CHAPTER 3

DEVELOPMENT OF REPRESENTATIVE TERM-DOCUMENT MATRIX

3.1 INTRODUCTION

In the field of Information Retrieval (IR), it is very common to represent a piece of text as a feature vector, in which the individual terms are assigned as features. When a collection of reviews are expressed as vectors, it is called as Term-Document Matrix (TDM), in which the entries correspond to the frequency of individual terms. Feature extraction algorithms are used for identifying the features of the text document. After identifying the features, singular value decomposition (SVD) technique is used to reduce the dimension of the term-document matrix. In a typical TDM, the rows represent the number of text documents and each column represents the features selected for the analysis after the dimension reduction using SVD.

Generally the features selected for analysis may be unigrams (single words), bi-grams (two words), tri-grams (three words) and in general n-grams. These are captured using the NLP tools. In this research, a new format is proposed for representing the text documents (review documents) using fewer features.

3.2 REPRESENTATIVE TERM-DOCUMENT MATRIX (RTDM) CREATION

The objective of sentiment analysis in this study is to assign a class for the review documents as either positive or negative based on the overall sentiment expressed in them. The rationale for the proposed format of representing the text documents using fewer features originated from how a human mind would perceive a review document based on the words and phrases used in it. Depending on the words and phrases used in a document, a reader would mentally assign any of the following eight sentiments viz. Good, Very good, Excellent, Recommended, Bad, Very bad, Disgusting, Never recommended. Within a document, a reviewer would have expressed multiple sentiments using words and phrases and hence the reader also would perceive those words and phrases among any of the sentiments mentioned. Hence the proposed matrix format consists only those eight categories of sentiments as its columns. These eight categories of sentiments are named as "Representative Terms" (RT).

For example, if a reader finds the word "exciting" in a review document, his mind would assign it to "excellent" category and if the phrase "worth watching" is found in the review, his mind would assign it to "recommended" category. A PERL program has been developed for extracting those sentiment bearing words and phrases from the review documents. The rules for capturing the sentiment bearing words and phrases have been manually coded by reading a sufficient number of review documents. The hypothesis here is that, only certain words and combination of them are repeatedly used for expressing the sentiments. Those patterns are captured and assigned to the appropriate RT in order to create the new matrix format for representing the review documents. The new matrix format is named as "Representative Term-Document Matrix" (RTDM).

Hence the RTDM consists of only eight columns viz. Good, Very good, Excellent, Recommended, Bad, Very bad, Disgusting, Never recommended. Each row of the RTDM denotes a review document with eight features representing it. The need for dimensional reduction does not arise here due to the fewer columns and certainly all of them are important.

A typical positive review will have entries in the first four columns and a typical negative review would have entries in the later four columns. In general, a review would consist of both positive and negative sentiments as the reviewer would comment on many aspects of the product or service. The reviewer may express happiness over some aspects and dissatisfaction over some other aspects. For example, in a movie review, a reviewer might express his happiness over the story and screen play and dissatisfaction over sound effects. Such reviews will have entries on all or some of the columns depending on the various sentiments expressed.

A sample set of manually coded rules used for capturing the relevant RT from the review documents are given below:

Note: The rules are given in the same format as found in the PERL program developed for this purpose. Since the “stop word” option is used in the PERL program, while coding the rules the stop words were not included as the program would capture the words/phrases without those stop words. The stop words are given in Table 3.1.

```
$_ =~ s/\s+joy watch\s+/ good /g;
$_ =~ s/\s+last forever\s+/ very_good /g;
$_ =~ s/\s+looks good\s+/ very_good /g;
$_ =~ s/\s+best actor\s+/ excellent /g;
$_ =~ s/\s+not forget\s+/ excellent /g;
$_ =~ s/\s+pleasently surprised\s+/ excellent /g;
```

```
$_ =~ s/\s+hooked me\s+/ recommended /g;
$_ =~ s/\s+incredibly brilliant\s+/ recommended /g;
$_ =~ s/\s+must see\s+/ recommended /g;
$_ =~ s/\s+oscar worthy\s+/ recommended /g;
$_ =~ s/\s+not smart\s+/ bad /g;
$_ =~ s/\s+poorly planned\s+/ bad /g;
$_ =~ s/\s+no suspense\s+/ very_bad /g;
$_ =~ s/\s+not able grab\s+/ very_bad /g;
$_ =~ s/\s+nothing perfect\s+/ very_bad /g;
$_ =~ s/\s+never exemplify\s+/ disgusting /g;
$_ =~ s/\s+no fizzing\s+/ disgusting /g;
$_ =~ s/\s+poor performance\s+/ disgusting /g;
$_ =~ s/\s+incredibly tiresome\s+/ never_recommended /g;
$_ =~ s/\s+most awful\s+/ never_recommended /g;
$_ =~ s/\s+nasty\s+/ never_recommended /g;
```

