New Topic Discovery Using LLM Analysis and Entropy Based Clustering of Short Texts

1st Allen Detmer
Department of Computer Science
University of Cincinnati
Cincinnati, OH USA
0009-0002-1659-8935

2nd Raj Bhatnagar
Department of Computer Science
University of Cincinnati
Cincinnati, OH USA
0000-0001-5222-6800

3rd Jillian Aurisano
Department of Computer Science
University of Cincinnati
Cincinnati, OH USA
0000-0003-4761-1741

***Abstract*—Our goal in this paper is to process short-text collections, labeled by their topics, and identify those groups of text items that may be better characterized by new labels and those text items that may be mislabeled. When a labeled text collection is clustered based on the syntactic and semantic contents of text items, it is expected that each cluster will also share the same label for text items contained in it. Our approach presented here embeds the short texts in a lower-dimensional space and then, using spatial label entropy as a guide, finds spatial clusters that are contiguous but have a wide diversity in their label assignments. We use LLMs to process such impure label clusters to discover new labels for them. We demonstrate promising results obtained by this approach for two different collections of short texts.**

***Index Terms*—Natural Language, Large Language Models (LLM), Text Clustering**

I. Introduction

Large language models are revolutionizing the influence of various AI and NLP tools in our daily lives. Many text processing systems include the important task of tagging individual text items by their topic-defining labels [21], [25], [34]. Typical tagging systems work with fixed sets of labels and assign one or more of these pre-selected labels to each text item. This process is particularly challenging for correctly classifying short-texts (e.g., Tweets, Chatbot inputs) from streaming or dynamically changing data sources. For such short text collections, new topics keep arising, requiring the static set of labels to be updated with the new emerging topic labels. In addition, label assignment systems can make mistakes and assign incorrect labels to some text items. In this paper, we present a solution that seeks to address these two problems with the help of LLMs. Our solution both detects text items that have potentially incorrect labels and proposes labels for sets of short texts that are labeled differently despite having textual similarities.

Label assignment systems that use classification models for natural language documents often suffer from overconfidence and lack of interpretability. When faced with novel or ambiguous types of text inputs, these models can produce misleadingly high confidence labels and scores [14], necessitating a labor intensive manual review to assess their performance and discover misclassifications. Existing approaches to model evaluation of Machine Learning systems to match user perception of conceptual categories do not effectively handle these cases of ambiguous input texts and output labels [18].

We present here a clustering-based method to identify those groups of text items whose members are very similar in terms of their text content such that they appear together in the embedding space, but whose preassigned topic labels have a large diversity. These texts may contain an emerging topic that was not considered in the initial set of labels or may contain ambiguous content that spans multiple categories that may be better reflected by refinement of the labels. For such impure clusters, we deploy LLMs to seek suggestions for new topic labels that better reflect the shared text contents of the cluster's items. The degree of diversity of labels within a cluster is captured by us in terms of the entropy of the labels occurring within a cluster. Our approach is derived from an adaptation of the basic DBSCAN algorithm, which is a robust approach that has been adapted to meet other problem-specific needs [2], [3], [22], [39]. Our novel approach enables the automatic discovery of emerging or ambiguous topics that the existing label taxonomy misses, providing a powerful way to refine taxonomies and reveal hidden structure in text collections. To our knowledge, our work is the first to combine entropy- based cluster impurity detection, density-based spatial clustering, and LLM-driven interpretation into a single framework for topic discovery in text collections.

A. Motivation

Text classification plays an important role in various domains, including sentiment analysis, topic modeling, and information retrieval. However, text classification is challenging due to a lack of clear separability among topic classes and large overlap areas among topics in classification spaces. One of the primary goals for such data situations is diagnosing misclassifications and understanding the reasoning behind model's decisions. Text classification models often struggle to adapt to the drifts that occur in the text content of their inputs and also in their topic areas, especially in the situation of streaming text items [24].

Another challenge in text classification is determining areas of interest for model updates. Given the dynamic nature of text-based data, particularly in domains such as news, social networks, and the scientific literature, identifying where

improvements should be focused is crucial. Without systematic methods to pinpoint evolving topics or shifting classification trends, models risk becoming obsolete [16]. The challenge extends to leveraging large language models (LLMs) effectively, as they provide powerful analysis capabilities, but require a structured approach to focus on relevant insights within vast text corpora. Methods for reducing and refining a corpus to concentrate on specific areas of interest remain an open research question [8].

In addition, the ability to identify new topics from textual data is of significant importance, particularly for applications such as market analysis, policy monitoring, and scientific discovery. Traditional topic modeling techniques such as Latent Dirichlet Allocation (LDA) and neural-based models often fail to capture novel topics effectively without retraining on updated data [4]. Thus, investigating whether LLMs and advanced clustering techniques, can enhance new topic detection and classification remains a pressing concern.

