Approximating Manifold Projected Hierarchical Clustering with Neural Nets

Armen Manukyan

Yerevan State University

armen.manukyan9@edu.ysu.am

May 23, 2024

Overview

Neural Network Based Clustering Methodology

Hypothesis:

A neural network can effectively approximate the outcomes of sequential dimensionality reduction and clustering algorithms.

Methodology:

We choose to cluster textual sentences using:

- **MPNet**[6]
- **PCA**[3] $->$ **UMAP**[5]
- **HDBSCAN**[4]

In the mentioned order which we call ***Initial Stack***, and use the obtained clusters as labels to train an approximation Network, with an objective to improve inference speed while maintaining the precision of the ***Initial Stack***.

Workflow Details

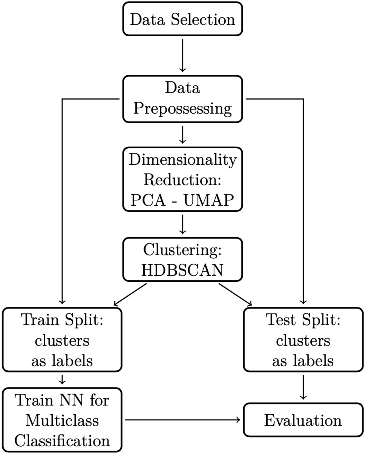


Figure: The workflow preprocesses selectet dataset(embeds sentences in our case), applies PCA and UMAP for dimensionality reduction, clusters with HDBSCAN, and trains a Neural Network on pairs : ***[Initial embedding, obtained cluster]***.

Problems with *Initial Stack*

1. The need for refitting the whole stack

- Each new datapoint requires the entire sequence of embedding, dimensionality reduction, and clustering to be assigned a cluster.
- This is computationally expensive and inefficient, especially for large datasets or real time applications.

2. Additional bottleneck of the HDBSCAN and UMAP implementations working on CPU

- HDBSCAN and UMAP are primarily CPU-based implementations.
- This prevents leveraging GPU acceleration, making the refitting process even slower.
- Significant delays in clustering new data points, which is problematic for dynamic or streaming data scenarios.

How We Address the Mentioned Problems

Solution Overview

We develop a process that transitions the unsupervised clustering task into a supervised classification problem by using the obtained clusters as labels to train a simple feed-forward Neural Network, which aims to approximate the ***Initial Stack***.

Key Benefits

- **Fast Inference:** Our Neural Network allows for rapid inference for new datapoints.
- **How:** Fit the ***Initial Stack*** once and keep the network trained on its results for inference on GPU.

Data Selection and Objectives

For our empirical evaluation, we selected three diverse textual datasets: AG News[2], TweetEval [1], and Yelp Reviews [7].

- **AG News Dataset**:
  - Source: 120,000 news articles from over 2000 sources.
  - Objective: Train our approximation network. Compare the identified latent clusters, leading to top performing NN, with the 4 primary categories of News - *World, Sport, Business, Sci/Tech.*
- **TweetEval Dataset**:
  - Source: Selected subset of 100,000 tweets categorized with 20 distinct emojis.
  - Focus: Train our approximation network. Compare the identified inherent clusters, leading to top performing NN, with the 20 distinct emojies.
- **Yelp Reviews Dataset**:
  - Source: Focused on 130,000 texts with 5-star reviews form the offered 700,000 reviews from the Yelp Dataset Challenge 2015 having 1 to 5 stars.
  - Objective: Train our approximation network. Understand which type of topics(Mexican restaurant, barbershop etc.) lead to top performing NN.

Ground truth labels were not used in conjunction with text embeddings.

HDBSCAN Epsilon Value Experiments

Objective: Explore how varying the $cluster - selection - epsilon$ parameter of HDBSCAN influences neural network performance.

