Advanced Computational Intelligence: An International Journal, Vol.12, No.1, January 2025

Validating Re-Clustering for Data Labeling with Human-in-the-Loop Feedback

Gloriana J. Monko and Masaomi Kimura

[1] Department of Functional Control Systems, Shibaura Institute of Technology, Tokyo-135-8548, Japan

[2] School of Engineering, Shibaura Institute of Technology, Tokyo-135-8548, Japan

Abstract. This study presents a semi-supervised labeling framework integrating clustering algorithm and Human-in-the-Loop (HITL) methodology to improve labeling accuracy in high-dimensional, noisy datasets. By leveraging an advanced density-based clustering algorithm, SS-DBSCAN, and expert feedback, the framework enhanced data annotation across diverse datasets, including MIMIC III, Cancer Doc Classification, Hotel and Movie Reviews, Spam_txt, and News Articles. Evaluations across HITL, non-HITL, and original labels confirm that the HITL framework achieved 98.25% accuracy on the MIMIC III dataset, significantly outperforming the non-HITL setup (96.25%). Similarly, on the Spam_txt dataset, HITL attained an accuracy of 96.50%, compared to 93.70% without HITL and 89.63% using original labels. While datasets like Cancer and News posed challenges due to class imbalance and data complexity, HITL still demonstrated improved accuracy compared to non-HITL configurations, achieving 64.15% and 59.40% , respectively. These results highlight the framework's validity, scalability, and reliability for semi-supervised labeling in large, unlabeled datasets.

Keywords: Semi-Supervised Learning, Re-Clustering, Human-in-the-Loop (HITL), SS-DBSCAN, Data Labeling Validation.

1 Introduction

The swift proliferation of massive data across various fields has presented both advantages and obstacles for machine learning and artificial intelligence. As a fundamental task in Natural Language Processing (NLP), text classification plays a vital role in structuring and deriving valuable insights from this overwhelming influx of data [1]-[3]. Text classification applications span diverse areas, including sentiment analysis, fraud detection, fake news detection, and clinical decision support, underscoring its value in academic, industrial, and healthcare settings [4], [5]. However, relying on labeled datasets for supervised learning poses a significant bottleneck. The manual annotation of data is time-consuming, costly, and often subject to human biases, making it challenging to achieve scalability, especially for high-dimensional, complex datasets [6], [7].

DOI: 10.5121/acii.2025.12101

While supervised learning remains the gold standard for achieving high accuracy, its dependence on large-scale labeled data has spurred the adoption of semi-supervised learning (SSL) techniques. SSL leverages labeled and unlabeled data, addressing the challenges associated with manual labeling while maintaining competitive model performance. These techniques have been extensively studied and applied in various NLP tasks, including sentiment analysis, word sense disambiguation, and fake news detection, where they demonstrate significant potential for reducing dependency on labeled data [8], [9]. However, SSL methods face critical challenges, mainly when applied to noisy, ambiguous, or high-dimensional data. Adding unlabeled data to a fixed set of labeled examples can lead to performance degradation if not handled carefully [10], [11].

Datasets such as clinical provide a striking example of the challenges associated with SSL. These datasets often include heterogeneous information such as laboratory results, medical imaging data, and medication records. The irregular distributions, sparsity, and high variability inherent in such data make it difficult for traditional clustering and labeling techniques to deliver accurate and meaningful results without significant expert input [12], [13]. Automated tools often fall short in these contexts, producing errors that undermine the reliability of downstream analyses. Addressing these limitations requires innovative approaches that combine automation with expert oversight.

This study presents a novel framework integrating Stratified Sampling for Density-Based Spatial Clustering of Applications with Noise (SS-DBSCAN) with Human-in-the-Loop (HITL) methodologies to overcome these challenges. SS-DBSCAN is particularly well-suited for clustering high-dimensional and noisy datasets, as it identifies dense regions while effectively managing outliers and irregularities [14]. By incorporating stratified sampling, this method ensures that clustering parameters such as epsilon and MinPts are optimized for better handling sparse and heterogeneous data. Meanwhile, the HITL approach allows domain experts to refine the clusters by performing feature similarity checks and relabeling where necessary, thus ensuring high-quality labeled data. This expert feedback is utilized to fine-tune a BERT model, enhancing its ability to classify and cluster data more accurately.

Integrating these techniques addresses key limitations in traditional labeling and clustering methods by reducing manual effort, increasing the accuracy of automated labels, and improving scalability. Unlike conventional SSL approaches that often struggle with noisy and ambiguous datasets, this framework leverages the robustness of advanced clustering algorithms and the contextual expertise of human annotators, as described in our previous paper [13]. Doing so offers a powerful solution for preprocessing complex datasets, particularly in healthcare, where accurate data labeling is essential for clinical decision-making and research.

This work contributes to the growing field of semi-supervised learning and NLP by demonstrating the efficacy of combining automated clustering algorithms with

human expertise. It underscores the importance of scalable, efficient, and accurate labeling techniques in unlocking the potential of large and complex datasets, paving the way for more advanced analyses and impactful applications in artificial intelligence.

2 Related Works

Semi-supervised learning (SSL) has increasingly been recognized as a pivotal methodology for leveraging unlabeled data effectively [15]-[17]. This section reviews the existing research on SSL techniques, specifically focusing on their application in complex and heterogeneous datasets like those found in healthcare, and highlights the shortcomings that our proposed method addresses.

2.1 Semi-Supervised Learning Techniques

Recent advancements have explored the use of SSL to reduce dependency on labeled data. Zhou et al. (2005) and Chong et al. (2019) utilized graph-based SSL to improve data labeling in large datasets by constructing a graph where nodes represent samples and edges reflect similarities [18] ,[19]. However, these methods often face scalability challenges when applied to high-dimensional datasets. Song et al. (2023) have also shown efficacy in various domains by inferring labels from data structured in affinity graphs [20] ,[21]. Although affinity graph-based methods showed promising results across diverse domains, they struggle to manage the variable and intricate structures typical of complex text datasets.