This paper aims to address these challenges by exploring methods for improving text classification, identifying emerging topics, and focusing LLM-based analysis on relevant segments of large text corpora. Using advances in machine learning, natural language processing (NLP), and information retrieval, we seek to develop more effective strategies to refine classification models and gain deeper insights into textual data.

Modern text embeddings (e.g. transformers [35]) reside in very high-dimensional spaces with complex nonlinear structures. Traditional clustering algorithms often struggle to effectively capture the complex, nonlinear structures and varying densities inherent in high-dimensional embedding spaces. Traditional clustering struggles with the "curse of dimensionality" [17] and the inherent complexity of these spaces.

B. Contribution

In this paper, we contribute a novel framework for identifying emerging topics and mislabeled items in short-text collections.

We develop a novel clustering approach which shifts the focus from density of points to diversity of labels in a spatial region. This entropy-driven clustering in the embedding space offers a powerful approach for uncovering hidden structures and patterns within complex data. By incorporating spatial entropy, the method highlights "impure regions" that coincide with classification uncertainty, mislabeled text, or new emerging topics. We leverage Large Language Models to analyze impure clusters to produce summaries, identify mislabeled text items, and suggest new topic labels that capture the coherence among the text items in the cluster. By unifying unsupervised clustering and LLM-driven analysis, this work helps produce semantically meaningful clusters with suitable topic labels for all clusters.

II. Related Work

A. Using LLMs for Label Prediction

Large language models (LLMs) [27] provide human-like natural language capabilities. A recent study finds that LLMs outperform traditional methods like TF-IDF and Latent Dirichlet Allocation (LDA) in capturing meaningful and contextually relevant topics for sets of text items [10]. The key challenge with the LLM compared to traditional methods is the computational cost and bias. LLMs can be biased, based on the training data used, which can potentially influence the inference of topics for sets of text items [7].

B. Determining Need to assign New Labels

A critical aspect of maintaining high-quality labeled datasets is determining when existing taxonomies of labels fail to adequately describe the content and new labels must be discovered. Prior research has explored various strategies for detecting such needs of intra-class semantic diversity, inter-class confusion, and cluster purity [30].

One approach is to apply a novel clustering technique to high-dimensional embeddings of text items and identify clusters with low label purity, indicating potential for new category creation [36]. More recent work leverages representation learning and unsupervised anomaly detection to uncover content not well-represented by existing labels [6], [31].

In natural language processing, topic modeling [29] and embedding-based semantic clustering [28] have been used to reveal latent topics that may warrant new labels. Several studies integrate large language models (LLMs) into this process to provide human-interpretable descriptions of potential new categories [5], [20].

These approaches highlight the value of combining quantitative measures (e.g. spatial entropy, purity scores) with qualitative assessments (e.g., LLM-generated summaries) to make informed decisions about when and how to expand a label set.

C. Entropy Measure for Impurity Detection

Entropy has long been used as a statistical measure of uncertainty in information theory [32]. In the context of classification and clustering, entropy quantifies the diversity of class labels within a set of samples. High entropy indicates label heterogeneity, which can signal impurity in a cluster or local neighborhood of points. In clustering, label entropy has been applied as a post hoc quality metric to evaluate the purity of discovered clusters [30]. In anomaly detection, high-entropy neighborhoods have been leveraged to flag potential outliers or emerging patterns [6], [31]. Our work builds upon this line of research by computing *spatial label entropy* directly in the sentence embedding space. This allows us to isolate dense but label-diverse regions, which we term *impure regions*, for further semantic analysis using large language models.

D. Novel Ideas of Our Approach

While prior work has explored label impurity detection, cluster analysis, and topic discovery separately, our approach integrates these components into a unified pipeline. Specifically, we combine:

1) **Spatial Label Entropy in Embedding Space:** Building on entropy-based impurity metrics [30], we compute local

entropy directly in high-dimensional sentence embeddings [28] to detect dense yet label-diverse regions.
2) **DBSCAN Adaptation for Impure Clusters:** Unlike traditional density-based clustering [12], which groups all dense regions, our method isolates only those with high semantic impurity, ensuring that downstream analysis focuses on the most informative areas.
3) **LLM-Driven Semantic Analysis:** Leveraging recent advances in large language models [13], we prompt the model with representative examples from impure clusters to obtain interpretable summaries, re-labeling suggestions, and candidate new topics.

To our knowledge, no prior work has combined entropy-based cluster impurity detection, density-based spatial clustering, and LLM-driven interpretation into a single framework. This framework sets the stage for iterative refinement of the set of labels. This integration enables both quantitative identification of problematic regions and qualitative generation of actionable label refinements.

