An Improved Taxonomy Re-structuring Using a Modified K-means Clustering for Efficient Large-scale Text Classification

Abubakar Ado[1], Abdurra'uf Garba Sharifai[2], Mansir Abubakar[3], Bashir Salisu[4] , Usman Mahmud[5] , Abdulkadir Abubakar Bichi[6]

[1,2,4,5,6]Department of Computer Science, Northwest University Kano

[4]Department of Computer Science, Al-Qalam University, Katsina

MDPI ASEC-2025

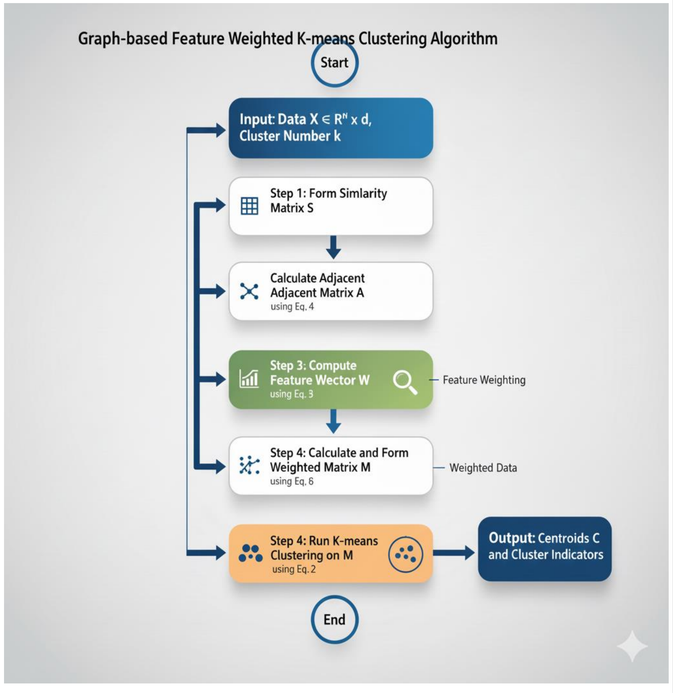
Introduction

Textual classification to a hierarchical taxonomy of classes is a common and known problem associated with Large-Scale Text classification (LSTC). When there are many labels, hierarchical restructuring of classes has been recognized as a natural and effective way to organize similar classes and it has been well studied in the past two decades.

Problem Statement

When there are many classes with increase number of features, the exiting hierarchy re-structuring methods tend to produce many nodes with similar granularities. This results in mis-classification, computationally expensive, not scalable for many classification models.

mKmeans Algorithm



Proposed Method

Algorithm: Res-mKMeans Algorithm

Input: Original hierarchy H, dataset X, class labels y,
number of clusters K for restructuring

Output: Modified hierarchy *H_new*

```
H_new ← H
for each parent node p in H_new do
    C_p ← children classes of p
    X_p ← samples belonging to classes in C_p
    if |C_p| > K then
        (cluster_ids, centroids) ← mKMeans(X_p, K)
        NewGroups ← {G_1, G_2, ..., G_K}
        for each class c ∈ C_p do
            assign class c to G_{cluster_ids(c)}
        end for
        replace children(p) with NewGroups
    end if end for
for each parent p in H_new do
    children_list ← children(p)
    for each child g in children_list do
        if size(g) == 1 then continue
        end if
        if purity(g) < threshold then
            (cid, cent) ← mKMeans(samples(g), K_refine)
            ReGroups ← {Rg_1, ..., Rg_K_refine}
            for each class c in g do
                assign c to Rg_{cid(c)} end for
            replace g in children_list with ReGroups
        end if end for end for
for each node n in H_new do
    if number_of_children(n) < min_children then
        H_new ← NDD(n, H_new)
    else if number_of_children(n) > max_children then
        (cid, cent) ← mKMeans(samples(n), K_split)
        SplitGroups ← assign classes using cid
        replace children(n) with SplitGroups
    end if end for
return H_new
```

Results Analysis

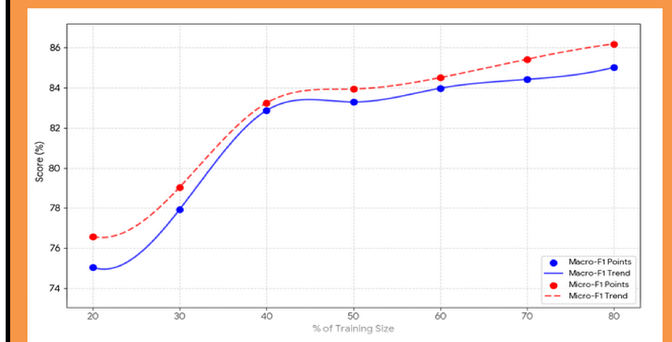


Fig.2: Macro-f1 and Micro-f1 Performance of the proposed method with varyimg % of training size

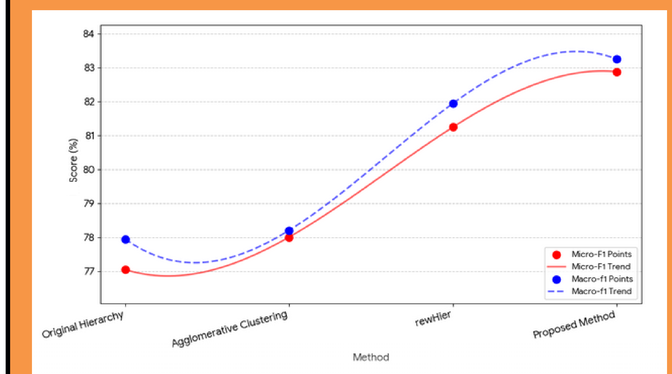


Fig 2: Micro-f1 and Macro-f1 performance comparison using different hierarchy restructuring approaches on newsgroup dataset

Conclusion

We propose an approach for restructuring hierarchy that is more suited for HC. In comparison to existing approaches, our method gives better performance that allow HC approaches to significantly scale to LSHC

Acknowledgments

We would like to thank the TetFund, Nigeria and Northwest University, kano for supporting this research under Institution based research, IBR-2024.