"All the statistics in the world can't measure the warmth of a smile"  

(Chris Hart)

4.1 PROPOSED APPROACH

Background of proposed work

The work is based on using probability distribution of words for identifying keyword and 
stopword. The work is based on the assumption that

1. Keywords and stopwords follow Gaussian distribution
2. Each of them gives a separate distribution curve.

If these two assumptions hold then we are able to differentiate between keywords and 
stopwords.

As discussed in section 2.1 of chapter 2, there are different term weighting techniques to 
measure term representative ness. We propose to identify the keywords and stopwords 
using probability distribution function for some selected measures and see which measure 
is more suitable for distinguishing between keywords and stopwords.

Overall Approach

Proposed approach is a supervised learning method. The aim is to identify keywords and 
stopwords from a Hindi text corpus. An attempt is made to learn Gaussian distribution 
curve for keywords and stopwords separately during training. The testing is performed by
comparing the fitness of the word for both of the curves obtained during training and accordingly the higher fitness curve determines the result.

From the corpus some representative keywords and stopwords are identified manually. This provides data set for training and testing. 90 % of data comprised of training corpus where as rest 10 % was used for testing purposes. From the training data separate Gaussian distribution curves are drawn for keywords and stopwords using formula

$$p = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left\{ -\frac{(x - \mu)^2}{2\sigma^2} \right\}$$

Testing data again consists of keywords and stopwords. For each word in the testing corpus we computed its probability of being a keyword and stopword according to the following formula,

$$p_k = \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp \left\{ -\frac{(x - \mu_k)^2}{2\sigma_k^2} \right\}$$

$$p_s = \frac{1}{\sqrt{2\pi\sigma_s^2}} \exp \left\{ -\frac{(x - \mu_s)^2}{2\sigma_s^2} \right\}$$

where $p_k$ = Gaussian probability for a word to be a keyword
$\sigma_k$ = standard deviation of the keywords
$\mu_k$ = mean of keywords

$p_s$ = Gaussian probability for a word to be a stopword
$\sigma_s$ = standard deviation of the stopwords
$\mu_s$ = mean of stopwords
Based on the values of $P_k$ and $P_s$ (whichever is greater) we decide whether the word belongs to keyword or stopword category. The accuracy of result is determined by comparing the result with its actual label (keyword or stopword).

### 4.2 EXPERIMENTS AND RESULTS

**Problem statement:** To identify keyword and stopword from probability distribution function

**Steps**
1. Data collection: collect domain specific text.
2. Identify keywords and stopwords in training corpus.
3. Prepare data files for training and testing.
4. Make a Gaussian probability distribution function keyword and stopword using formula,

\[
p_k = \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp \left\{ -\frac{(x - \mu_k)^2}{2\sigma_k^2} \right\} \quad \text{and} \quad p_s = \frac{1}{\sqrt{2\pi\sigma_s^2}} \exp \left\{ -\frac{(x - \mu_s)^2}{2\sigma_s^2} \right\}
\]

5. For the test words in testing data set find whether they are keyword or stopword by checking if $P_k > P_s$ or $P_k < P_s$ respectively.
6. Test the accuracy of result by comparing the actual label of the word.
Implementation

Data collection

1. Cricket was chosen to be the domain for our purpose and relevant sports news were downloaded from various Hindi news websites and saved in text format.
2. These Hindi text files were transliterated into English. Transliteration was done on file by file basis by an application from Sanskrit centre section of JNU website. http://sanskrit.jnu.ac.in/ile/index.jsp
3. For the initial statistical purpose we used the limited period fully functional WordStat software for analyzing English text. It provided us with the following measures.
   - Frequency count of each token in the corpus.
   - Tf.idf measure of different tokens in the corpus.
   - Number of cases (no. of documents in the corpus in which a particular token occurs).
4. Besides these measures following measures were also calculated for each word.
   - Relative Frequency Count
   - Average Frequency Count
   - $X'$
   - IDF.
5. Keywords and stopwords were identified from the list to make two separate lists of keywords and stopwords so that training sets for keywords and stopwords can be generated.
6. Mean and Standard Deviation were computed for both of the training sets of keywords and stopwords for each of the term weighting measures viz. frequency, relative doc freq, avg doc freq, $X_{i}$, idf, tf.idf.
Assuming now that the distribution of keywords and stopwords in a corpus is Gaussian we have separate mean and variance values for each of the keywords and stopwords. Our objective is now to compute the resultant Gaussian probabilities for each measure in the keywords list as well as stopwords list e.g. taking frequency as the first measure in the keyword list we will compute the normal probabilities of all the keywords in the list. This procedure would be repeated for each of the measures in the keyword list. Similar actions would be taken for the stopwords list.

**Testing procedure**

There are following options to perform the testing on testing corpus

1. Randomly select a subset of words (approx 50) from the testing corpus (as the no of words is large) and perform testing on them and then finally check the accuracy manually.

2. A better option is to manually select a subset of words from the testing corpus which are known to be keywords and then perform testing on those words and find out the accuracy of different measures i.e. which measure classifies the most no of words as keywords.