III. Proposed Approach

The main flow of our methodology can be summarized as the following sequence of steps: (i) Perform *SentenceTransformer* embedding of all short texts into a lower-dimensional space, (ii) Compute entropy of spatial distribution of document labels around each point in the embedding space, (iii) identify those contiguous regions of embedding space that have a high concentration of high entropy points, and call them *impure regions*, (iv) use LLMs to analyze text contents of documents in impure regions to identify new possible topics and misclassified documents.

A. Embedding Text Items in a Lower Dimensional Space

The first step in our clustering pipeline is to convert each short-text document into a fixed-size vector representation. We employ the `SentenceTransformer` framework with the pre-trained model `all-MiniLM-L6-v2` [11]. We selected `all-MiniLM-L6-v2` over other models, due to its suitability for short text encoding [19], [33], [37]. It is known to works well with Euclidean or cosine distances [19]. Other embedding models could be utilized for this step, provided they met these criteria.

We have developed an adaptation of the DBSCAN algorithm [12] to find clusters of documents that have similar text and semantics but different assigned labels. This requires that the embedding space be such that semantically similar items are positioned close together under a given distance metric. The embedding model used by `all-MiniLM-L6-v2` [11] produces embeddings that preserve semantic similarity, which:

- Improves separation between text items with distinct semantics.
- Creates dense neighborhoods for semantically similar, high-entropy points, enabling our entropy-based DBSCAN extension to identify dense regions with diverse labels.

The 384-dimensional output provides a computationally efficient yet semantically rich representation, facilitating large-scale clustering.

B. Spatial Entropy of Label Distributions

We seek to identify regions of the embedding space where documents are semantically close yet have highly diverse preassigned labels. Such regions are often indicative of:

1) **Potential Misclassifications:** Points where the assigned label does not align with the majority of semantically similar texts.
2) **Emerging Topics:** Novel concepts not adequately represented in the existing set of labels.
3) **Ambiguous Content:** Texts that genuinely span multiple topics and require human review or label refinement.

Entropy, in the information-theoretic sense, quantifies the diversity (or uncertainty) of labels within a neighborhood in the embedding space. By focusing on *spatial entropy*, we explicitly capture the variation in label distribution around each point, allowing the algorithm to target regions where label assignments are least consistent.

Our contribution focuses on using spatial entropy. Let $\mathcal{E} = \{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n\}$ be the set of normalized embeddings for n documents, and let $L = \{\ell_1, \ell_2, \ldots, \ell_n\}$ be their corresponding labels. For a given point $\mathbf{e}_i$:

1) Identify the set of neighbors $\mathcal{N}_\epsilon(\mathbf{e}_i)$ within radius ϵ according to a chosen distance metric $d(\cdot, \cdot)$.
2) Extract the label set $L_\epsilon(\mathbf{e}_i)$ for these neighbors.
3) Compute the Shannon entropy:

$$H(\mathbf{e}_i) = -\sum_{c \in C} p(c) \log_2 p(c)$$

where C is the set of unique labels in $L_\epsilon(\mathbf{e}_i)$ and $p(c)$ is the empirical probability of label c in that neighborhood.

The result $H(\mathbf{e}_i)$ measures the local diversity of labels around $\mathbf{e}_i$ in embedding space.

The spatial entropy computation depends on three key parameters:

- **Neighborhood Radius (ϵ):** Defines the spatial extent of the local neighborhood.
- **Minimum Points for Entropy Calculation (min_points):** The minimum number of neighbors required to compute a meaningful entropy value; otherwise $H(\mathbf{e}_i)$ is set to zero. The results presented in this paper use a minimum value of eight.
- **Distance Metric (metric):** Determines neighborhood membership (e.g., cosine or Euclidean). The choice affects both neighborhood composition and the resulting entropy distribution. For our test results we have used the Euclidean distance.

Spatial entropy is sensitive to its parameter settings:

- *ϵ:* Too small an ϵ may produce sparse neighborhoods, leading to unstable entropy estimates dominated by noise. Too large an ϵ may merge semantically distinct regions,

reducing entropy contrast and obscuring high-uncertainty zones.

- *min_points:* If set too low, entropy values may be computed from insufficient samples, inflating variability. If set too high, genuinely high-entropy regions in sparse areas may be ignored.
- *metric:* A poor choice can distort neighborhood boundaries. For example, Euclidean distance in unnormalized spaces may overemphasize vector magnitude, while cosine distance better reflects semantic similarity in normalized embedding spaces.

Careful calibration of these parameters is essential for balancing sensitivity to label inconsistency against robustness to noise. In practice, parameter sweeps combined with visual inspection of high-entropy regions provide an effective tuning strategy. We have tested minor variations and the results remained stable.

C. Clustering Impure Regions in Embedding Space

In our approach, we use our developed adaptation of the DBSCAN algorithm [12] to identify *impure regions* in the embedding space - areas where documents are spatially close but have highly diverse assigned labels. Such regions are of particular interest because they often indicate:

1) **Mislabeled Data:** Documents semantically aligned but incorrectly labeled.
2) **Emerging Topics:** Novel concepts not represented in the current taxonomy.
3) **Ambiguous Content:** Items that legitimately span multiple topics, useful for refining label definitions.

