

CHAPTER 6

DESIGN AND DEVELOPMENT OF MAHALANOBIS-FUZZY SYSTEM CLASSIFIER (MFSC) FOR SENTIMENT ANALYSIS

6.1 INTRODUCTION

Representing the text documents in the RTDM format enabled the development MDC and FLC. The classification performance of them was discussed in Chapter 4 and Chapter 5 respectively. The proposed Mahalanobis Fuzzy System Classifier (MFSC) has been designed by utilizing the best of MDC and FLC. The steps involved in the process of sentiment classification by MFSC are exactly the same as MDC. The only difference is the entries in the RTDM. For the MDC, the RTDM entries represent the frequency of RT in a document. For the MFSC, the frequencies of RT are converted as a fuzzy score based on the procedure described in section 5.4.5. RTDM with fuzzy score (FS) entries will be denoted as RTDM-FS in the subsequent discussions.

6.2 SENTIMENT CLASSIFICATION USING MFSC

As mentioned in the previous paragraph, the sentiment classification is carried out exactly similar to the procedure followed in MDC. The final sentiment class assignment is based on the threshold-MD score. MFSC performed well with the MS selected from the negative reviews.

For this analysis, instead of using eight RT, only four of them were considered viz. Excellent, Recommended, Disgusting and Never recommended. Hence the RTDM-FS has only four columns and the rows represent the documents as usual.

Table 6.1 is the abridged version of RTDM-FS for the dataset LDS403. The MS selected for determining the MD of each review document for the MFSC is shown in Table 6.2. Table 6.3 shows the abridged version of RTDM-FS with MD score for the LDS403 dataset.

Table 6.1 Abridged RTDM-FS for LDS403

Document No.	RT			
	Excellent	Recommended	Disgusting	Never Recommended
1	0.00	0.00	0.04	0.03
2	0.91	0.00	0.04	0.04
3	0.92	0.95	0.04	0.00
4	0.00	0.00	0.05	0.04
5	0.91	0.00	0.04	0.03
.
.
.
.
.
399	0.94	0.00	0.00	0.00
400	0.91	0.95	0.05	0.00
401	0.95	0.00	0.00	0.00
402	0.95	0.95	0.05	0.04
403	0.93	0.96	0.00	0.04

Table 6.2 MS for the Mahalanobis-Fuzzy system classifier

Document No.	RT				
	Excellent	Recommended	Disgusting	Never Recommended	MD
874	0.00	0.00	0.04	0.03	0.00
875	0.91	0.00	0.00	0.04	0.22
876	0.91	0.95	0.05	0.04	0.57
877	0.91	0.00	0.04	0.03	0.23
878	0.91	0.00	0.04	0.04	0.23
879	0.92	0.00	0.04	0.03	0.23
880	0.91	0.00	0.04	0.04	0.23
881	0.91	0.00	0.03	0.00	0.22
882	0.93	0.97	0.00	0.00	0.58
883	0.91	0.00	0.05	0.00	0.23
884	0.00	0.96	0.04	0.04	0.24
885	0.00	0.96	0.05	0.04	0.24
886	0.93	0.00	0.00	0.04	0.23
887	0.91	0.00	0.04	0.00	0.22
888	0.00	0.00	0.00	0.03	0.00
889	0.00	0.00	0.04	0.00	0.00
Average					0.23

Table 6.3 Abridged RTDM-FS with MD score for LDS403 dataset.

Document No.	RT				
	Excellent	Recommended	Disgusting	Never Recommended	MD
1	0.00	0.00	0.04	0.03	3.50
2	0.91	0.00	0.04	0.04	1.94
3	0.92	0.95	0.04	0.00	6.65
4	0.00	0.00	0.05	0.04	4.08
5	0.91	0.00	0.04	0.03	1.64
.
.
.
.
.
399	0.94	0.00	0.00	0.00	5.42
400	0.91	0.95	0.05	0.00	6.73
401	0.95	0.00	0.00	0.00	5.43
402	0.95	0.95	0.05	0.04	5.40
403	0.93	0.96	0.00	0.04	6.95

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 6.2 is 0.23. 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 5.38, the maximum accuracy was reached. After this value, the accuracy started decreasing.