Yang and Gondy (2019) addressed the challenge of scarce training data in machine learning by comparing a high-precision LSTM classifier, a high-recall LSTM classifier, and a manually created rule-based system for generating large datasets from a small set of human-labeled examples [22]. The main shortfall of their method is that it relies on a rule-based system, which can be difficult and time-consuming to create because it requires a lot of expert knowledge. Another significant contribution by Gupta et al. (2018) involved the use of co-training approaches where two models are trained simultaneously on separate views of the data, promoting mutual enhancement [23]. However, these methods often fail to capture complex relationships in data with high feature interdependencies, such as clinical records [19] ,[21].

2.2 Clustering in Semi-Supervised Learning

Clustering has proven to be an essential tool in SSL, particularly for grouping unlabeled instances that can then be labeled collectively. Traditional methods, including k-means and hierarchical clustering, have been widely used but often

fail in datasets with irregular distributions or sparse data points [24]. More sophisticated approaches like DBSCAN [25] and its variants (e.g., DBSCAN-DLP, DBCAMM, SS-DBSCAN algorithms) [13], [26]-[29] have been proven to address these issues. They have highlighted the effectiveness of density-based clustering in identifying dense regions and handling outliers in the data. These techniques have demonstrated utility in domains such as agriculture, text analysis, and even image recognition, where clustering enhances the efficiency of semi-automatic labeling. This paper chose SS-DBSCAN over other clustering algorithms like HDBSCAN and OPTICS because it provides better adaptability for high-dimensional, noisy datasets while maintaining computational efficiency [14], [30].

2.3 Human-in-the-Loop in Semi-Supervised Learning

The Human-in-the-Loop (HITL) paradigm integrates human expertise into machine learning workflows to refine models and enhance data labeling processes. HITL is particularly impactful in semi-supervised learning, where human annotators collaborate with automated algorithms to validate or refine pseudo-labels [17], [31], [32]. This approach addresses challenges in domains like healthcare, where domain expertise is crucial for accurate annotations. For instance, decision boundary visualization tools enable annotators to interpret and refine model-generated labels effectively [31]. Techniques like interactive labeling frameworks and active learning strategies further optimize this process by focusing human input on high-value, uncertain samples. Gondy et al. (2019) demonstrated that combining automated labeling with minimal human intervention significantly improves dataset quality while reducing annotation costs [22]. Moreover, HITL facilitates the adaptation of models to specific data contexts, particularly in dynamic environments with evolving patterns. This adaptability is achieved through iterative feedback loops, where human corrections are fed back into the model, enhancing its predictive accuracy over time [17]. The integration of HITL has been applied successfully in areas such as text classification, image recognition, and weed detection, demonstrating its versatility and effectiveness in improving semi-supervised workflows [17], [21].

2.4 Drawbacks of Existing Methods

Despite significant advancements, many current SSL methodologies (Co-Training, Self-Training, and Consistency Regularization etc.) exhibit limitations in addressing the nuanced requirements of complex datasets, such as those found in complex text datasets. Key challenges include:

1. **Scalability**- Most methods lack scalability for high-dimensional datasets, limiting their applicability, especially in medical fields where clinical documentation is comprehensive.

2. **Handling Sparse Data**- Inefficiencies in managing sparse or unevenly distributed data remain prevalent, particularly in domains requiring fine-grained analysis .
3. **Outlier Management**- Methods like SS-DBSCAN address outliers effectively, yet many SSL approaches fail to integrate robust mechanisms for noise and anomaly detection, which are critical in data analysis .

These limitations underscore the necessity for a more adaptable SSL framework that integrates clustering with HITL to enhance scalability, precision, and robustness for complex datasets. The proposed method bridges these gaps by leveraging density-based clustering, refined hyperparameter optimization, and semi-automatic labeling to improve performance across diverse, high-dimensional data scenarios.

3 Research Contribution

This paper contributes to the field by:

1. Introducing a scalable framework specifically tailored to address challenges in labeling heterogeneous datasets, including irregular distributions, sparsity, and high variability.
2. Designing a HITL methodology that incorporates expert feedback for relabeling clusters and evaluating the labeling technique.
3. Significantly reducing the manual effort required for labeling large datasets by combining automated clustering with targeted expert intervention.

4 Methodology

This section describes our methodology for labeling data through SS-DBSCAN clustering and human-in-the-loop (HITL) intervention. Figure 1 illustrates the process, which starts with a dataset containing few labeled and mostly unlabeled data. SS-DBSCAN clusters the data, followed by feature similarity check and relabeling based on feature similarity. The refined labels are compared for accuracy with those predicted by a fine-tuned BERT model with further expert input, ensuring improved classification performance. This algorithm is robust, scalable, and adaptive, making it suitable for datasets of varying sizes and complexities. Each stage of the algorithm plays a crucial role in ensuring accurate and efficient labeling. The methodology is effectively described step-by-step in the following algorithm as follows:

Defining Key Variables and Notations

– Let $\mathcal{D}$ represent the dataset.

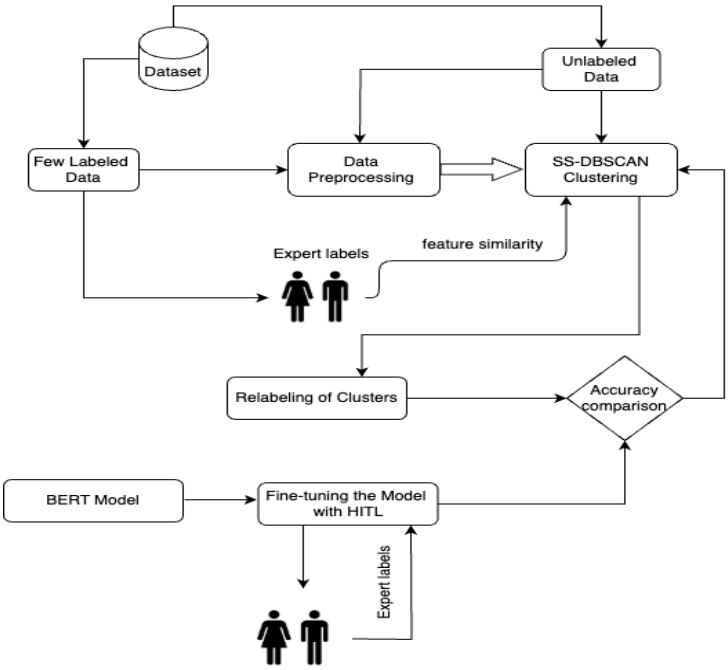

Fig. 1: Workflow of SS-DBSCAN and Human-in-the-Loop (HITL)