Given a set of normalized embeddings $\mathcal{E} = \{\mathbf{e}_1, \ldots, \mathbf{e}_n\}$ and corresponding labels $L = \{\ell_1, \ldots, \ell_n\}$, the DBSCAN-based impure region clustering proceeds as follows:

1) **Compute Spatial Entropy:** For each $\mathbf{e}_i$, calculate the local label entropy $H(\mathbf{e}_i)$ over its ϵ-neighborhood.
2) **Mark High-Entropy Points:** Classify points as high-entropy if $H(\mathbf{e}_i) > \tau$, where τ is a predefined threshold of .55 to find inherently fuzzy or overlapping classes.
3) **Identify Core Points:** Mark a high-entropy point as a *core point* if at least `min_samples_high` of its neighbors are also high-entropy. For our test results, we used 55% of the points within the ϵ hypersphere.
4) **Cluster Formation:** Starting from each unassigned core point, grow a cluster by recursively including neighboring high-entropy points and their reachable neighbors (standard DBSCAN expansion).
5) **Assign Cluster Labels:** All points in the same connected high-entropy region receive the same cluster ID; unassigned points are labeled -1 (noise).

Traditional DBSCAN forms clusters solely based on spatial *density* — measured by number by points within an ϵ-radius of a point. Our adaptation replaces pure density with a measure of diversity of labels for text items with the ϵ-radius.

High-entropy regions are disproportionately informative for evaluation and model refinement.

- They often contain points near decision boundaries in classifier space.
- They reveal inconsistencies between human labeling and semantic similarity.
- They can identify emerging or merged topics that static taxonomies fail to capture.

While our algorithm operates in a 384-dimensional embedding space, Figure 1 illustrates the concept in a 2-D projection for easy visualization. Here, each point represents a document embedding, color indicates its assigned high or low entropy rating, and dotted outline mark high-entropy clusters of high-entropy points detected by our algorithm.

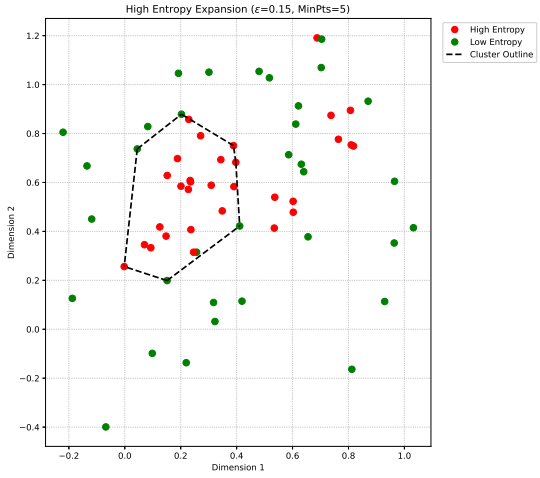

Fig. 1. **Cluster of high entropy points** This figure illustrates an example of a high-entropy cluster identified by our method. Individual points are colored (e.g., red for 'H' for high entropy, green for 'L' for low entropy) according to whether their entropy value exceeds a predefined threshold.

D. LLMs' Processing of Impure Clusters

Once impure regions, dense clusters in embedding space with high label diversity are identified the next step is to interpret and act upon them. Traditional post-processing approaches typically rely on statistical summaries, manual inspection, or rule-based relabeling [9], [15], [23]. These methods can be labor-intensive and limited in their ability to generalize. Large Language Models (LLMs) offer a promising alternative: by leveraging their extensive pretraining on diverse textual data, they can:

1) Provide nuanced semantic interpretations of the mixed-content clusters.
2) Suggest plausible re-labelings or subtopic splits without exhaustive manual coding.
3) Identify new topic candidates or reveal ambiguous phrasing that leads to label confusion.

This makes LLMs particularly effective for rapidly extracting actionable insights from high-entropy clusters.

The process of applying the LLM `GPT-4o-mini` [26] LLM to impure regions proceeds as follows:

1) **Cluster Selection:** Choose impure clusters detected by the entropy-based DBSCAN method.

2) **Data Extraction:** Retrieve the original short texts and their preassigned labels for all points in the cluster.
3) **Prompt Construction:** Create a structured prompt for the LLM that includes:
   - Texts and labels from the cluster.
   - Instructions to analyze semantic themes, label consistency, and potential refinements.
   - Optional constraints such as limiting suggestions to a fixed label vocabulary.
4) **LLM Processing:** Pass the prompt to the LLM for analysis.
5) **Output Parsing:** Extract relevant sections from the LLM's output for downstream tasks (e.g., relabeling, taxonomy update).

