Text Clustering as Classification with LLMs

Natural Language Processing

Salveen Singh Dutt
Patryk Prusak

January 29, 2025

Warsaw University of Technology

Paper

- Title: Text Clustering as Classification with LLMs
- Authors: Chen Huang, Guoxiu He
- Affiliations: StatNLP Research Group (SUTD), East China Normal University
- Year: 2024
- Contact: chen_huang@mymail.sutd.edu.sg, gxhe@fem.ecnu.edu.cn

Introduction

- Background:
    - Text clustering organizes and identifies patterns in unlabeled data.

Introduction

- Background:
    - Text clustering organizes and identifies patterns in unlabeled data.
    - Traditional methods rely on fine-tuning embedders and sophisticated metrics.

Introduction

- Background:
  - Text clustering organizes and identifies patterns in unlabeled data.
  - Traditional methods rely on fine-tuning embedders and sophisticated metrics.
- Objective:
  - Transform text clustering into a classification task using LLMs.

Introduction

- Background:
  - Text clustering organizes and identifies patterns in unlabeled data.
  - Traditional methods rely on fine-tuning embedders and sophisticated metrics.
- Objective:
  - Transform text clustering into a classification task using LLMs.
- Key Contributions:
  - No fine-tuning or hyperparameter tuning required.
  - State-of-the-art performance on multiple datasets.

Methodology Overview

- Framework: Two-Stage Process
    1. Stage 1: Label Generation
        - Generate labels in mini-batches.
        - Merge similar labels for granularity.

Methodology Overview

- Framework: Two-Stage Process
    1. Stage 1: Label Generation
        - Generate labels in mini-batches.
        - Merge similar labels for granularity.
    2. Stage 2: Label Classification
        - Classify data samples based on the generated labels.

Methodology Overview

- Framework: Two-Stage Process
  1. Stage 1: Label Generation
     - Generate labels in mini-batches.
     - Merge similar labels for granularity.
  2. Stage 2: Label Classification
     - Classify data samples based on the generated labels.
- Advantages:
  - Utilizes LLM's in-context learning ability.
  - Bypasses input length and clustering algorithm complexity.

Task Definition

- **Input:** Unlabeled dataset $D = \{d_i\}_{i=1}^{N}$.
- **Goal:** Partition data into $C = \{c_j\}_{j=1}^{K}$ clusters.
- **Transformation:**
  - Generate potential labels $L = \{l_k\}_{k=1}^{K'}$.
  - Classify each $d_i \in D$ into one label $l \in L$.

Methodology

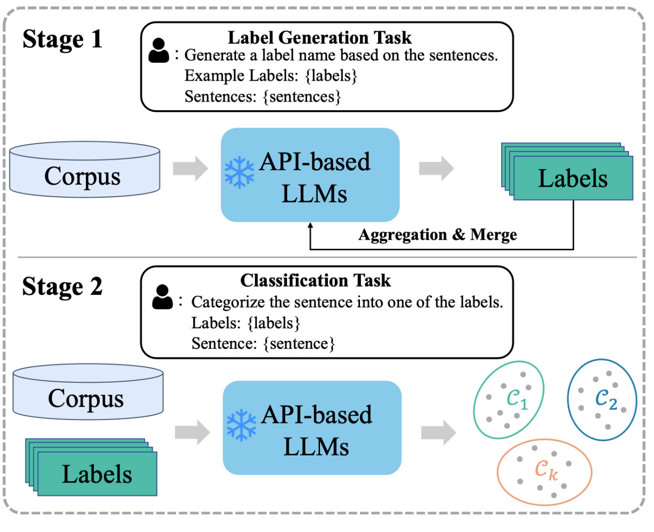

Figure 1: Clustering Methodology from [Huang and He(2024)]

Label Generation

- Process:
    1. Divide dataset into mini-batches.

Label Generation

- Process:
  1. Divide dataset into mini-batches.
  2. Prompt LLM to generate labels for each mini-batch.

Label Generation

- Process:
  1. Divide dataset into mini-batches.
  2. Prompt LLM to generate labels for each mini-batch.
  3. Aggregate and merge similar labels to reduce redundancy.