The sentiment classification is carried out as per the procedure mentioned in section 4.3.5. Since the MS is chosen from the negative reviews, the MD of the negative review will be less than or equal to the threshold-MD value and the positive review will have larger MD values.

6.3 RESULTS

Table 6.4, Table 6.5, Table 6.6, Table 6.7, Table 6.8 and Table 6.9 show the confusion matrix for camera reviews, cell phone reviews, LDS403, LDS2000, LDS11000 and LDS25000 respectively. Table 6.10 shows the classification performance of MFSC on various datasets. For camera and cell phone reviews MFSC performed with a comparatively higher accuracy than for the movie reviews.

Table 6.4 Confusion matrix for camera reviews (MFSC)

	Classified Negative	Classified Positive
Actual Negative	71	51
Actual Positive	17	399

Table 6.5 Confusion matrix for cell phone reviews (MFSC)

	Classified Negative	Classified Positive
Actual Negative	75	47
Actual Positive	71	345

Table 6.6 Confusion matrix for LDS403 (MFSC)

	Classified Negative	Classified Positive
Actual Negative	157	45
Actual Positive	37	164

Table 6.7 Confusion matrix for LDS2000 (MFSC)

	Classified Negative	Classified Positive
Actual Negative	746	254
Actual Positive	262	738

Table 6.8 Confusion matrix for LDS11000 (MFSC)

	Classified Negative	Classified Positive
Actual Negative	3979	1521
Actual Positive	1682	3818

Table 6.9 Confusion matrix for LDS25000 (MFSC)

	Classified Negative	Classified Positive
Actual Negative	8997	3503
Actual Positive	4141	8359

Table 6.10 Classification performance of MFSC on various datasets

S.No	Dataset	P	R	F-Measure	A
1	Camera	0.89	0.96	0.92	0.87
2	Cell phone	0.88	0.83	0.854	0.78
3	LDS 403	0.78	0.82	0.80	0.796
4	LDS 2000	0.74	0.74	0.74	0.742
5	LDS 11000	0.72	0.69	0.70	0.708
6	LDS 25000	0.70	0.67	0.68	0.694

P-Precision; R-Recall; A-Accuracy

Figure 6.1 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 7.7. From the figure it can be clearly seen that, the negative reviews have smaller MD compared to the positive documents, thus clearly paving the way for classification of review documents.

Figure 6.2, Figure 6.3, Figure 6.4, Figure 6.5 and Figure 6.6 shows the MD plot of cell phone, LDS 403, LDS 2000, LDS 11000 and LDS 25000 respectively. Table 6.11 shows the details of MS for each dataset and its polarity, threshold MD identified for sentiment classification for each dataset.

