**Hybrid Dimensionality Reduction Scheme for Text Classification**

The purpose of this research paper is to propose a two-stage dimensionality reduction (DR) technique for effectively classifying textual content. The first step involves performing feature selection on the original feature space to obtain a reduced feature set. In this step, the features are ranked using the chi square method based on their scores in the original space, and the top-ranking features are selected for the next step.

In the second step, feature extraction is conducted on the reduced feature set obtained from the previous step. Principal Component Analysis (PCA) is utilized to generate a new feature space with significantly reduced dimensions. The performance of various learning algorithms is then evaluated on this new feature space. The experimental results demonstrate a remarkable enhancement in the classification performance, specifically in terms of the f-score, when using our proposed technique on IMDB dataset.

**Algorithm**

An algorithm for text classification that involves feature selection using chi-square in the first step and feature extraction using PCA in the second step:

Input:

- Training dataset (text documents with corresponding class labels)

- Number of desired features after feature selection (k)

Output:

- Predicted class labels for the test dataset

Algorithm:

Step 1: Feature Selection using Chi-Square

1.1: Preprocess the training dataset (tokenization, stemming, etc.) to obtain a collection of text documents.

1.2: Calculate the chi-square value for each term (feature) in the training dataset, comparing its frequency in each class.

1.3: Rank the features based on their chi-square scores in descending order.

1.4: Select the top-k (k=50%) features with the highest chi-square scores.

Step 2: Feature Extraction using PCA

2.1: Preprocess the training dataset again, including the selected top-k features only.

2.2: Construct a term-document matrix, where each entry represents the frequency of a feature in a document.

2.3: Perform PCA on the term-document matrix to reduce its dimensionality.

2.4: Retain the principal components that capture a significant amount of variance ( 95%).

2.5: Transform the training dataset using the retained principal components to obtain the reduced feature space.

Step 3: Training and Classification

3.1: Train a text classification model (Naive Bayes, Support Vector Machine, Logistic Regression, Artificial Neural Network etc.) using the reduced feature space and the corresponding class labels.

3.2: Apply 10 fold cross validation technique on the training dataset.

3.4: Predict the class labels for the dataset using the trained classification model.

3.5: Output the predicted class labels for the dataset.

End of Algorithm

**Result**

Without applying the proposed algorithm

|  |  |  |
| --- | --- | --- |
| **Classifier** | **Accuracy** | **f-score** |
| **SVM** | 82.34 | 82.45 |
| **NB** | 83.76 | 84.94 |
| **LR** | 79.25 | 81.43 |
| **ANN** | 88.83 | 89.48 |

After applying the proposed algorithm

|  |  |  |
| --- | --- | --- |
| **Classifier** | **Accuracy** | **f-score** |
| **SVM** | 93 | 92.20 |
| **NB** | 91.8 | 88.94 |
| **LR** | 84.99 | 85.68 |
| **ANN** | 90.88 | 86.48 |