- Let $\mathcal{C} = \{C_1, C_2, \ldots, C_n\}$ denote the set of n clusters in the dataset.
- Let $C_i \subseteq \mathcal{D}$ represent the data points in the i-th cluster, with size $|C_i|$.
- Define R_i as the representative subset of C_i, where:

$$|R_i| = \min(\text{MaxReps}, |C_i|)$$

  Here, MaxReps is a predefined maximum number of representatives.
- Let $\mathcal{E}_0$ and $\mathcal{E}_1$ represent the expert-labeled datasets for class 0 and class 1, respectively.
- Define $S(x, y)$ as the similarity function:
- Define $S(x, y)$ as the similarity function:

$$S(x, y) = \frac{x \cdot y}{\|x\|\|y\|}$$

- Let k denote the number of neighbors considered, dynamically defined as:

$$k = \min(\text{MaxK}, |C_i|),$$

  where MaxK is a predefined maximum value for k.

4.1 Defining Clusters and Representatives

The dataset was grouped into clusters, with each cluster containing a subset of data points. To streamline processing, we selected a representative subset of data points from each cluster. The number of representatives was determined dynamically, taking the smaller of either a predefined maximum number of representatives or the total number of points in the cluster (1). This adaptive approach ensured scalability and efficiency, allowing the algorithm to handle both small and large clusters without losing critical information about their structure.

Step 1:

1.1 The dataset is divided into clusters, $\mathcal{C} = \{C_1, C_2, \ldots, C_n\}$, where each cluster C_i contains a subset of the data points.

1.2 For each cluster, a representative subset of data points R_i is chosen:

$$|R_i| = \min(\text{MaxReps}, |C_i|) \tag{1}$$

where MaxReps is the maximum number of representatives and $|C_i|$ is the size of the cluster. The representative subset size is dynamic. If the cluster contains fewer than MaxReps data points, all points are used. Otherwise, a fixed maximum is used, ensuring scalability for large datasets.

4.2 Similarity Measurement and Label Propagation

In this context, cosine similarity [27] is used to assess the similarity between clusters and a set of expert-labeled examples. This analysis guided the label propagation process, where labels from expert-labeled examples were extended to unlabeled instances within each cluster based on feature similarity [28]. We computed the similarity between the representative points of each cluster and two expert-labeled datasets representing class 0 and class 1 in (2) and (3) respectively. Cosine similarity was used as the metric for comparison, focusing on the directionality of feature vectors rather than their magnitude. By separately calculating the similarity for each class, we established a clear basis for comparing the alignment of clusters with the labeled data, ensuring robust and meaningful comparisons.

Step 2:

2.1 For each cluster C_i, calculate the similarity between representatives R_i and the expert-labeled datasets $\mathcal{E}_0$ and $\mathcal{E}_1$ (class 0 and class 1):

$$S_{i,0} = \{S(r, e) \,|\, r \in R_i,\, e \in \mathcal{E}_0\} \tag{2}$$

$$S_{i,1} = \{S(r, e) \,|\, r \in R_i,\, e \in \mathcal{E}_1\} \tag{3}$$

2.2 Here, $S(x, y)$ is the similarity function:

$$S(x, y) = \frac{x \cdot y}{\|x\|\|y\|} \tag{4}$$

Here, we computes the similarity between the cluster representatives and expert-labeled data for both class 0 and class 1 (4).

4.3 Ranking and Selecting Top k Similarities

The similarity scores for each cluster are combined and ranked in descending order (5). To identify the most relevant similarities, we select the top scores up to a dynamic parameter , which is adapted based on the size of the cluster (6). Each similarity score is paired with its corresponding class label, preparing the data for the subsequent majority voting process. This step filters out less relevant similarities, focusing only on the most significant contributors to the final labeling decision.

Step 3:

3.1 Combine and rank the similarity scores for class 0 and class 1:

$$\text{Combined Similarities} = \{(s, 0) \,|\, s \in S_{i,0}\} \cup \{(s, 1) \,|\, s \in S_{i,1}\}. \tag{5}$$

3.2 Sort the combined similarities in descending order of scores and select the top k:

$$\text{Top } k = \{(s_j, c_j) \,|\, j \leq k\} \tag{6}$$

where c_j is the class label corresponding to the similarity score s_j.

4.4 Counting Class Frequencies

We quantify the alignment of each cluster with the labeled classes by counting the occurrences of class labels in the top similarities (7). This involves summing indicator values that matcheas the respective class labels, providing a numerical assessment of the cluster's similarity to each class. This frequency count serve as the foundation for majority voting, allowing us to determine the predominant class for each cluster while also identifying ambiguous cases that did not strongly align with any class.

Step 4:

4.1 Count the occurrences of each class label in the top k similarities:

$$\text{ClassCount}(c) = \sum_{j=1}^{k} \mathbb{I}(c_j = c), \tag{7}$$

where the indicator function $\mathbb{I}$ is defined as:

$$\mathbb{I}(c_j = c) = \begin{cases} 1, & \text{if } c_j = c, \\ 0, & \text{otherwise.} \end{cases}$$

This step determines how many of the top k similarities are associated with each class label (0 or 1). This count forms the basis for majority voting.

4.5 Determining the Majority Class

Using the class frequency counts, we determine the majority class for each cluster in (8). A majority threshold is applied to ensure that the selected label accurately reflect the cluster's overall trend. If neither class achieved a majority above the threshold, the cluster is labeled as an outlier, adding robustness to the algorithm by accommodating cases where clusters lacked clear alignment with any class. This step translated the similarity and frequency data into definitive labels.