The rules were written considering all the individual opinion words, opinion phrase, positive patterns and negative patterns of the words. POS tagger has not been used for the purpose of capturing the opinion words or phrases. The entire process of writing the rules was based on how the human mind would understand while reading a review. Figure 3.1 shows a sample review, the corresponding feature vector using RT. Figure 3.1 also shows the opinion words, sentences captured by the PERL program and the appropriate RT assigned to them. A sample RTDM is shown in Table 3.2.

Table 3.1 List of stop words

Stop Words
a, about, an, and, are, as, at, be, been, by, for, from, how, in, is, of, on, or, that, than, the, they, these, this, to, was, what, when, where, who, will, with.

neg78_1.txt (A negative review from LDS 25000)

“I found this movie really hard to sit through, my attention kept wandering off the tv. As far as romantic movies go..this one is the worst I've seen. Don't bother with it.”

Step 1: Uppercase to lowercase conversion

i found this movie really hard to sit through, my attention kept wandering off the tv. as far as romantic movies go..this one is the worst i've seen. don't bother with it.

Step 2: Removal of stop words

i found movie really hard sit through my attention kept wandering off tv far romantic movies go one worst i ve seen don’t bother it

Step 3 : Opinion words/phrases captured by the PERL program and the appropriate RT assigned

1. really hard sit through → very bad

2. worst i ve seen → never recommended

3. don’t bother → disgusting

Step 4: Substitute the RT in the place of opinion words/phrases

i found movie **very bad** my attention kept wandering off tv far romantic movies go one **never recommended disgusting** it.

Step 5: Count the RT and create the feature vector

Feature vector using RT for the review

File name	Good	Very good	Excellent	Recommended	Bad	Very bad	Disgusting	Never recommended
neg78_1.txt	0	0	0	0	0	1	1	1

Figure 3.1 A sample review and the procedure for creating feature vector using RT

Table 3.2 A sample Representative Term-Document Matrix (RTDM)

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.	12	1	0	4	4	0	4	0
4.	9	7	4	0	13	5	1	2
5.	4	3	2	4	4	1	1	0
6.	7	2	1	2	6	1	2	0
7.	0	0	0	1	2	1	1	0

3.3 DATASET USED IN THIS THESIS

- The initial experiments were carried out on the dataset consisting of 538 reviews (122 negative reviews and 416 positive reviews) on camera and cell phone collected from www.epinions.com and www.cnet.com respectively.
- The benchmark movie review dataset provided by Andrew L Mass et al (2011) consisting of 25000 labeled movie reviews were used to evaluate the performance of various classifiers. This dataset will be referred as “large movie review dataset” (LDS) in this thesis. Also small subsets of the LDS have been used to evaluate the performance of classifiers. For example, LDS403 refers to 403 reviews from the LDS consisting of 25000 reviews.
- In the entire collection of movie reviews no more than 30 reviews are allowed for any given movie to avoid highly correlated ratings. The overall distribution of labels is balanced i.e. 12500 positive and 12500 negative reviews. Movie reviews of the entire categories viz. thriller, romance, science fiction, action and comedy etc. are included in the review corpus.

3.4 PERFORMANCE MEASURES

In this thesis, the following performance measures have been used to evaluate the performance of a classifier. All the performance measures discussed in this section are based on the following parameters:

- True Positive (TP) → The number of positive documents classified as positive by the classifier. (Correct classification)
- True Negative (TN) → The number of negative documents classified as negative by the classifier. (Correct classification)
- False Positive (FP) → The number of negative documents classified as positive by the classifier. (Incorrect classification)
- False Negative (FN) → The number of positive documents classified as negative by the classifier. (Incorrect classification)

Based on these parameters, the following performance measures are defined.

