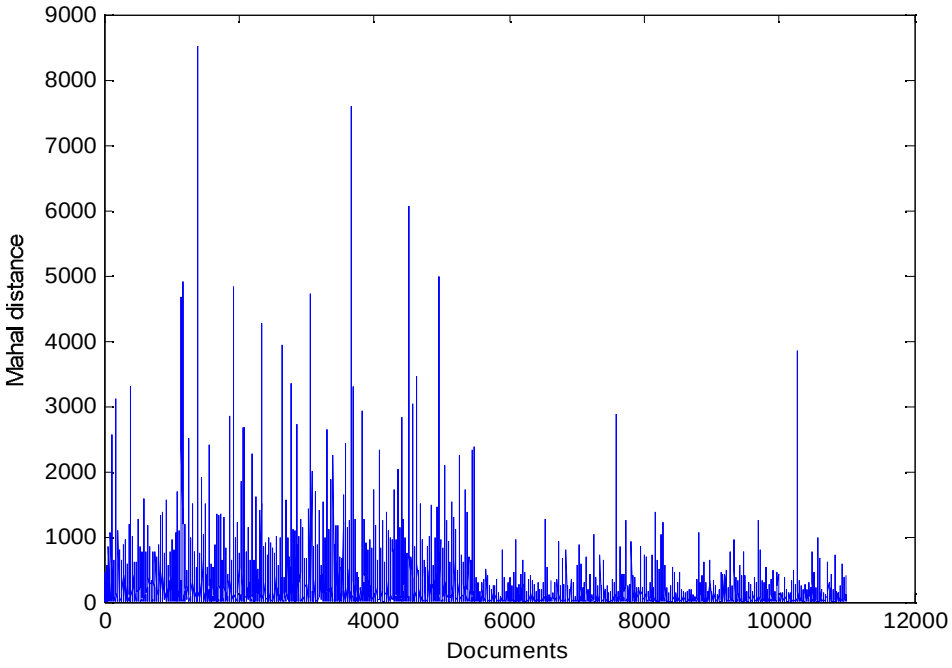

Figure 4.5 MD plot for LDS 11000

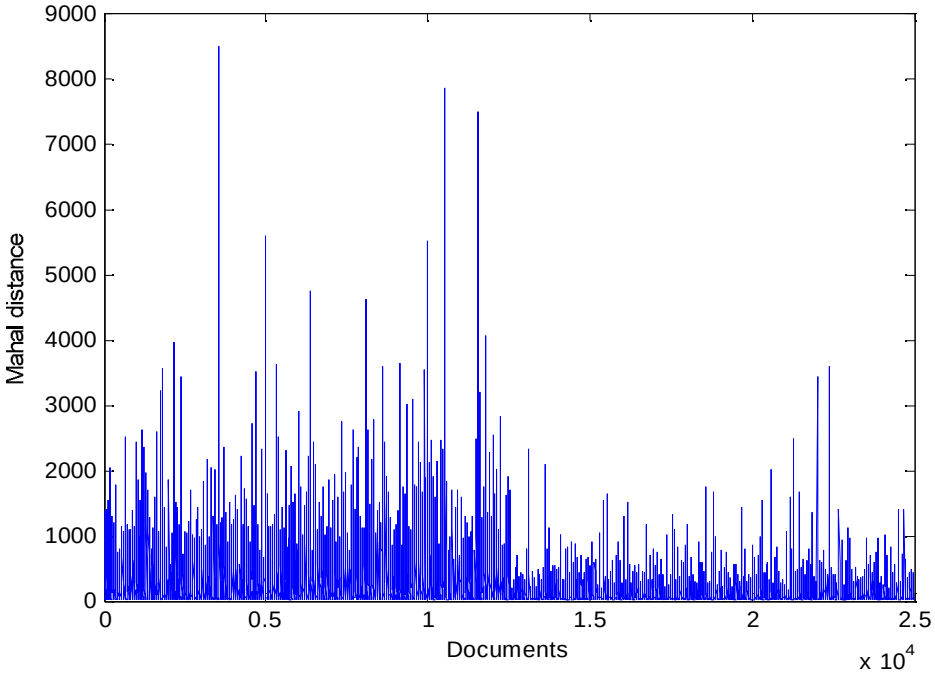

Figure 4.6 MD plot for LDS 25000

Table 4.11 Details of MS for various datasets

S. No	Dataset	MS (Document Number)	MS belong to Positive /Negative	Threshold
1	Camera (document No. 1-122 negative, 123-538 positive)	1-50	Negative	4.5
2	Cell phone (document No.1-122 negative, 123-538 positive)	151-200	Positive	19
3	LDS 403 (document No.1-202 negative, 203-403 positive)	350- 361 of LDS 403	Positive	56
4	LDS 2000 (document No. 1-1000 negative, 1001-2000 positive)	350- 361 of LDS 403	Positive	56
5	LDS 11000 (document No.1-5500 negative, 5501-11000 positive)	350- 361 of LDS 403	Positive	56
6	LDS 25000 (document No.1-12500 negative, 12501-25000 positive)	350- 361 of LDS 403	Positive	56