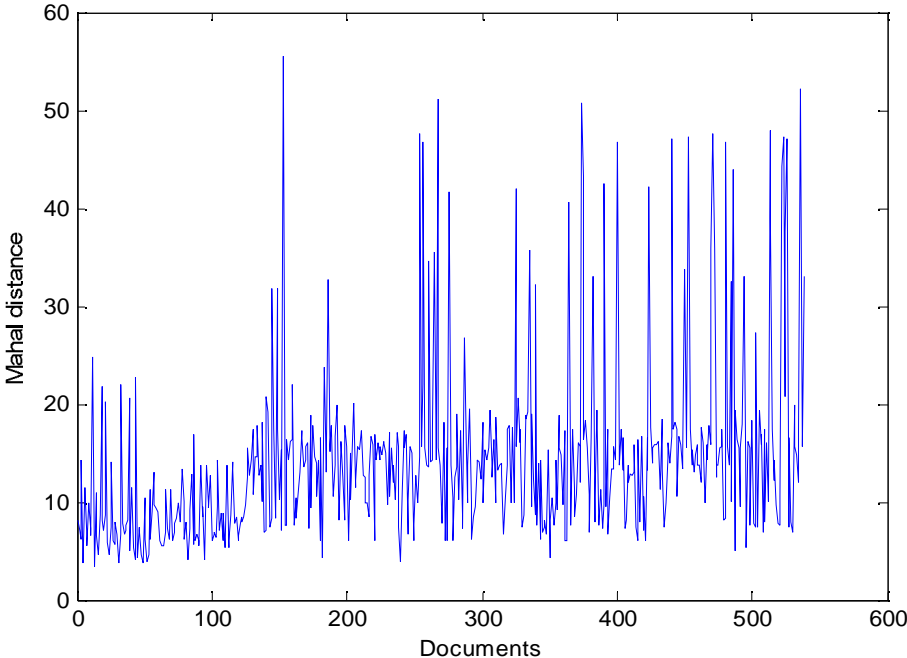


Figure 6.1 MD plot for camera reviews (MFSC)

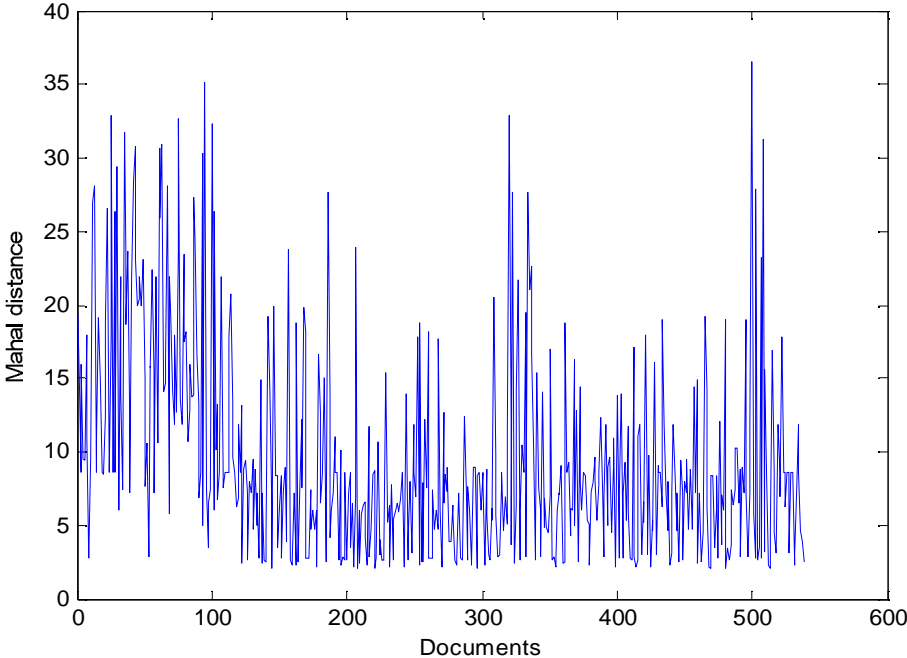


Figure 6.2 MD plot for cell phone reviews (MFSC)

