**National University of Computer & Emerging Sciences, Karachi Campus**

**Spring 2024**



**Project Report**

**Information Retrieval**

**Section: BCS-6K**

**Term Based Document Classification**

Instructor: Sir Basit

| **Group Members:** |
| --- |
| **Hajra Mutaal 21K-3310** |
| **Sareem Farooqui 21K-3381** |
| **Ali Hunain 21K-3397** |

**Project Objective**

Comparison of an acclaimed efficient and a robust classification method and four-well known classification techniques

**Methodology**

With the popularity of the Web, there has been an increase in the utility of text classification in applications like document filtering and sentiment categorization. Many of these applications demand that the classification method be efficient and robust, yet produce accurate categorizations by using the terms in the documents only. In the shared paper, the author proposed a novel and efficient method using terms-based discriminative information space for robust text classification. Terms in the documents are assigned weights according to the discrimination information they provide for one category over the others. These weights also serve to partition the terms into category sets. A linear opinion pool is adopted for combining the discrimination information provided by each set of terms to yield a feature space (discriminative information space) having dimensions equal to the number of classes. Subsequently, a discriminant function is learned to categorize the documents in the feature space.

**Highlights**

method A document vector is prepared on the basis of term occurrence and not term frequency to produce more accurate classifiers

Stopword removal and all preprocessing

To produce discrimination information for each term, we must first compute the likelihood of it appearing in a category k document which should be greater than its likelihood in appearing in other category documents.

The RR and OR will have their minimum value 1 and the log functions of both should have minimum of 0

To avoid probability of a term in a document being zero, add-one laplacing is done to avoid division-by-zero error during weight computation.

Although weights of every English word is computed, the terms are partitioned into relevant and irrelevant terms, later are removed from classification model. The terms are found relevant if the discrimination value is higher than the value of the threshold.

On the basis of set of terms whose occurrence chances in category k is greater than that in category not k and the set of terms whose occurrence chances are greater in category not k than occurrence chances in category k, we computed two scores of each document relative to each category, score with category k and score with category not k. Using these scores, we built term partition space.

Once we have a graph plot of the feature space, we can easily compute the slope and bias parameter. We also assign a document’s category using the linear discriminant function.

| Data | Training | DIST-RR % | DIST-LRR % | DIST-KLD % | DIST-OR % | DIST-LOR % |
| --- | --- | --- | --- | --- | --- | --- |
| 20NewsGroup | 14000  (70%) | 60.714 | 62.5 | 64.286 | 67.0714 | 42.857 |
|
| Movie Review | 1000(each) | 79.3 | 82.14 | 78.214 | 80.36 | 79.6 |
| SRAA | 50000 | 89.3 | 81.39 | 82.14 | 83.57 | 87.14 |

76.43 % (Overall RR)

75.34 % (Overall LRR)

74.88 % (Overall KDL)

76.223 % (Overall OR)

69.8 % (Overall LOR)

|  |  | NB | ME | BW |  |
| --- | --- | --- | --- | --- | --- |
| SRAA | 1500 | 71.4 | 96.3 | 85.2 |  |
|  | 1000 | 63.3 | 92.5 | 82.95 |  |
|  | 500 | 59.7 | 90.6 | 80.79 |  |
|  | 250 | 56.9 | 82.6 | 75.4 |  |
|  | 150 | 56.7 | 64.9 | 73.3 |  |
|  | Avg | 61.6 | 85.38 | 79.528 |  |

|  |  | NB | ME | BW |  |
| --- | --- | --- | --- | --- | --- |
| Movie Review | 600 | 77.6 | 78 | 74.56 |  |
|  | 500 | 78.46 | 78.06 | 73.29 |  |
|  | 400 | 78.6 | 77.1 | 65.78 |  |
|  | 300 | 61.7 | 75.2 | 62.13 |  |
|  | 200 | 51.7 | 71.5 | 60.15 |  |
|  | Avg | 69.612 | 75.972 | 67.182 |  |

|  |  | NB | ME | BW |  |
| --- | --- | --- | --- | --- | --- |
| 20NewsGroup | 20,000 | 72.3 | 72.7 | 60.32 |  |

**Conclusion**

Relative Risk performs relatively better among all term weighing techniques. Although ME and BW performed better than DIST with SRAA dataset, overall DIST performance is significantly better in classification with higher accuracy percentage. The log involving techniques performed poorly in comparison with their non-log counterparts but slightly better than KDL.