3. To perform the testing on the whole corpus.

We would be employing the second method for our testing purposes since the last method is very calculation intensive as well as it would require checking for each of the about 1100 words in the testing corpus whether they were correctly classified or not. The measure which would give us the least no of stopwords during testing would be the better one.
4.2.1 Results and analysis

4.2.1.1 Preliminary analysis of training corpus

Total no of transliterated Hindi documents in training corpus = 88.
Total no of unique words in the training corpus = 3957
Total no of tokens in the training corpus = 38171.
Maximum Term frequency in the training corpus = 1468 for the word के along with the
document frequency of 88 (it occurred in all the documents) and a minimum term
frequency of 1 for a large no of words.
These low frequency words comprised the bulk of unique words in the corpus.
Words with term frequency up to 3 were removed.
No of words in the list after removal of low frequency words = 1330 words
No of keywords in the training corpus = 335
No of stopwords in the training corpus = 995
Ratio of keywords to all the tokens in the training corpus ≈ 335/4000 ≈ 1/12.
No of unique words in testing corpus = 1066

Tables 4.1 (a) and 4.1 (b) below shows the mean and standard deviation values computed
from the training corpus for the various measures for keywords and stopwords.

Table 4.1 (a) Training values for keywords

<table>
<thead>
<tr>
<th></th>
<th>Frequency</th>
<th>Tf.Idf</th>
<th>Idf</th>
<th>Average Frequency</th>
<th>Relative Frequency</th>
<th>X^2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average</td>
<td>25.59477</td>
<td>15.03987</td>
<td>1.058473</td>
<td>0.29085</td>
<td>1.949604</td>
<td>13.79085</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>49.87723</td>
<td>9.572614</td>
<td>0.391583</td>
<td>0.566787</td>
<td>1.086411</td>
<td>38.8245</td>
</tr>
</tbody>
</table>

45
Table 4.1(b) Training values for stopwords

<table>
<thead>
<tr>
<th></th>
<th>Frequency</th>
<th>Tf.Idf</th>
<th>Idf</th>
<th>Average Frequency</th>
<th>Relative Frequency</th>
<th>X^i</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average</td>
<td>25.30151</td>
<td>11.11116</td>
<td>1.092317</td>
<td>0.287517</td>
<td>1.604303</td>
<td>14.00603</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>92.53553</td>
<td>5.609148</td>
<td>0.381556</td>
<td>1.05154</td>
<td>1.119985</td>
<td>82.79247</td>
</tr>
</tbody>
</table>

Graphical representation of Gaussian frequency distribution of keywords and stopwords utilizing the values of mean and standard deviation from tables 4.1 (a) and 4.1 (b), is shown following figures. For further details please refer to the appendix.

Fig 4.1(a) Gaussian probability distribution of keywords in training data.
Fig 4.1(b) Gaussian probability distribution of stopwords in training data.

4.2.1.2 Results from testing corpus

Table of keywords selected tested from the testing corpus.

Table 4.2

| Team, Run, Rain, India, Cricket, Win, End, Cup, Game, Over, Player, Sri Lanka, Team, Dhoni, Pakistan, Tendulkar, Indian, Afridi, Crowd, West Indies, Ball, Sachin, Sahara, Ball, Player, Match, India, Matter, England, Bangladesh, Match, Step, Success, Approach, Ball, Player, Field, Ground, International, Country, Australia, Amavay |
### Table 4.3

<table>
<thead>
<tr>
<th>STATISTICAL MEASURE</th>
<th>NUMBER OF KEYWORDS</th>
<th>KEYWORDS CORRECTLY IDENTIFIED</th>
<th>PRECISION</th>
</tr>
</thead>
<tbody>
<tr>
<td>Term Frequency</td>
<td>50</td>
<td>41</td>
<td>82%</td>
</tr>
<tr>
<td>Inverse Document Frequency</td>
<td>50</td>
<td>23</td>
<td>46%</td>
</tr>
<tr>
<td>Tf.Idf</td>
<td>50</td>
<td>0</td>
<td>0%</td>
</tr>
<tr>
<td>Relative Frequency</td>
<td>50</td>
<td>34</td>
<td>68%</td>
</tr>
<tr>
<td>Average frequency</td>
<td>50</td>
<td>33</td>
<td>66%</td>
</tr>
<tr>
<td>$X^i$</td>
<td>50</td>
<td>43</td>
<td>86%</td>
</tr>
</tbody>
</table>

As shown in the table of results term frequency and $X^i$ measure are better discriminator for keywords and stopwords.

### Observation

Following observation were made during testing

1. Measures like frequency and $X^i$ which are whole integers and whose values are repeated frequently for many words i.e. two different terms can have the same frequency values, would be given the same probability and hence would be assigned to the same category even if they are from different categories.

2. Only those measures whose values are not likely to be repeated would be providing us with different probability values and hence different categories. Simple additive or subtractive measure would not be sufficient instead measures involving multiplication, division, logarithm etc would be giving better discriminatory results.
3. If we plot the frequency distribution of words then the important corpus representative keywords are found in the mid-section of the curve.

4. Tail end portion of the curve is not very informative for extracting keywords.

5. List size reduced to almost 1/3rd of the original list without any significant loss of information.

6. Zipf's Law is followed even in Hindi corpus suggesting that it's an inherent universal characteristic of languages.

7. In any corpus, the no of keywords are a fraction of total words in the corpus.