4.5 DISCUSSIONS

Identification of MS is the most important task for the sentiment analysis using MDC. At present, no prescribed method is available for selecting the MS. In the case of medical diagnosis using MD, it is possible to prescribe a set of features representing a healthy condition and the data is common for all the people to be declared as healthy. Hence the data of healthy people can be used as MS. The MD can be calculated between the test set and the sample in MS. If the test set is from a healthy person, the MD will be very small and it the test set is from an unhealthy person, the MD will be large.

Sentiment analysis is a pattern recognition task of a special kind. In a positive review, there will be certainly some features negatively commented and in a negative review, there will be some features with a positive comment. This makes the problem of sentiment classification an interesting

yet challenging issue. Selecting a Mahalanobis Space for a medical application is straightforward as the healthy data can be clearly defined but for sentiment analysis it is done by trial and error method, which is a time consuming process. It must be clearly noted that, the MS can be from either positive review or from the negative review. If MS is selected from the negative reviews, the MD between the negative documents and samples in the MS will be small and positive documents MD will be large. From the Figure 4.1 it can be clearly seen that the MD values of negative documents are small compared to the positive documents. This is due to the fact that, for the camera reviews, the MS chosen from the negative documents yielded a better accuracy.

4.6 IMPORTANT OBSERVATIONS

- RTDM creation is based on the rules manually coded for each domain. For example, in this research, three different types of reviews are considered for the sentiment analysis viz. Camera, Cell phone and movie reviews. Hence three set of rules specific to each category of reviews have been coded manually.
- The classification performance of the MDC is high for camera and cell phone reviews, whereas for the movie reviews, the performance is moderate. This may be due to the number of reviews read, and the number of rules coded for capturing the RT. Camera and Cell phone reviews consisted of only 538 reviews in total. The rules were written after carefully reading 200 reviews containing both positive and negative reviews and hence the higher accuracy.
- The movie reviews are complex and there are lots of varieties of movies like comedy, action, thriller etc. Each type of movie will be described using some set of words and hence the rules must be written after carefully reading all types of reviews.

- The rules for classifying the movie reviews from LDS were written based on just 200 positive reviews and 200 negative reviews. LDS is a specially prepared dataset that contains only 30 reviews per movie and all are highly polar reviews. MDC's accuracy of 70.8% on the LDS25000 dataset is based on the rules written by reading just 400 reviews. The accuracy can be greatly increased by adding more rules by thoroughly reading sufficient reviews.
- The proposed system can handle the direct opinions only. Opinion expressed directly on an entity/aspect is known as direct opinion. For example, "The picture quality is great" is a direct opinion. In this example, our proposed system will capture the word "great" and an appropriate RT will be assigned and then using the MD value the classification can be performed. At the present level, the proposed system can not handle the comparative opinions. The current study aims at the sentiment detection at document level, which assumes that the opinion is expressed on a single entity only.
- The proposed system does not consider time when the opinion is expressed. In general an opinion is a quintuple, ($\mathbf{e_i}$, $\mathbf{a_{ij}}$, $\mathbf{S_{ijkl}}$, $\mathbf{h_k}$, $\mathbf{t_l}$), where $\mathbf{e_i}$ is the name of an entity, $\mathbf{a_{ij}}$ is an aspect of e_i, $\mathbf{S_{ijkl}}$ is the sentiment on a_{ij} of e_i, $\mathbf{h_k}$ is the opinion holder, and $\mathbf{t_l}$ is the time when the opinion is expressed. **For a document level sentiment analysis**, the general format of the problem is expressed by the following quintuple, **(_, General, S, _, _)**. The entity "e", opinion holder "h" and time when the opinion is expressed "t" are assumed known or irrelevant (Bing Liu 2012). It should be noted that in document level sentiment classification system no specific aspect of the entity is analyzed, only the overall sentiment about the entity is determined, which is mentioned as "**General**" in the quintuple.