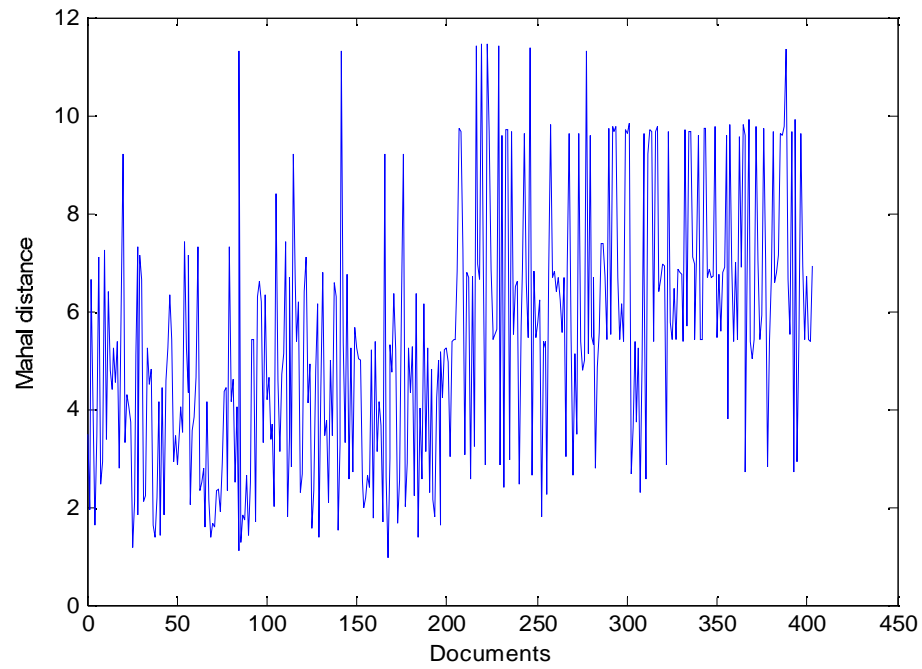


Figure 6.3 MD plot for LDS 403 (MFSC)

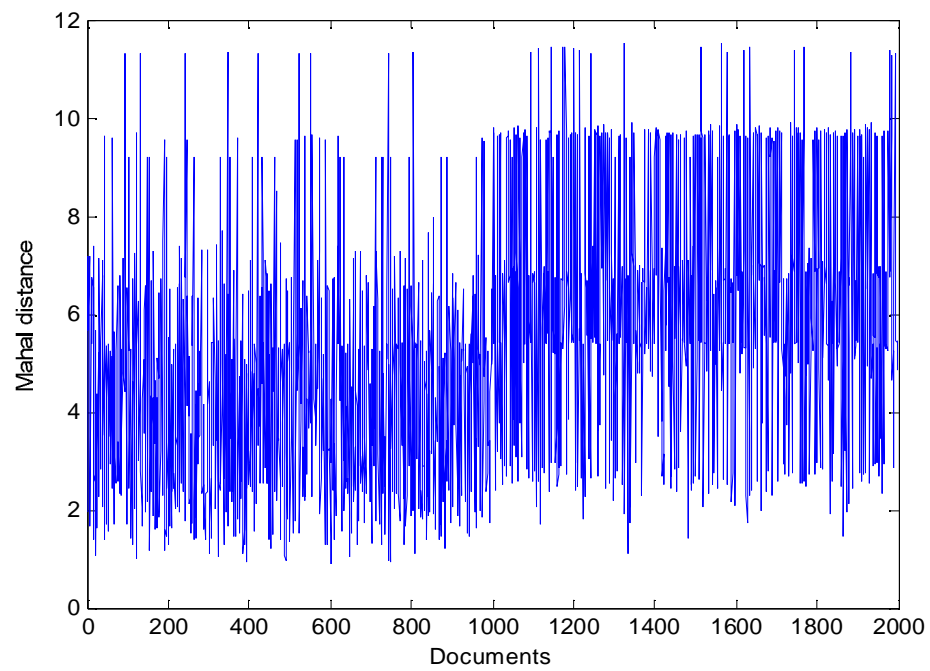


Figure 6.4 MD plot for LDS 2000 (MFSC)