Label Generation

- Process:
    1. Divide dataset into mini-batches.
    2. Prompt LLM to generate labels for each mini-batch.
    3. Aggregate and merge similar labels to reduce redundancy.
- Prompt Examples:
    - Generate labels: ”Given sentences: {sentences}. Suggest labels.”

Label Generation

- **Process:**
  1. Divide dataset into mini-batches.
  2. Prompt LLM to generate labels for each mini-batch.
  3. Aggregate and merge similar labels to reduce redundancy.
- **Prompt Examples:**
  - Generate labels: ”Given sentences: {sentences}. Suggest labels.”
  - Merge labels: ”Analyze and merge synonymous labels: {label_list}.”

Experiment Setup

- Datasets:
  - Tasks: Topic mining, emotion detection, intent discovery, domain discovery.
  - Examples: ArxivS2S, GoEmo, Massive-I/D, MTOP-I.

Experiment Setup

- Datasets:
  - Tasks: Topic mining, emotion detection, intent discovery, domain discovery.
  - Examples: ArxivS2S, GoEmo, Massive-I/D, MTOP-I.
- Evaluation Metrics:
  - Accuracy (ACC), Normalized Mutual Information (NMI), Adjusted Rand Index (ARI).

Experiment Setup

- Datasets:
    - Tasks: Topic mining, emotion detection, intent discovery, domain discovery.
    - Examples: ArxivS2S, GoEmo, Massive-I/D, MTOP-I.
- Evaluation Metrics:
    - Accuracy (ACC), Normalized Mutual Information (NMI), Adjusted Rand Index (ARI).
- Baseline Methods:
    - K-means, IDAS, PAS, Keyphrase Clustering, ClusterLLM.

Evaluation Metrics

- Accuracy (ACC)
  - Measures the percentage of correctly assigned cluster labels
  - Formula: $ACC = \frac{1}{N} \sum_{i=1}^{N} \delta(y_i, map(c_i))$
  - Where $\delta(x, y)$ is 1 if x=y and 0 otherwise
  - $map(c_i)$ finds the best mapping between clusters and true labels

Evaluation Metrics

- Rand Index (RI)
  - Measures similarity between two data clusterings
  - Defined as the fraction of correctly grouped or separated pairs.

Cluster labels by an algorithm	Ground truth cluster labels
0	2
0	2
1	3
1	3
2	1

Figure 2: Clustering Example

Evaluation Metrics

- Rand Index (RI)
  - Measures similarity between two data clusterings
  - Defined as the fraction of correctly grouped or separated pairs
  - Formula:

$$RI = \frac{TP + TN}{\binom{n}{2}}$$

    where:
    - TP = Number of pairs in the same cluster in both partitions
    - TN = Number of pairs in different clusters in both partitions
    - $\binom{n}{2}$ = Total number of pairs
  - Ranges from 0 to 1:
    - 1: Perfect agreement between clusterings
    - 0: No agreement beyond random chance

Evaluation Metrics

- Adjusted Rand Index (ARI)
  - Measures similarity between two data clusterings. [Hubert and Arabie(1985)]
  - Adjusts for chance - accounts for random label assignments
  - Formula: $ARI = \frac{RI - E[RI]}{max(RI) - E[RI]}$
  - Where RI is the raw Rand index
  - Ranges from -1 to 1:
    - 1: Perfect match between clusterings
    - 0: Random labeling
    - Negative: Worse than random

Evaluation Metrics

- Normalized Mutual Information (NMI)
  - Measures the mutual dependence between true labels and predicted clusters. [Vinh et al.(2010)Vinh, Epps, and Bailey]
  - Formula: $NMI(Y, C) = \frac{2 \times I(Y;C)}{H(Y)+H(C)}$
  - $I(Y; C)$ is mutual information, $H(Y)$ and $H(C)$ are entropies
  - Ranges from 0 (no mutual information) to 1 (perfect correlation)

Results

- Performance Highlights:
  - Outperforms baselines in ACC, NMI, and ARI across all datasets.
  - Example: ArxivS2S ACC: 38.78% (Ours) vs. 26.34% (ClusterLLM).