The quality of LLM-based analysis depends on LLM parameters:

- **Number of Examples in Prompt ($n_{\mathbf{ex}}$):** Too few examples may yield generic analysis; too many can exceed token limits or dilute focus.
- **Prompt Specificity:** Overly broad prompts may produce unfocused results; overly constrained prompts may limit discovery of novel topics.
- **Temperature:** Higher temperature can encourage creative label suggestions, but may reduce reproducibility.
- **Model Size:** Larger LLMs tend to capture more nuanced semantics, but incur higher computational cost.

LLM outputs are structured into distinct parts, each serving a specific purpose:

1) **Cluster Summary:** Concise semantic description of the dominant and minority themes within the cluster.
2) **Label Consistency Analysis:** Quantitative or qualitative assessment of how well existing labels align with content.
3) **Suggested Re-labelings:** Revised label assignments for each text, either from the original label set or expanded with new topics.
4) **Emergent Topic Candidates:** Suggested new categories that better capture underrepresented themes.

Below is an excerpt of LLM output for an impure cluster containing "sports" and "world" and "Sci/Tech" items:

```
Upon analyzing the 'text' column, several
entries labeled as 'Sports' contain content
that primarily discusses technology or
security aspects related to the Olympics,
which may indicate misclassification. The
identifiers of potentially misclassified
rows are: [197, 328, 372, 399, 409, 2350].

Suggested new topic labels in order:
1. Technology
2. Security
3. Event Management
```

This type of output is valuable because:

- **Cluster Summary** accelerates human understanding of the thematic scope.
- **Label Consistency Analysis** identifies systematic mislabeling patterns.
- **Suggested Re-labelings** provide actionable corrections to improve dataset quality.
- **Emergent Topic Candidates** support taxonomy evolution and model retraining.

By structuring the prompt to request these specific outputs, we ensure that the LLM delivers insights directly usable for quality assurance, taxonomy refinement, and model improvement.

IV. Experiments and Analysis

Overview of the evaluation: In this section, we evaluate our approach. We selected two suitable datasets with emerging topics. Then we used our method to identify impure clusters, and describe their characteristics. We present a set of example impure clusters, the texts they contain and the proposed labels, to illustrate the application of our method. We did not validate through comparison against alternative clustering methods, because to our knowledge no other algorithm creates clusters driven by entropy.

A. Evaluation Dataset Selection and Justification

The goal is to evaluate the effectiveness of our methodology in collections of general-purpose short-texts. To do this we selected two complementary datasets: (1) the *News Topic Classification* dataset derived from AG News, and (2) the *Twitter Financial News* dataset[1]. These datasets are well-suited for our purpose due to their topical diversity, balanced label distributions, and suitable sizes.

Tweets and news headlines serve distinct but complementary roles in this evaluation. Tweets represent short user-generated content with informal language, real-time reactions, and varying semantics, making them ideal for testing the robustness of the algorithm to noise and ambiguity. In contrast, news headlines are concise, curated summaries of structured events, offering cleaner semantic signals and more formal language. This dichotomy allows us to evaluate clustering performance across both high-noise and low-noise text environments, supporting a comprehensive analysis of entropy-based embedding learning.

Dataset 1: News Topic Classification (AG News):

- **Source:** Derived from the AG corpus of news articles, as introduced by Zhang et al. [38]
- **Number of Articles:** Over 1 million
- **Number of Category Labels:** 4
- **Topic Labels:** *World*, *Sports*, *Business*, *Sci / Tech*
- **Example Headlines:**
  - *World:* "Japan urges North Korea to cancel missile launch"
  - *Sports:* "Federer wins fifth Wimbledon title in epic final"
  - *Business:* "Oil prices rise amid OPEC output cut speculation"

[1]https://www.kaggle.com/datasets/sulphatet/twitter-financial-news

  - *Sci / Tech:* "NASA's Mars rover sends new panoramic images"
- **Label Clarity:** This dataset has clearly defined, mutually exclusive topic labels with relatively balanced distribution, offering an effective testbed for clustering algorithms on structured and clean data.
- **Selection Method for Experiments:** To ensure a balanced representation across categories, we selected the first 1000 short texts for each of the following four labels: Business, Sci/Tech, Sports, and World. This resulted in a total dataset size of 4000 short texts.