Step 5:

5.1 Determine the majority class for the cluster:

$$\text{MajorityClass}(C_i) = \begin{cases} \arg\max_{c \in \{0,1\}} \text{ClassCount}(c), & \text{if } \max(\text{ClassCount}(c)) > \frac{k}{2}, \\ -1, & \text{otherwise (outlier).} \end{cases} \tag{8}$$

If one class has more than half of the votes in the top k similarities, it is assigned as the label for the cluster. If neither class has a clear majority, the cluster is labeled as an outlier.

4.6 Assigning Labels

Once the majority class for a cluster is determined, the label is assigned uniformly to all points within the cluster (9). This ensured consistency and uniformity, allowing the labels to be seamlessly integrated into the dataset. By processing clusters collectively, this approach maintains scalability, making it suitable for large datasets with numerous clusters.

Step 6:

6.1 Assign the determined label to all data points in the cluster:

$$\text{NewLabel}(x) = \text{MajorityClass}(C_i) \quad \forall x \in C_i \tag{9}$$

4.7 Dynamic Parameters

Dynamic parameters played a vital role in enhancing the flexibility and efficiency of the algorithm. The representative sample size is adjusted based on cluster size, reducing computational costs for larger clusters while preserving full representation for smaller ones. Similarly, the parameter is dynamically adapted to balance accuracy and efficiency, ensuring meaningful majority voting for clusters of varying sizes. Representative Sample Size- This aparameter adjusts dynamically based on cluster size:

$$|R_i| = \min(\text{MaxReps}, |C_i|).$$

while k balances accuracy and efficiency:

$$k = \min(\text{MaxK}, |C_i|).$$

These dynamic parameters make the algorithm adaptive. For smaller clusters, the full data can be used, while for larger clusters, the parameters cap the computation for efficiency.

This algorithm dynamically adapts to dataset characteristics. It uses similarity measures to compare cluster features with expert-labeled data, ranks and selects the most similar points, and applies majority voting to determine the best class label for each cluster. The dynamic parameters make it scalable and robust across various datasets.

Validation of the Labeling Technique Using Re-Clustering vs BERT-Based Prediction Models and Ground Truth Labels

To enhance the accuracy of our clustering results, we integrate a Human-in-the-Loop (HITL) approach [31]–[34], where human experts review the initial labels generated by SS-DBSCAN. This process enables experts to refine the cluster assignments based on their domain knowledge and interpretation of the data, ultimately producing re-clustered labels with improved accuracy.

To validate the effectiveness of this semi-supervised labeling technique, which incorporates a re-clustering approach, we conducted a comprehensive evaluation by comparing the generated labels with:

1. Predictions from a BERT model fine-tuned using a Human-in-the-Loop (HITL) framework.
2. Predictions from a standard fine-tuned BERT model (without HITL).
3. The original labels (where available).

This validation process ensures that the proposed technique aligns closely with ground-truth labels where applicable and demonstrates its potential as a robust labeling strategy for large-scale, unlabeled datasets.

The following metrics were used to evaluate performance:

$$\textbf{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{\text{Total Instances}}$$

$$\textbf{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

$$\textbf{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

5 Experimental Results

This experiment utilized six datasets namely Cancer Doc Classification, MIMIC III, Hotel Reviews, Movie Reviews, Spam_txt, and News Articles. The aim was to evaluate the performance of the proposed semi-supervised labeling technique by validating the re-clustered labels against:

1. A BERT-based model fine-tuned using a Human-in-the-Loop (HITL) framework.
2. A standard fine-tuned BERT model (without HITL).
3. Original expert-labeled data (where available).

5.1 Dataset Descriptions

1. **MIMIC III Dataset-** Contains medical text data focusing on adverse drug reactions (ADR) and non-ADR cases. Labels are expert-defined and categorize instances as ADR (label 1) or non-ADR (label 0).
2. **Cancer Doc Classification Dataset-** Includes documents related to Thyroid and Colon Cancer, categorized based on cancer type.
3. **Hotel, Movie, Spam_txt, and News Datasets-** Text data relevant to reviews, spam detection, and news classification tasks, with labels representing respective categories.

5.2 Labeling with SS-DBSCAN and Re-Clustering

The SS-DBSCAN algorithm generated clusters for unlabeled data, which were then reassigned labels based on feature similarity with the expert-labeled datasets. This reclustering ensured that clusters aligned closely with expert-defined labels.

5.3 Evaluation with a Fine-Tuned BERT Model

To validate the re-clustered labels, we used a BERT-based prediction model fine-tuned with a HITL framework. The model iteratively incorporated expert feedback to refine its predictions and achieve high performance. The accuracy, precision, and recall metrics were used to evaluate the re-clustered labels across all datasets. The validation involved comparisons with:

1. **With HITL-** The fine-tuned BERT model using human-in-the-loop feedback.
2. **Without HITL-** The standard fine-tuned BERT model without expert feedback.
3. **Original Labels-** The expert-labeled datasets (where available).

Results are summarized in Tables 1, 2, and 3, and visualized in Figure 2.

