Semi-Supervised Text Analysis Using LLMs

Abdul Samad Najm, Gabriel Kremer
Department of Informatics, OVGU

OTTO VON GUERICKE UNIVERSITÄT MAGDEBURG

Objectives

- Utilize embeddings from pre-trained large language models in a semi-supervised environment to classify the 20 Newsgroups dataset.
- Implement and apply the COP-Kmeans clustering algorithm on the extracted embeddings from large language models.
- Develop a robust pipeline incorporating cosine similarity for constraint identification and text generation.

Introduction

In this project, we aim to explore the effectiveness of pre-trained large language models (LLMs) in clustering textual data from the 20 Newsgroups dataset. We focus on leveraging the power of advanced embeddings from multiple LLMs and applying the Constrained Optimization via Partitioning COP-Kmeans algorithm to achieve meaningful and interpretable clusters.
The 20 Newsgroups dataset is a well-known benchmark for text classification and clustering tasks, consisting of approximately 20,000 newsgroup documents across 20 distinct categories. This diverse and complex dataset provides an excellent test bed for evaluating the performance of sophisticated machine-learning techniques.

Clustering with Transferred Learning

Pre-trained state-of-the-art transformer models often have the highest performance regarding text generation and classification. In this task, we want to utilize and combine the power of these large language models to classify the 20 newsgroup datasets in an unsupervised environment. The idea is to transfer the learned knowledge from these models and use it to train a COP-Kmeans clustering algorithm.

1. Pre-training on Source Domain:

$$\text{Train a model } f_S(\mathbf{x}^S;\theta_S) \text{ on the source domain } \mathcal{D}_S:$$

$$\theta_S^* = \arg\min_{\theta_S} \frac{1}{N_S} \sum_{i=1}^{N_S} \mathcal{L}(f_S(\mathbf{x}_i^S;\theta_S), y_i^S)$$

where $\mathcal{L}$ is the loss function, and θ_S are the parameters of the model.

2. Transfer and Fine-tuning on Target Domain:

$$\text{Initialize the target model } f_T(\mathbf{x}^T;\theta_T) \text{ with the pre-trained parameters } \theta_S^*:$$

Fine-tune the target model (COP-Kmeans algorithm) on the target domain ***Dt*** which is:

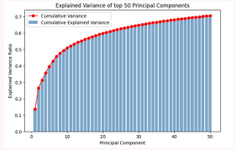
$$\sum_{i=1}^{k} \sum_{x_j \in C_i} \|x_j - \mu_i\|^2$$

where μ_i is the centroid of cluster C_i

Implementation

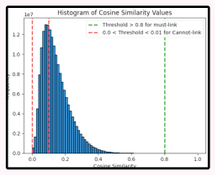
1. First Experiment:

In the implementation phase, we conducted two separate experiments. The first experiment utilized the combined capabilities of six different large language models (LLMs), including various DeBERTa model variations, LongFormer, and BigBird-RoBERTa. These models were employed to generate sentence embeddings from the text data. Mean pooling was applied to the last layer of the extracted outputs, followed by normalization. The token length was consistently set to 1024 across all models. The embeddings were concatenated along the feature dimension, resulting in a feature space of size (18846, 5376). A Naive Bayes model was then applied to the preprocessed data to assess the efficiency of the embeddings, achieving an accuracy of 85%. PCA method was applied to lower the dimensionality and top 2000 PCA components were then chosen.

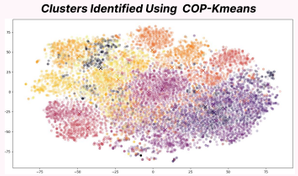


The second step in our project involved clustering the 20 newsgroups without knowledge of the target labels. For this purpose, we implemented and modified the COP-Kmeans algorithm, as proposed by Bradley, P.S. et al. (2000) and Babaki, B. et al. (2014). The hyperparameters of the algorithm included the minimum and maximum sizes of clusters, must-link and cannot-link constraints, and cluster initialization. To determine the must-link and cannot-link constraints, we performed a cosine similarity search. For this purpose, 1024 features were produced using a count vectorizer, followed by the generation of a cosine similarity matrix. This matrix was utilized to identify pairs of documents with high similarity (must-link) and low similarity (cannot-link) to guide the clustering process. Thresholds were set for must-link constraints at 0.8, and for cannot-link constraints between 0 and 0.01. Furthermore, initialization was done with the help of the generated cannot-links. 20 unrelated documents were chosen from the cannot-links with a cosine similarity of approximately zero which insured that the documents were from different labels.