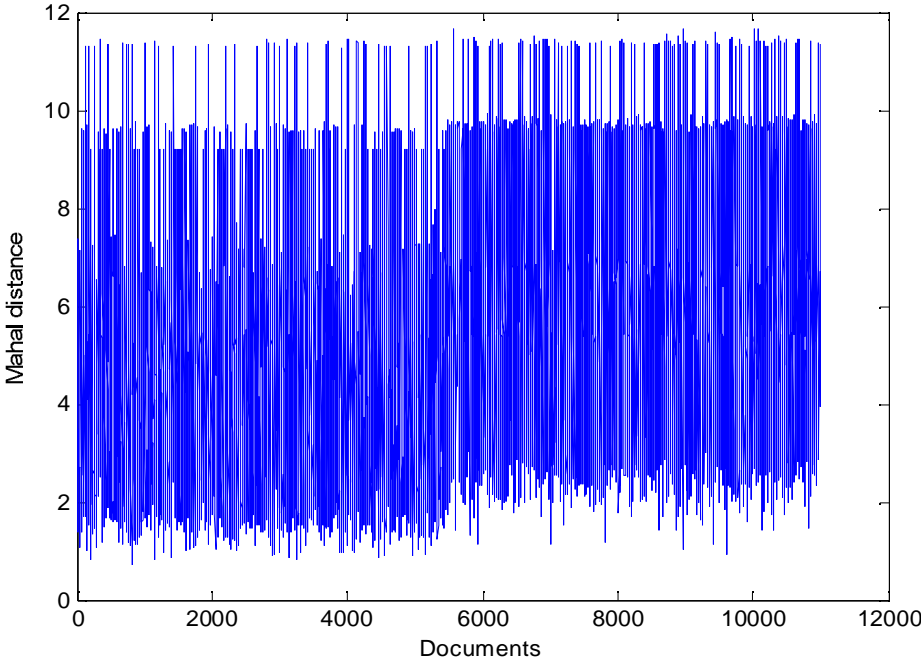


Figure 6.5 MD plot for LDS 11000 (MFSC)

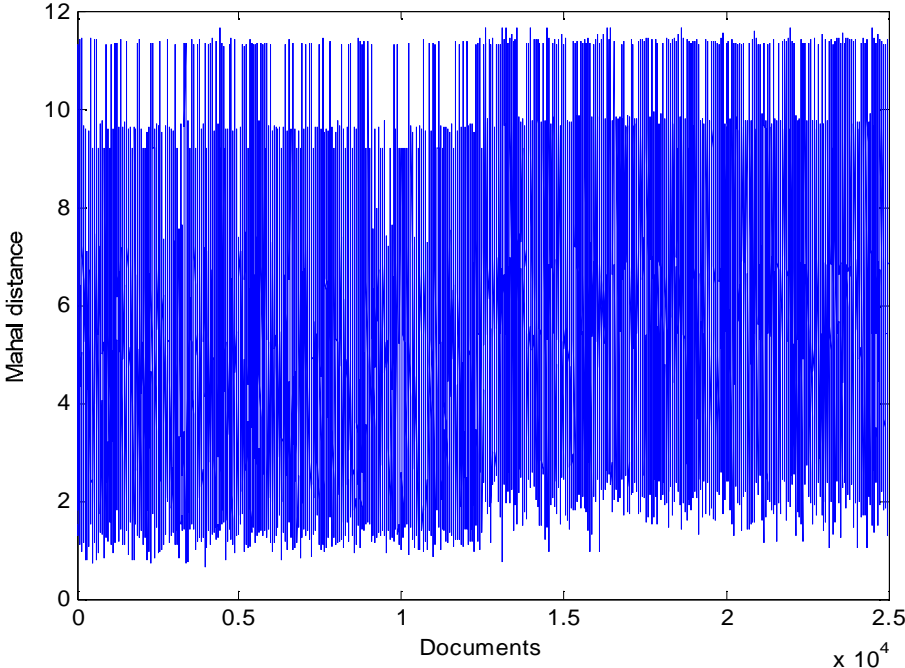


Figure 6.6 MD plot for LDS 25000 (MFSC)

Table 6.11 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)	1-60	Negative	7.7
2	Cell phone (document No.1-122 negative, 123-538 positive)	151-175	Positive	14
3	LDS 403 (document No.1-202 negative, 203-403 positive)	874-889 of LDS25000	Negative	5.38
4	LDS 2000 (document No. 1-1000 negative, 1001-2000 positive)	874-889 of LDS25000	Negative	5.38
5	LDS 11000 (document No.1-5500 negative, 5501-11000 positive)	874-889 of LDS25000	Negative	5.38
6	LDS 25000 (document No.1-12500 negative, 12501-25000 positive)	874-889 of LDS25000	Negative	5.38

6.4 DISCUSSIONS AND IMPORTANT OBSERVATIONS

- The performance of MFSC is slightly less than the performance of MDC. But it should be noted that, only four of the eight RT used in MDC were used in MFSC for achieving this result.
- The initial experiments carried out with all the eight RT resulted in a comparatively lower performance and hence the classification was tried with less number of RT.
- The reason for this better performance with less number of columns is yet to be ascertained and left to the future research.
- The performance of the classifier can be still improved by adding more rules for capturing the opinion phrases and opinion words.