Dataset 2: Twitter Financial News:

- **Source:** Used in *InstructNet* [1]
- **Number of Category Labels:** 20
- **Topic Labels:** *Analyst Update*, *Fed — Central Banks*, *Company — Product News*, *Treasuries — Corporate Debt*, *Dividend*, *Earnings*, *Energy — Oil*, *Financials*, *Currencies*, *General News — Opinion*, *Gold — Metals — Materials*, *IPO*, *Legal — Regulation*, *M&A — Investments*, *Macro*, *Markets*, *Politics*, *Personnel Change*, *Stock Commentary*, *Stock Movement*
- **Example Tweets:**
  - *Earnings:* "Apple Q4 earnings beat expectations with record iPhone sales"
  - *Fed — Central Banks:* "Powell signals possible pause in interest rate hikes"
  - *Stock Movement:* "TSLA up 7% after delivery numbers beat estimates"
- **Label Clarity:** Labels are more granular and sometimes overlapping, introducing ambiguity. This makes the dataset ideal for evaluating clustering in noisy, real-world scenarios.
- **Selection Method for Experiments:** This dataset originally contained 16,990 records. For this study, we reduced the dataset to focus on specific topics, resulting in a total of 6,186 records across several labels. The distribution of labels in this reduced dataset is as follows: Stock - 3344, Trading - 471, Withdraw - 321, Personnel - 495, Politics - 985, Markets - 501, Metals - 69.

Comparative Utility and Sensitivity Analysis: The combined use of these two datasets enables a comprehensive assessment of clustering performance across different text domains. The AG News dataset serves as a clean, structured benchmark with clearly separated topics, while the Twitter Financial News dataset introduces complexity through informal language and topic overlap. To evaluate the stability and reliability of the clusters, a sensitivity analysis was performed. This included:

- Varying the number of clusters (k) and observing changes in cluster purity.
- Measuring cluster stability using different parameters.
- With slight changes in entropy threshold and ϵ, the core clusters remained intact with slight variations in size of less than 5%.

Clustering in the AG News dataset appears robust to minor perturbations and variations in k, with frequent words and cluster structures remaining stable throughout the runs. Similarly, the Twitter Financial News dataset—despite its informal syntax and semantic overlap—showed less than a 5% change in frequent terms and cluster composition, indicating that its core clustering structure also remained consistent. This stability across both clean (AG News) and noisy (Twitter) text reinforces the value of evaluating our model in diverse environments to ensure effectiveness under realistic conditions.

B. Results and Analysis

This section presents the experimental results of applying our methodology to assign new labels to documents in impure clusters.

1) Clustering Results on News Topic Dataset: For the News Topic Dataset, the clustering step of our methodology yielded the following results:

2) Parameters and Initial Filtering (News Topic):

- The mean distance to the nearest 8^{th} neighbor for all data points in the embedding space was 0.9588. Subsequently, this value was used as the ϵ parameter for all steps that involve neighborhood definitions for this data set.
- **Entropy per Data Point:** All those data points whose ϵ-neighborhood contained at least 8 data points were considered as dense enough for entropy computation. Out of the 4000 total points, **779 points** met this minimum neighborhood size criterion.
- **High Entropy Point Identification:** An entropy ***threshold_high*** of 0.5 was applied to distinguish high-entropy points. **575 points** were marked as having an entropy value greater than 0.5.

3) Core Point Identification and Cluster Formation (News Topic):

- **Core Point Determination:** A point was marked as a core point if at least **50%** of its neighbors within its ϵ hypersphere were also marked as high entropy points. A total of **527 points** were marked as core points out of a total **4000 points**.
- **Cluster Generation:** Finally, these 527 core points were used as seeds to form clusters. Starting from a core point as a seed, all its neighbors within its ϵ-neighborhood were included in the cluster, and the cluster was then recursively expanded by including all points within ϵ-neighborhood of each core point already in the cluster. All clusters containing at least eight points were considered reasonable for subsequent attention. The algorithm successfully formed **five distinct clusters**. The distribution of points among these clusters is detailed in Table I.

TABLE I
DISTRIBUTION OF POINTS PER FORMED CLUSTER

ID(#Pts)	Entropy	Labels
1 (371)	1.0185	Business(284), Sci/Tech(48), World(39)
2 (46)	1.3869	Business(21), World(20), Sci/Tech(5)
3 (307)	0.6377	Sports(267), World(34), Sci/Tech(6)
4 (46)	0.8987	World(35), Business(10), Sci/Tech(1)
5 (15)	0.7219	Business(12), World(3)

C. Clustering Results on Twitter Financial News Dataset

The label-entropy based clustering algorithm was subsequently applied to the Twitter Financial News Dataset.

1) Parameters and Initial Filtering (Twitter Financial News):

- **Mean Nearest Neighbor Distance:** The mean of the distances to the 8th nearest neighbors of all data points is 0.8421. This value served as ϵ for the Twitter Financial News dataset.
- **Entropy per Data Point:** The entropy for labels in each point's neighborhood was computed using the calculated ϵ (0.8421) and computing only for those points that had a minimum of eight points in their neighborhoods. **564 points** qualified for the computation. ϵ, enabling entropy calculation.
- **High Entropy Points: 328 points** were identified as having an entropy value greater than 0.55.