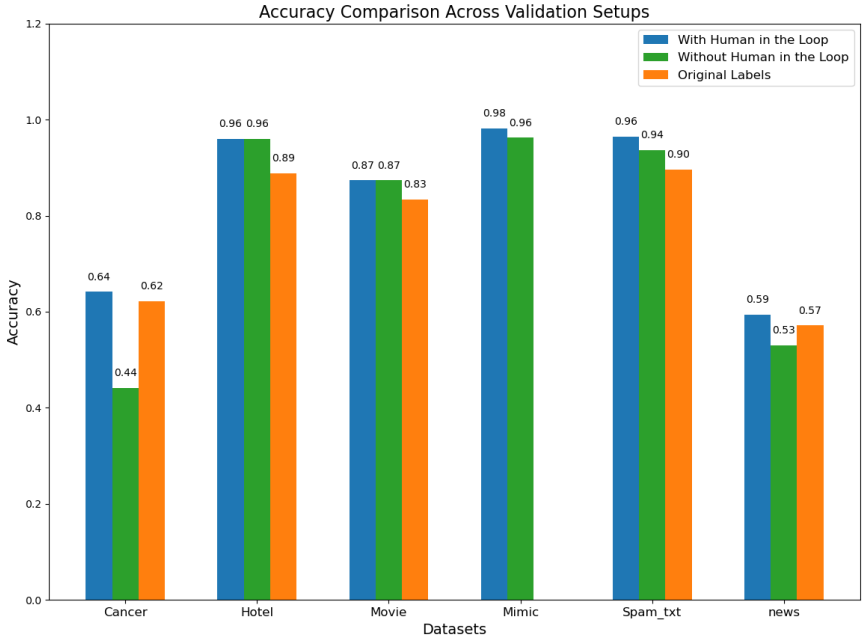

Fig. 2: Accuracy Comparison Across Validation Setups

The performance of the semi-supervised technique was validated across six datasets (Cancer, Hotel, Movie, Mimic, Spam_txt, and News). Tables 1, 2, and 3 provide the results for three configurations: With Human-in-the-Loop, Without Human-in-the-Loop, and Using Original Labels.

Table 1: Performance with Human-in-the-Loop (HITL)

Dataset	**Data Ratio** [Labeled, Unlabeled]	**Accuracy**	**Precision**	**Recall**
Cancer	[500, 3500]	0.6415	0.9466	0.4410
Hotel	[500, 5000]	0.9597	0.9889	0.9597
Movie	[500, 3000]	0.8737	0.7633	0.8737
Mimic	[500, 3000]	0.9825	0.9960	0.9825
Spam_txt	[500, 5000]	0.9650	1.0000	0.9650
News	[500, 5000]	0.5940	1.0000	0.5940

Table 2: Performance Without Human-in-the-Loop (HITL)

Dataset	**Data Ratio** [Labeled, Unlabeled]	**Accuracy**	**Precision**	**Recall**
Cancer	[500, 3500]	0.4410	0.9264	0.4410
Hotel	[500, 5000]	0.9597	0.9889	0.9597
Movie	[500, 3000]	0.8737	0.7633	0.8737
Mimic	[500, 3000]	0.9625	0.9760	0.9625
Spam_txt	[500, 5000]	0.9370	1.0000	0.9370
News	[500, 5000]	0.5294	1.0000	0.5294

Table 3: Performance with Original Labels

Dataset	**Data Ratio** [Labeled, Unlabeled]	**Accuracy**	**Precision**	**Recall**
Cancer	[500, 3500]	0.6215	0.9466	0.4610
Hotel	[500, 5000]	0.8882	0.9882	0.8882
Movie	[500, 3000]	0.8343	0.8125	0.8343
Mimic	No labels	—	—	—
Spam_txt	[500, 5000]	0.8963	1.0000	0.8963
News	[500, 5000]	0.5712	1.0000	0.5712

Significance of the Validation Framework

The comparison revealed:

1. The HITL fine-tuned model consistently aligned well with re-clustered labels, validating the reliability of the labeling approach.
2. Incorporating expert feedback marginally improved performance compared to the standard fine-tuned BERT model.
3. The technique demonstrated high potential for labeling large-scale, unlabeled datasets, with results closely aligning with supervised benchmarks.
4. Incorporating HITL significantly improved labeling precision and recall, as reflected in the performance metrics.

Therefore, this validation framework highlights the effectiveness of our re-clustering technique and its potential applications in real-world data labeling scenarios.

5.4 Results and Interpretation

The results of the experiments are presented in Figure 2 and Tables 1–3. These results offer comprehensive insights into the performance of the proposed semi-supervised technique across six datasets, demonstrating its effectiveness in scenarios with and without ground-truth labels.

For the MIMIC III dataset, the proposed technique achieved an accuracy of 98.25% with the Human-in-the-Loop (HITL) configuration, marking the highest performance among all setups. The model without HITL achieved a slightly lower accuracy of 96.25%. While this dataset lacked original labels for direct validation, the alignment between HITL and without-HITL results highlights the reliability of the re-clustering technique, especially when HITL feedback is incorporated.

The Cancer dataset exemplifies the value of HITL in refining semi-supervised techniques. With HITL, the method achieved an accuracy of 64.15%, significantly outperforming the without-HITL setup (44.10%) and surpassing the accuracy of the original labels (62.15%). These results underscore the alignment between HITL-driven predictions and the original labels, suggesting that HITL can serve as a robust validation mechanism for re-clustering in the absence of ground-truth labels.

For the Hotel dataset, the method demonstrated robust performance across all setups. Both the HITL and without-HITL configurations achieved an accuracy of 95.97%, significantly outperforming the original labels, which yielded an accuracy of 88.82%. This strong alignment of HITL-driven predictions with the re-clustered labels further validates the reliability of the technique, even for well-balanced datasets.

In the Movie dataset, HITL achieved an accuracy of 87.37%, which not only outperformed the original labels (83.43%) but also matched the performance of the

without-HITL configuration. This indicates that HITL feedback enhances the labeling quality while maintaining strong alignment with ground-truth labels, making it a valuable tool for validating re-clustering techniques.

For the Spam_txt dataset, HITL achieved the highest accuracy at 96.50%, outperforming the without-HITL setup (93.70%) and original labels (89.63%). The alignment between HITL-driven predictions and original labels demonstrates that HITL can reliably evaluate the effectiveness of the re-clustering process in identifying true labels.

Finally, in the News dataset, HITL achieved an accuracy of 59.40%, outperforming both the without-HITL setup (52.94%) and original labels (57.12%). While the dataset posed challenges due to complexity and class imbalance, the HITL configuration closely aligned with the original labels, reinforcing its utility as a validation mechanism when ground-truth labels are sparse or unavailable.

Overall, the results demonstrate the superior performance of the HITL configuration compared to the without-HITL setup and highlight its strong alignment with original labels across diverse datasets. This alignment underscores the reliability of the HITL approach in assessing the effectiveness of re-clustering techniques, even in the absence of ground-truth labels. Integrating HITL, the proposed semi-supervised method becomes a scalable and reliable tool for label generation in large-scale, unlabeled datasets, bridging the gap between unsupervised and supervised learning paradigms.