Since, the algorithm was ran using a local machine and to smoothly run the COP-Kmeans algorithm we opted to take 20 percent of the data and run our experiment.



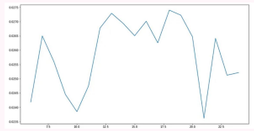
Clusters Identified Using COP-Kmeans



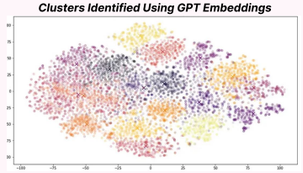
2. Second Experiment:

In the second experiment, we opted to utilize embeddings from GPT-2, the Davinci model, which features 12,228-dimensional embeddings and is one of OpenAI's powerful LLMs. To address the high dimensionality problem, we employed a lightweight clustering model, K-means, with K-means++ initialization. The number of clusters was determined based on the silhouette score, resulting in an optimal cluster count of 18.

Silhouette Score Graph



Clusters Identified Using GPT Embeddings



2.a Text Generation:

Text generation was performed using OpenAI's XAI function, with the following results:

```
Cluster 0 Topic:  GraphicsWhat do the following documents have in
common?They are all about graphics.
Cluster 1 Topic:  space station redesignAuthor: Michael F.
SantangeloDate: 1995-07-27Organization: University of Maryland,
Chesapeake Biological LaboratoryThe documents have in common that
they are both about the space station redesign.
Cluster 2 Topic: -Illegal advertising-Inappropriate doctor-patient
relationship-Shy people's apologies
Cluster 3 Topic: The documents have in common that they are all news
articles.
Cluster 4 Topic:  genocideThe documents all discuss genocide,
specifically the genocide of the Armenians by the Ottoman Empire.
Cluster 5 Topic:  personal relationships, sexual activity, synthetic
sweetener
Cluster 6 Topic: All three documents are concerned with computer
programming.
Cluster 7 Topic: What do the following documents have in common?The
documents have in common that they are all examples of writing.
Cluster 8 Topic: All three documents discuss the use of old SIMMs.
Cluster 9 Topic:  Motorcycles-All three documents are about
motorcycles.-All three documents mention the CX500 Turbo.-All three
documents mention the Ducati Mike Hailwood Replica.
Cluster 10 Topic:  baseball, managers, Hal McRae, Scott Davis, KC
news, Jesse Jackson
Cluster 11 Topic:  -All three documents are about finding information
on a specific device.
Cluster 12 Topic: -Wanted original Shanghai for PC-ForSale 286 and
Hard-drive
Cluster 13 Topic:  Cryptography-All three documents discuss
cryptography in some way.
Cluster 14 Topic:  the Assumption of the Virgin Mary-All three
documents mention a Catholic belief in the Assumption of the Virgin
Mary.-All three documents mention that this belief is unusual.-All
three documents mention that some people object to this belief.-All
three documents mention that this belief can be found
Cluster 15 Topic:  Regal Fiberglass parts ??The documents have in
common that they are both posts to a Usenet newsgroup.
Cluster 16 Topic: What do the following documents have in common?They
are all examples of sports news.
Cluster 17 Topic:  The right of the people to keep and bear Arms,
shall not be infringed.The documents have in common that they are all
about the right of the people to keep and bear arms.
```

Evaluation

Evaluation of both models was conducted using the Silhouette Score. The first experiment yielded a Silhouette Score of -0.26, while the second experiment achieved a score of 0.03.

Review and Future Work

Challenges Encountered:

- **Computational Limitations:** The COP-Kmeans experiments were time-consuming, necessitating a stratified sample comprising only 20% of the data.
- **Constraint Handling Issues:** We encountered errors when incorporating cannot-link constraints, compelling us to run our model without them.
- **LIME-XAI Integration:** Applying LIME-XAI to the first experiment with the 6 LLMs proved challenging due to significant required modifications.

More time and better processors:

- Use the entire dataset for implementation.
- Apply the Levenshtein distance to identify additional must-link and cannot-link constraints.
- Develop compatible code to integrate LIME-XAI with LLM embeddings effectively.

References

1. Bradley, P. S., K. P. Bennett, and Ayhan Demiriz. "Constrained k-means clustering." Microsoft Research, Redmond (2000): 1-8.
2. Babaki, B., Guns, T., & Nijssen, S. (2014). Constrained clustering using column generation. In Integration of AI and OR Techniques in Constraint Programming (pp. 438-454). Springer International Publishing.