Results

- Performance Highlights:
  - Outperforms baselines in ACC, NMI, and ARI across all datasets.
  - Example: ArxivS2S ACC: 38.78% (Ours) vs. 26.34% (ClusterLLM).
- Granularity:
  - Closer alignment to true cluster counts after label merging.
  - Example: MTOP-I clusters: 83 (Ours) vs. 43 (ClusterLLM).

Advantages and Limitations

- Advantages:
  - Simplifies clustering into a classification task.
  - Improves interpretability with meaningful labels.
  - Eliminates fine-tuning and hyperparameter tuning.

Advantages and Limitations

- Advantages:
    - Simplifies clustering into a classification task.
    - Improves interpretability with meaningful labels.
    - Eliminates fine-tuning and hyperparameter tuning.
- Limitations:
    - Higher API costs due to LLM usage.
    - Challenges in managing label granularity and polysemy.

Conclusion

- Summary:
  - Transformed text clustering into a classification task using LLMs.
  - Achieved superior performance compared to state-of-the-art methods.

Conclusion

- Summary:
  - Transformed text clustering into a classification task using LLMs.
  - Achieved superior performance compared to state-of-the-art methods.
- Future Work:
  - Incorporate user feedback for improved labels.
  - Explore cost-efficient and fine-grained clustering methods.

References i

Chen Huang and Guoxiu He.
Text clustering as classification with llms.
arXiv preprint arXiv:2410.00927, 2024.
URL `https://anonymous.4open.science/r/Text-Clustering-via-LLM-E500`.

Lawrence Hubert and Phipps Arabie.
Comparing partitions.
Journal of Classification, 2(1):193–218, 1985.
URL `https://link.springer.com/article/10.1007/BF01908075`.

Nguyen Xuan Vinh, Julien Epps, and James Bailey.
Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance.
Journal of Machine Learning Research, 11:2837–2854, 2010.
URL `https://www.jmlr.org/papers/volume11/vinh10a/vinh10a.pdf`.

Thank You

Question 1

What is the primary goal of text clustering?

A) To generate labeled datasets
B) To group similar texts based on their representations
C) To create embeddings for text analysis
D) To reduce dataset size

Question 2

What are the two stages of the proposed framework?

A) Label Generation and Clustering
B) Label Generation and Classification
C) Clustering and Embedding Fine-Tuning
D) Classification and Hyperparameter Tuning

Question 3

Which challenge does the proposed method address in traditional text clustering approaches?

A) Lack of datasets
B) Complexity of fine-tuning embedders and hyperparameter tuning
C) Low accuracy of clustering results
D) High computational requirements of small models

Question 4

What evaluation metrics were used in the experiments?

A) Precision, Recall, F1-Score

B) Accuracy, Normalized Mutual Information (NMI), Adjusted Rand Index (ARI)

C) BLEU, ROUGE, METEOR

D) Log-Loss, Cross-Entropy

Question 5

Which baseline methods were compared against the proposed framework?

A) IDAS, PAS, K-means, Keyphrase Clustering, ClusterLLM
B) Word2Vec, FastText, BERT
C) GANs, Transformers, Autoencoders
D) Sentence Transformers, T5, GPT-4

Question 6

What advantage does label merging provide in the proposed method?

A) Reduces API usage
B) Increases the number of clusters
C) Eliminates redundant labels and improves granularity
D) Lowers the need for training data

Question 7

Which dataset tasks were used in the experiments?

A) Sentiment analysis and machine translation
B) Topic mining, emotion detection, intent discovery, domain discovery
C) Summarization and text generation
D) Image recognition and text classification

Question 8

What is a key limitation of the proposed method?

A) Inconsistent results across datasets
B) High dependency on embeddings
C) Higher API costs due to reliance on LLMs
D) Limited dataset availability

Answer Key

1. **B)** To group similar texts based on their representations
2. **B)** Label Generation and Classification
3. **B)** Complexity of fine-tuning embedders and hyperparameter tuning
4. **B)** Accuracy, Normalized Mutual Information (NMI), Adjusted Rand Index (ARI)
5. **A)** IDAS, PAS, K-means, Keyphrase Clustering, ClusterLLM
6. **C)** Eliminates redundant labels and improves granularity
7. **B)** Topic mining, emotion detection, intent discovery, domain discovery
8. **C)** Higher API costs due to reliance on LLMs