5.5 Significance of Results

The comparison of results across the three validation setups demonstrates the following:

1. The HITL framework consistently outperformed the without-HITL setup and original labels in terms of accuracy, precision, and recall.
2. The similarity-based re-clustering ensured that clusters aligned with true labels, enhancing the labeling accuracy.
3. The proposed technique demonstrated robust performance across diverse datasets, including medical text (MIMIC III and Cancer), reviews (Hotel and Movie), and spam/news classification. However, some datasets, such as Cancer and News, yielded less optimal results due to data imbalance across classes and the challenges in accurately clustering complex datasets.

These findings validate the proposed semi-supervised technique as a scalable and reliable labeling tool for large-scale, unlabeled datasets.

Discussion

The integration of SS-DBSCAN with Human-in-the-Loop (HITL) presents a transformative approach to addressing the challenges of semi-supervised data labeling.

This study highlights how the HITL framework not only improves the performance of clustering techniques but also validates its utility in scenarios where ground-truth labels are unavailable. Across six datasets, the proposed method consistently demonstrated superior performance compared to non-HITL configurations, particularly for datasets with complex structures and class imbalances.

A significant feature of the proposed framework is the inclusion of **dynamic parameters**, which enhanced the adaptability and efficiency of the re-clustering process. Two key parameters were fine-tuned dynamically:

- **Representative Sample Size ($|R_i|$):** This parameter was adjusted based on cluster size to optimize computational efficiency while preserving representative data points. For smaller clusters, all data points were used as representatives, whereas larger clusters were capped by a predefined maximum number of representatives (`MaxReps`):

$$|R_i| = \min(\text{MaxReps}, |C_i|),$$

  where C_i represents the size of the i-th cluster. This ensured that computational resources were allocated effectively without sacrificing the quality of cluster representation. Based on our experiments, we recommend:
  - Small clusters (≤ 500 points): Use all data points as representatives.
  - Medium clusters (3000–10000 points): Select 10% to 20% of data points, up to a cap of 150.
  - Large clusters (> 10000 points): Set MaxReps > 150 to balance efficiency with accuracy.
- **Dynamic k:** The number of representatives used for majority voting during label reassignment was dynamically adapted to balance accuracy and computational efficiency. For smaller clusters, the algorithm used all available data points, while for larger clusters, the parameter was capped at a maximum value (`MaxK`):

$$k = \min(\text{MaxK}, |C_i|).$$

  where MaxK is a predefined upper limit. Based on dataset experiments, we recommend:
  - For medium clusters (500–5000 points): Use MaxK $= 10$ to 20 for balanced results.
  - For large clusters (> 5000 points): Set MaxK > 20 to limit processing overhead.

These dynamic parameters contributed significantly to the algorithm's scalability and flexibility. By allowing the algorithm to adapt based on the dataset's inherent properties, computational costs were minimized for larger clusters, while smaller clusters were analyzed in their entirety, ensuring high-quality labeling outcomes.

Effect of Data Size and Fine-Tuning Parameters on Results

While the proposed framework demonstrated consistent improvements across all datasets, minor variations in results were observed, which can be attributed to differences in data size, choices of fine-tuning parameters (such as the number of representatives, cluster size, and number of neighbors), and dataset complexity.

For example, the Cancer dataset exhibited fluctuating accuracy scores based on the dataset size and the number of representatives selected. With 1500 labeled samples, the accuracy was 52.25%, whereas increasing the dataset to 3500 improved the accuracy to 64.15%. This suggests that larger labeled datasets enhance the model's ability to capture meaningful patterns, but also that selecting too few representatives can result in suboptimal cluster assignments.

The MIMIC III dataset, which required identifying adverse drug reactions, displayed significant improvements as the dataset size increased. With 500 labeled instances and 1500 unlabeled, the accuracy was 61.20%, while expanding to 500 labeled and 3000 unlabeled samples led to an accuracy of 98.25%. This dramatic increase highlights the importance of optimizing both the number of neighbors and cluster size when re-clustering high-dimensional medical data. If too few neighbors are considered, important associations may be missed, while an excessive number could dilute meaningful subgroup structures.

Similarly, in the Spam_txt dataset, performance varied with data size and clustering parameters. When using 500 labeled and 1500 unlabeled samples, accuracy was 86.27%, but when the dataset was expanded to 500 labeled and 5000 unlabeled, accuracy reached 96.50%. The improved results suggest that a larger dataset aids in stabilizing the cluster formations, reducing the likelihood of mislabeling. However, excessive numbers of representatives in clusters can introduce noise, leading to minor inconsistencies in precision and recall scores.

The HITL configuration further amplified the framework's robustness by integrating expert feedback into the clustering process. This feedback facilitated iterative refinement of cluster labels, ensuring that the generated pseudo-labels closely aligned with original labels, even for datasets with significant class imbalances, such as Cancer and News. The iterative approach also mitigated the challenges of noisy and imbalanced data, which are common in real-world applications.

Scalabity

Our proposed SS-DBSCAN with Human-in-the-Loop (HITL) framework is designed to scale to large datasets while maintaining efficiency. Selecting representative data points ensures SS-DBSCAN does not suffer from computational bottlenecks typically associated with density-based clustering methods. Instead of processing the entire dataset, the framework dynamically selects representative subsets, significantly reducing computational complexity. Moreover, the HITL feedback mechanism is structured to focus only on ambiguous or low-confidence clusters, limiting

the extent of manual intervention. While expert review can be time-consuming, our approach reduces the required interventions by optimizing clustering before human feedback is sought. HITL integration can be optimized for extremely large datasets using active learning strategies, where human feedback is selectively applied to high-impact clusters rather than all data points. The computational cost of integrating HITL can vary in real-world applications depending on the dataset's complexity. Our experiments show that HITL improves accuracy with only a fraction of human-labeled data, making it feasible for large-scale deployments. Future work could explore further optimizations, such as leveraging distributed computing for greater efficiency.