- Precision $(P) = TP/(TP+FP)$ (3.1)
- Recall $(R) = TP/(TP+FN)$ (3.2)
- Accuracy $(A) = TP+TN/(TP+TN+FP+FN)$ (3.3)
- F-Measure $= 2*P*R/(P+R)$ (3.4)

Note: Accuracy is also referred as Overall Success Rate (OSR)

For the performance measurement of all the classifiers discussed in this thesis, the performance measures discussed in section 3.4 have been used.

3.5 CLASSIFICATION PERFORMANCE OF VARIOUS CLASSIFIERS

RTDM for the LDS has been created as per the procedure discussed in section 3.2. The rules were manually coded after carefully reading 200 positive and 200 negative reviews from the dataset containing 25000 reviews.

Some of the popular classifiers like NB, BLR, MLP, SMO and CART available in WEKA 3.6.3 have been used to classify the reviews in the RTDM format. To evaluate the performance of these classifiers on various sizes of data set, subsets of LDS have been used. The performance of various classifiers is shown in Table 3.3, Table 3.4, Table 3.5 and Table 3.6. All the results are based on a ten-fold cross validation test.

Table 3.3 Performance of various classifiers for LDS403 (202 negative and 201 positive)

Classifier	P	R	F-Measure	Accuracy
NB	0.89	0.89	0.89	0.89
BLR	0.895	0.89	0.89	0.89
MLP	0.89	0.87	0.88	0.88
SMO	0.90	0.88	0.89	0.89
CART	0.84	0.80	0.82	0.82

Table 3.4 Performance of various classifiers for LDS2000 (1000 negative and 1000 positive)

Classifier	P	R	F-Measure	Accuracy
NB	0.74	0.88	0.81	0.79
BLR	0.84	0.83	0.83	0.83
MLP	0.84	0.82	0.83	0.83
SMO	0.83	0.84	0.83	0.83
CART	0.797	0.77	0.78	0.79

Table 3.5 Performance of various classifiers for LDS11000 (5500 negative and 5500 positive)

Classifier	P	R	F-Measure	Accuracy
NB	0.72	0.85	0.78	0.76
BLR	0.81	0.82	0.81	0.81
MLP	0.81	0.795	0.80	0.80
SMO	0.81	0.81	0.81	0.81
CART	0.78	0.78	0.78	0.78

Table 3.6 Performance of various classifiers for LDS25000 (12500 negative and 12500 positive)

Classifier	P	R	F-Measure	Accuracy
NB	0.71	0.85	0.77	0.75
BLR	0.797	0.797	0.797	0.797
MLP	0.78	0.81	0.796	0.79
SMO	0.795	0.80	0.797	0.797
CART	0.78	0.77	0.77	0.78

3.6 DISCUSSION ON THE CLASSIFICATION PERFORMANCE OF ML CLASSIFIERS WITH RTDM AS INPUT

- The NB classifier showed a consistently superior recall(R) compared to all the other classifiers.

- The accuracy of BLR, MLP and SMO has been consistent for all the sizes of dataset ranging from 403 to 25000. Accuracy level of near 90% for 403 reviews and a consistent 80% and above accuracy for all other sizes of dataset has been achieved by BLR, MLP and SMO. This consistently superior performance is certainly due to the RTDM format.

- Generally it is reported that, the performance of the NB classifier is the worst among the other classifiers and not dependable (Yang (1996) and Yang (1999)), but for the RTDM format, even the NB classifier performed with a comparable accuracy with respect to other classifiers.

3.7 ASSUMPTIONS AND SOME SPECIAL OBSERVATIONS

- The following assumptions were made with respect to the sentiment classification of review documents:

  a) Each review document focuses on a single object (target of opinion) expressed by a single person (opinion holder)

  b) The opinions expressed in a document are direct, explicit and genuine.

- RTDM creation requires creation of domain specific lexicon that requires some intelligence and prior knowledge. Though it is time consuming, the improved classification accuracy compensates it.

- While framing the rules for assigning the appropriate RT, the following special cases were also considered:

Rule	**Example**	**RT assigned**
Negative Negative (outcome is positive)	Not poor	Very good
Negative Positive (outcome is negative)	Not strong	Disgusting