Notes

- We use macro f1 score to describe the performance of our neural network
- We use the outliers generated by the ***Initial Stack*** as a separate class while training our network

Results: AG News Dataset

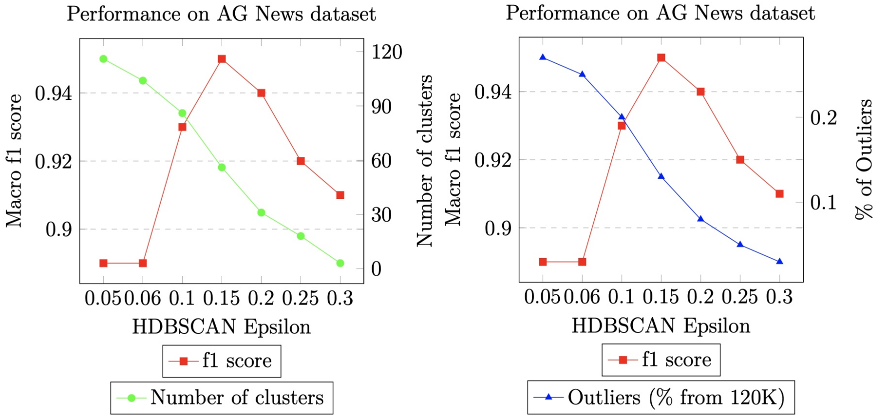


Figure 4.1: Comparison of NN's F1 score with Number of Clusters and Percentage of Outliers based on HDBSCAN Epsilon for AG News Dataset with 120K datapoints

Results: Yelp 5-star Reviews

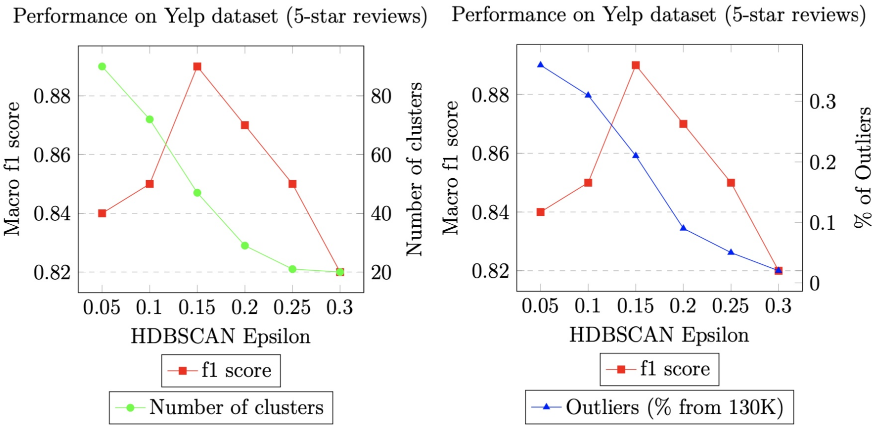
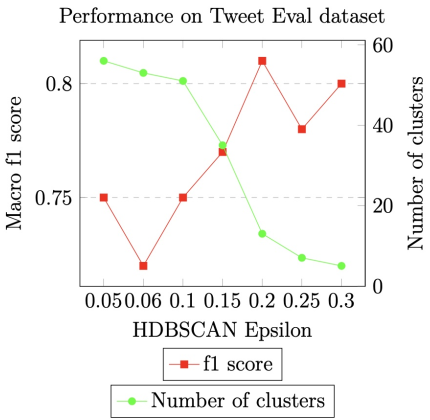


Figure 4.3: Comparison of NN's F1 score with Number of Clusters and Percentage of Outliers based on HDBSCAN Epsilon for Yelp Dataset (5-star reviews) with 130K datapoints

Results: TweetEval Dataset



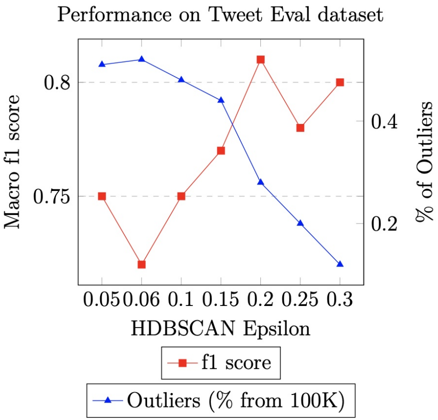


Figure 4.2: Comparison of NN's F1 score with Number of Clusters and Percentage of Outliers based on HDBSCAN Epsilon forr Tweet Eval Emoji Dataset with 100K datapoints

Results Summary: HDBSCAN

HDBSCAN's Optimal Parameter Leading to Peak NN Performance

- **Optimal Range:** Clusters obtained with $cluster - selection - epsilon$ values between 0.15 and 0.2 yield the best neural network performance across all datasets.