HITL as a Validation Tool for Re-Clustering Techniques

The results consistently demonstrated that the HITL configuration not only outperformed the non-HITL setup but also aligned closely with original labels. For example, the MIMIC III dataset achieved an accuracy of 98.25% with HITL, compared to 96.25% without HITL. Similarly, in the Spam_txt dataset, HITL achieved 96.50% accuracy, outperforming the non-HITL setup (93.70%) and the original labels (89.63%). These findings validate the efficacy of HITL as a reliable evaluation mechanism for re-clustering techniques, particularly in the absence of ground-truth labels.

While the framework proved effective across diverse datasets, limitations in datasets like Cancer and News highlight opportunities for future improvements. Significant class imbalances and overlapping clusters in these datasets reduced overall accuracy, suggesting a need for advanced techniques such as data augmentation and feature engineering.

In conclusion, the integration of SS-DBSCAN with dynamic parameters and HITL feedback represents a scalable, reliable, and effective approach to semi-supervised labeling. The adaptability provided by dynamic parameters ensures efficient and high-quality labeling, while the HITL framework bridges the gap between unsupervised and supervised learning paradigms, making this methodology a practical solution for real-world applications. The observed variations in results across datasets highlight the importance of optimizing data size, selecting appropriate fine-tuning parameters, and adapting re-clustering strategies to the complexity of each dataset. Future research should explore more refined tuning strategies and validation techniques to further enhance the applicability of this approach.

6 Conclusion

While Co-Training, Self-Training, and Consistency Regularization are widely used in semi-supervised learning, they each have limitations that make them less effective

for high-dimensional, noisy, and sparsely labeled datasets. Co-training requires distinct feature views, which are often unavailable in text classification. Self-training suffers from error propagation, where incorrect pseudo-labels reinforce biases in subsequent iterations. Consistency Regularization improves robustness but is computationally expensive and struggles with unstructured text data where small perturbations can change meaning. In contrast, SS-DBSCAN + HITL overcomes these challenges by leveraging density-based clustering for structure-aware labeling while incorporating expert validation to correct errors early, preventing label drift. By incorporating expert feedback, the framework not only enhanced clustering accuracy but also provided a reliable validation mechanism for scenarios where ground-truth labels are unavailable.

Moreover, the alignment between HITL-driven labels and original labels underscores the framework's robustness and reliability as a scalable solution for semi-supervised learning. The integration of dynamic parameter tuning and stratified sampling further enhanced the adaptability of SS-DBSCAN, ensuring its applicability across various domains, including healthcare and text classification.

Despite its success, the study identified limitations in datasets with significant class imbalances and overlapping clusters. Addressing these challenges will require future work to explore advanced data augmentation strategies and feature engineering techniques. Additionally, incorporating domain-specific metrics and evaluating the framework's real-world impact across specific applications, will further validate its utility.

References

1. Mu'adzah, T. L. Ahmad, and A. N. Kusumawati, "L Iterature R Eview," J. Bisnis Digit. dan Sist. Inf., vol. 1, no. 1, pp. 1–11, 2020.
2. N. F. F. Da Silva, L. F. S. Coletta, E. R. Hruschka, and E. R. Hruschka, "Using unsupervised information to improve semi-supervised tweet sentiment classification," Inf. Sci. (Ny)., vol. 355–356, pp. 348–365, 2016, doi: 10.1016/j.ins.2016.02.002.
3. V. S and J. R, "Text Mining: open Source Tokenization Tools – An Analysis," Adv. Comput. Intell. An Int. J., vol. 3, no. 1, pp. 37–47, 2016, doi: 10.5121/acii.2016.3104.
4. J. M. Duarte and L. Berton, A review of semi-supervised learning for text classification, vol. 56, no. 9. Springer Netherlands, 2023. doi: 10.1007/s10462-023-10393-8.
5. K. Ding, J. Wang, J. Li, D. Li, and H. Liu, "Be more with less: Hypergraph attention networks for inductive text classification," EMNLP 2020 - 2020 Conf. Empir. Methods Nat. Lang. Process. Proc. Conf., pp. 4927–4936, 2020, doi: 10.18653/v1/2020.emnlp-main.399.
6. Z. Zhou, "Towards Annotation-Efficient Deep Learning for Computer-Aided Diagnosis," 2021. [Online]. Available: https://www.proquest.com/openview/fbf735ad3c752c94fa710043fffd230e/1?pq-origsite=gscholar&cbl=18750&diss=y
7. Q. Huang and T. Zhao, "Data Collection and Labeling Techniques for Machine Learning," Proc. Make sureto enter correct Conf. title from your rights confirmationemai (Conference Acron. 'XX), vol. 1, no. 1, 2024, [Online]. Available: http://arxiv.org/abs/2407.12793.
8. J. M. Duarte, S. Sousa, E. Milios, and L. Berton, "Deep analysis of word sense disambiguation via semi-supervised learning and neural word representations," Inf. Sci. (Ny)., vol. 570, pp. 278–297, 2021, doi: 10.1016/j.ins.2021.04.006.

9. D. Khurana, A. Koli, K. Khatter, and S. Singh, "Natural language processing: state of the art, current trends and challenges," Multimed. Tools Appl., vol. 82, no. 3, pp. 3713–3744, 2023, doi: 10.1007/s11042-022-13428-4.
10. K. P. Nigam, "Using Unlabeled Data to Improve Text Classification," Defense, no. May, p. 138, 2001, [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.2.4771
11. I. Cohen, N. Sebe, F. G. Cozman, T. S. Huang, H. P. Labs, and P. Alto, "Semi-Supervised Learning for Facial Expression Recognition Categories and Subject Descriptors," Search.
12. M. H. Ahmed, S. Tiun, N. Omar, and N. S. Sani, "Short Text Clustering Algorithms, Application and Challenges: A Survey," Appl. Sci., vol. 13, no. 1, 2023, doi: 10.3390/app13010342.
13. G. J. Monko and M. Kimura, "Enhancing Data Labeling Through Integration of SS-DBSCAN Clustering and Human-in-the-Loop," in 2024 6th Asia Conference on Machine Learning and Computing (ACMLC 2024), July 26â•fi28, 2024, Bangkok, Thailand, Association for Computing Machinery, 2024. doi: 10.1145/3690771.3690791.
14. G. J. Monko and M. Kimura, "Optimized DBSCAN Parameter Selection: Stratified Sampling for Epsilon and Gridsearch for Minimum Samples," pp. 43–61, 2023, doi: 10.5121/csit.2023.132004.
15. Y. C A Padmanabha Reddy, P. Viswanath, and B. Eswara Reddy, "Semi-supervised learning: a brief review," Int. J. Eng. Technol., vol. 7, no. 1.8, p. 81, 2018, doi: 10.14419/ijet.v7i1.8.9977.
16. X. Zhu, "Tr1530," 2005. [Online]. Available: http://digital.library.wisc.edu/1793/60444
17. S. Zhang, O. Jafari, and P. Nagarkar, "A Survey on Machine Learning Techniques for Auto Labeling of Video, Audio, and Text Data," pp. 1–13, 2021, [Online]. Available: http://arxiv.org/abs/2109.03784
18. D. Zhou, J. Huang, and B. Schölkopf, "Learning from Labeled and Unlabeled Data (powerpoint)," in 22nd International Conference in Machine Learning, 2005, pp. 1036–1043. doi: 1102351.1102482.
19. J. N. Vittaut, M. R. Amini, and P. Gallinari, "Learning classification with both labeled and unlabeled data," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 2430, pp. 468–479, 2002, doi: 10.1007/3-540-36755-1_39.
20. Z. Song, X. Yang, Z. Xu, and I. King, "Graph-Based Semi-Supervised Learning: A Comprehensive Review," IEEE Trans. Neural Networks Learn. Syst., vol. 34, no. 11, pp. 8174–8194, 2023, doi: 10.1109/TNNLS.2022.3155478.
21. A. dos Santos Ferreira, D. M. Freitas, G. G. da Silva, H. Pistori, and M. T. Folhes, "Unsupervised deep learning and semi-automatic data labeling in weed discrimination," Comput. Electron. Agric., vol. 165, no. July, p. 104963, 2019, doi: 10.1016/j.compag.2019.104963.
22. Y. Gu and G. Leroy, "Machine Mechanisms for Automatic Training Data Labeling for Machine Learning," 2019.
23. S. Gupta, M. Gupta, V. Varma, S. Pawar, N. Ramrakhiyani, and G. K. Palshikar, "Co-training for Extraction of Adverse Drug Reaction Mentions from Tweets," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2018, pp. 59–71. doi: 10.1007/978-3-319-76941-7_5.
24. H. K. Kanagala and V. V. Jaya Rama Krishnaiah, "A comparative study of K-Means, DBSCAN and OPTICS," 2016 Int. Conf. Comput. Commun. Informatics, ICCCI 2016, pp. 1–6, 2016, doi: 10.1109/ICCCI.2016.7479923.
25. X. X. Martin Ester, Hans-Peter Kriegel, Jiirg Sander, "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise," in KDD-96 Proceedings, 1996, pp. 226–231.
26. G. J. Monko and M. Kimura, "SS-DBSCAN: Epsilon Estimation with Stratified Sampling for Density-Based Spatial Clustering of Applications with Noise," Proc. - 2023 Int. Conf. Autom. Control Electron. Eng. CACEE 2023, pp. 72–76, 2023, doi: 10.1109/CACEE61121.2023.00023.
27. V. K. Dubey and A. K. Saxena, "Cosine similarity based filter technique for feature selection," ICCCCM 2016 - 2nd IEEE Int. Conf. Control Comput. Commun. Mater., no. Iccccm, pp. 1–6, 2017, doi: 10.1109/ICCCCM.2016.7918222.

28. S. Sohangir and D. Wang, "Improved sqrt-cosine similarity measurement," J. Big Data, vol. 4, no. 1, 2017, doi: 10.1186/s40537-017-0083-6.
29. V. Dogra and S. Verma, "Challenges and Opportunities in Labeling for Text Classification," India J., vol. 4370, no. 17, pp. 971–1260, 2019.
30. G. Monko and M. Kimura, "Enhanced SS-DBSCAN Clustering Algorithm for High-Dimensional Data," Data Sci., 2024.
31. B. C. Benato, C. Grosu, A. X. Falcão, and A. C. Telea, "Human-in-the-loop: Using classifier decision boundary maps to improve pseudo labels," Comput. Graph., vol. 124, no. August, p. 104062, 2024, doi: 10.1016/j.cag.2024.104062.
32. X. Wu, L. Xiao, Y. Sun, J. Zhang, T. Ma, and L. He, "A survey of human-in-the-loop for machine learning," Futur. Gener. Comput. Syst., vol. 135, pp. 364–381, 2022, doi: 10.1016/j.future.2022.05.014.
33. E. Mosqueira-Rey, E. Hernández-Pereira, D. Alonso-Ríos, J. Bobes-Bascarán, and Á. Fernández-Leal, Human-in-the-loop machine learning: a state of the art, vol. 56, no. 4. Springer Netherlands, 2023. doi: 10.1007/s10462-022-10246-w.
34. C. Chai and G. Li, "Human-in-the-loop Techniques in Machine Learning," Bull. IEEE Comput. Soc. Tech. Comm. Data Eng., pp. 37–52, 2020.

Authors

Gloriana J. Monko received a Master of Information and Communication Science and Engineering from The Nelson Mandela African Institution of Science and Technology (NM-AIST), and a Bachelor's degree in Informatics from Sokoine University of Agriculture (SUA). She is currently pursuing her PhD in Engineering Science from Shibaura Institute of Technology. Her research interests include data science, machine learning and natural language.

Prof. Masaomi Kimura received Ph.D. degree from The University of Tokyo. After his career of a system engineer in IBM, he started his career as a researcher at Shibaura Institute of Technology (SIT) in 2004. Now, he is a professor in the Department of Computer Science and Engineering in the Faculty of Engineering, and Department of Electrical, Electronics and Computer Engineering in the Graduate School of Science and Engineering, SIT. His research interests are in the areas of data science and data engineering, with a particular focus on data analysis as an application of artificial intelligence (machine learning), especially using deep learning.