Peak NN Performance Correlates with Natural Clusters

- **AG News Dataset:** Optimal performance corresponds to more granular topics, such as Hockey, Basketball, NCAA, and Motorsports, as opposed to the broader 'sports' category provided by the dataset.
- **Yelp Reviews Dataset:** Top NN performance is achieved with natural clusters which align with highly rated locations and specific types of establishments, like hotels, steak houses in Vegas, and sushi bars.
- **TweetEval Dataset:** Cluster counts leading to peak NN performance closely match the 20 emoji labels, with topics including birthday, food, and hair.

UMAP Dimensions Experiments

Objective: Investigate the impact of varying UMAP dimensionality (16, 32, 48, 64, 80) on clustering outcomes and neural network performance.

Key Findings

- **AG News Dataset:**
  - Varying dimensions showed maximum of 1% change in network performance across all experiments.
- **Yelp Reviews Dataset:**
  - Varying dimensions showed maximum of 3% change in network performance across all experiments.
- **TweetEval Dataset:**
  - Higher dimensions - 64 and 80 resulted in slight performance drop (macro f1 score drop of around 5% in both cases) possibly due to the noisy structure of the tweets.

Results Summary: UMAP

UMAP's Optimal Dimension Before Clustering Leading to Peak NN Performance

- **Optimal dimension before clustering:** Network generalizes well across varied final dimensions before clustering, thus some applications could benefit from using the lowest end of final dimensions to aid the speed of initial clustering.

Experiments with Network Depth

Objective: Evaluate the impact of adding layers to the neural network on its performance.

Experimental Setup

- Original network: 5 hidden layers.
- New architecture: Adding 3 intermediate layers.
- Consistent training parameters to isolate depth impact.

Results

- **Yelp Reviews:** Macro F1 decreased from 0.89 to 0.87.
- **AG News:** Performance stable at Macro F1 $= 0.95$.
- **TweetEval:** Macro F1 dropped from 0.81 to 0.75.

Results Summary: Network Depth

Network Depth Leading to Peak Performance

- Our goal was to test if performance improvements could be achieved by solely increasing the network depth while keeping other factors constant, without intentionally increasing task difficulty.
- Increased depth did not consistently enhance performance; results were influenced by dataset complexity and the use of outlier clusters as a class during training.

Conclusion and Key Insights

Summary:

We demonstrated the viability of a neural network based framework for approximating sequential dimensionality reduction and clustering outcomes within textual data, achieving rapid inference without sacrificing precision.

Key Insights:

- **Optimal HDBSCAN Parameters:** The $cluster - selection - epsilon$ parameter consistently performed best between **0.15 and 0.2** across all datasets.
- **UMAP Dimensionality:** Varying final dimensionality with five different values in the range **16 and 80** had minimal impact, advocating for lower dimensionality to expedite fitting.
- **Network Depth:** Increasing neural network depth did not improve performance. Instead, global performance is influenced by the presence of outlier class when training the network.

Future Work:

Expanding this framework to other data domains, such as voice or images, and comparing the optimal settings with those identified in this study.

References I

[1] Francesco Barbieri et al. "TweetEval: Unified Benchmark and Comparative Evaluation for Tweet Classification". In: *Proceedings of Findings of EMNLP*. 2020, pp. 1644–1650. URL: https://github.com/cardiffnlp/tweeteval.

[2] ComeToMyHead. *AG News Dataset*. 2015. URL: https://github.com/zhangxiangxiao/Crepe.

[3] I.T. Jolliffe. *Principal Component Analysis*. 2nd. Springer, 2016.

[4] Claudia Malzer and M. Baum. "A Hybrid Approach To Hierarchical Density-based Cluster Selection". In: *arXiv preprint arXiv:1911.02282* (2019). Available at arXiv:1911.02282.

[5] Leland McInnes and John Healy. "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction". In: *Journal of Open Source Software* 3.29 (2018), p. 861.

[6] Kaitao Song et al. *MPNet: Masked and Permuted Pre-training for Language Understanding*. Available at arXiv:2004.09297. 2020.

References II

[7] Xiang Zhang, Junbo Zhao, and Yann LeCun. "Character-level Convolutional Networks for Text Classification". In: *Advances in Neural Information Processing Systems 28 (NIPS 2015)*. 2015. URL: https://www.yelp.com/dataset.

Thank You

Now I will happily answer your questions