2) Core Point Identification and Cluster Formation (Twitter Financial News):

- **Core Point Determination:** A point was marked as a core point if at least **50%** of its neighbors within its ϵ hypersphere were also marked as high entropy points. A total of **232 points** were marked as core points.
- **Cluster Generation:** Using these 232 core points as seeds, our algorithm identified **10 distinct impure clusters**. The distribution of points among these clusters is presented in Table II.

TABLE II
DISTRIBUTION OF POINTS PER FORMED CLUSTER

ID(#Pts)	Entropy	Labels
1 (390)	1.3554	stock(216), markets(145), withdraw(16), metals(13)
2 (70)	1.1157	stock(50), trading(16), markets(3), personnel(1)
3 (24)	0.9183	stock(16), trading(8)
4 (15)	0.9710	stock(9), trading(6)
5 (10)	0.7219	stock(8), trading(2)
6 (8)	0.9544	stock(5), markets(3)
7 (44)	1.5033	politics(21), withdraw(14), markets(9)
8 (8)	1.2988	markets(5), withdraw(2), stock(1)
9 (18)	0.9183	personnel(12), stock(6)
10 (16)	0.8960	politics(11), stock(5)

D. Analysis and Discussion

For the News Topic Dataset, the algorithm identified eight distinct clusters from 4000 embedded data points. The resulting clusters varied in size, suggesting the algorithm's capacity to identify both broader thematic groups (e.g., the large clusters 1 and 3) and smaller, more specific sub-topics or unique variations. The application to the Twitter Financial News Dataset yielded 12 distinct clusters.

Overall, these initial results are promising for an algorithm designed to reveal spatially localized but semantically diverse groupings of documents. The approach demonstrates a capacity to capture both general topics and more granular subthemes, depending on the characteristics of the input data and appropriate parameter settings.

E. LLM Analysis and Review of Impure Clusters

We perform analysis of each impure cluster by an LLM system to seek guidance about (i) topic labels that best capture the coherence of semantics in the text items included in a cluster, including possibly new labels that are outside the existing set of topic labels, and (ii) Text items in a cluster that are mislabeled.

We show here the LLM responses obtained for the representative impure clusters from both the News Topic Dataset and the Twitter Financial News Dataset. This analysis demonstrates the effectiveness of the label entropy-based clustering technique by highlighting content coherence, misclassifications insights, and key topics within each cluster. The LLM's ability to consistently identify thematic relevance and point out subtle misclassifications within the clusters underscores the utility of this new clustering algorithm in creating semantically meaningful groupings.

To test the robustness of the new label suggestions by the LLM for documents in a cluster, we repeated the prompt ten times. The LLM returned a list of three suggested topics, in order from best, second best to third best. The first topic (best) was assigned three merit points, the second topic (second best) two merit points and the third one (third best) was assigned one merit point. We added up the merit points earned by each new topic suggestions across the ten different runs to compute the most promising topic suggestion. Robustness of our topic suggestion methodology is demonstrated if topics are suggested consistently.

1) News Topic Dataset:

a) Cluster 1: Business-Focused News with IPO and Economic Trends:

- **LLM Response:** *"Upon analyzing the 'text' column, several entries appear to be misclassified. For instance, texts discussing Google's IPO and its implications on the market are classified under 'Sci/Tech' but could also fit under 'Business' due to their financial context. Similarly, some entries labeled 'World' discuss economic impacts, which are more aligned with 'Business'. Identifiers of misclassified rows (texts) include: 209 (Sci/Tech) - 215 (Sci/Tech) 220 (Sci/Tech) - 393 (Sci/Tech) - 397 (Sci/Tech). Suggested new topic labels in order: 1. Finance 2. Technology 3. Market Trends"*
- **Top Word Frequencies:** google (336), oil (226), prices (220), reuters (202), new (141), us (139), ipo (139), stocks (135), inc (122), offering (119)
- **Top N-grams:** oil price (136), new york (92), initial public (91), public offering (89), initial public offering (82), york reuters (58), new york reuters (57), google ipo (49), consumer price (44), price range (38)
- **Robustness of LLM Suggested Topics:** topic (priority), topic 2 (priority), topic 3 (priority)

The cluster's strong focus is on business-related terms like "oil," "stocks," "Google," and "ipo". The LLM's identification of "Sci/Tech" labeled articles as misclassified for this business-centric cluster demonstrates our methodology's effectiveness in grouping semantically similar content, regardless of initial

TABLE III
LLM SUGGESTED TOPIC LABELS AND THEIR MERIT POINTS

Suggested Label	Priority Points
Technology	20
Finance	17
IPO/Finance	6
Media	4
Seven other topics	1-3

broad labels. It also demonstrates that the new labels suggested by the LLM, in this case "technology" and "Finance", are very relevant for the terms and the N-grams occuriong in the cluster. The stability of the suggested labels across multiple runs of the prompt is also demonstrated by Table III

b) Cluster 3: Olympic Events and Sports News:

- **LLM Response:** *"Upon analyzing the 'text' column, several entries labeled as 'Sports' contain content that primarily discusses technology or security aspects related to the Olympics, which may indicate misclassification. The identifiers of potentially misclassified rows are: [197, 328, 372, 399, 409, 2350]. Suggested new topic labels in order: 1. Technology 2. Security 3. Event Management"*.
- **Top Word Frequencies:** athens (299), olympic (193), gold (183), olympics (137), greece (121), phelps (114), games (105), us (96), medal (95), reuters (89)
- **Top N-grams:** gold medal (84), athens reuters (79), athens greece (65), michael phelps (57), athens olympics (43), united state (39), greek sprinter (31), olympic game (29), athens game (27), paul hamm (26)

TABLE IV
LLM SUGGESTED TOPIC LABELS AND THEIR MERIT POINTS

Suggested Label	Priority Points
Sports	21
Technology	17
Sports Events	6
World Events	5
Five other topics	1-4

This cluster, characterized by "athens," "olympic," "gold," and "phelps", shows excellent thematic coherence. The LLM's explicit recommendation to reclassify 'World' or 'Sci/Tech' entries to 'Sports' within this cluster confirms that our methodology effectively groups articles by underlying sports themes, even when their original assigned labels are very different. The topics suggested across multiple runs of the prompt include "Sports" and "Technology" at the top with much higher merit compared to all other suggestions. These two top suggestions are very much in sync with the most frequent words and the N-grams occurring in the cluster's documents.

2) Twitter Financial News Dataset:

a) Cluster 1: General Market Trends and Economic Indicators:

- **LLM Response:** *"Upon analyzing the 'text' column, several entries appear misclassified, particularly those discussing broader economic trends or specific company performances that do not directly relate to stocks or metals. Misclassified identifiers include: 948, 949, 950, ..(more IDs) ...5905, 6067, 6068, 6070]. Suggested new topic labels: 1. 'economy', 2. 'financial news', 3. 'corporate earnings'."*
- ***Top Word Frequencies:*** *spy (126), qqq (93), stocks (93), market (76), stockmarket (74), economy (71), dia (62), spx (52), markets (49), compq (45)*
- ***Top N-grams:*** *spy qqq (76), http economy (57), qqq dia (54), spy qqq dia (54), economy stockmarket (51), http economy stockmarket (49), dia spx (43), dia spx compq (43), qqq dia spx (43), qqq dia spx compq (43)*

TABLE V
LLM SUGGESTED TOPIC LABELS AND THEIR MERIT POINTS

Suggested Label	Priority Points
Economic Trends	18
Economy	12
Market Analysis	8
Finance	5
Company Performance	5
Six other topics	1-4

This cluster, rich with terms "spy," "stocks," "qqq," and "market," shows strong thematic coherence around broad market and economic discussions. The LLM's recommendations for 'economic trends', 'economy' and 'market analysis' are broader terms and align well with high-frequency terms and N-grams such as 'economy stock market' and 'spy qqq'. The stability of these labels across runs (Table VI) further demonstrates that the suggested labels better capture the broader underlying economic content of this cluster instead of more specific preassigned labels such as stocks, metals, or trading.

b) Cluster 2: Elon Musk, Twitter Acquisition, and Legal Disputes:

- **LLM Response:** *"Upon analyzing the 'text' column, several entries appear to be misclassified. For instance, tweets discussing legal actions and implications of Elon Musk's acquisition of Twitter should be classified under 'legal' rather than 'trading' or 'stock.' The following identifiers are likely misclassified: 1864, 2926, 5043, 5083, 5085, 5086, 5191, 5200, 5206, 5234, 5565. Suggested new topic labels in order: 1. Legal 2. Corporate Governance 3. Mergers and Acquisitions"*
- **Top Word Frequencies:** twitter (60), musk (52), elon (37), twtr (26), deal (24), billion (14), tesla (19), trading (12), buy (11), business (11)
- **Top N-grams:** elon musk (36), twtr twitter (8), http http (7), http trading (6), twitter share (6), buy twitter (5), deal http (5), twitter http (5), twtr pre (5), billion deal (4)

TABLE VI
LLM SUGGESTED TOPIC LABELS AND THEIR MERIT POINTS

Suggested Label	Priority Points
Acquisition News	22
Stock Market Reactions	12
Legal Proceedings	11
Corporate Governance	4
Five other topics	1-3

The clear emphasis on "twitter," "musk," and "elon" highlights a tight narrative around the Twitter acquisition. The LLM's suggestions of new topic names "Acquisition News" and "Legal Proceedings" are consistent with N-grams like "elon musk" and "twitter share." As in prior clusters, repeated emergence of these labels across runs (Table VII) demonstrates that the clusters did need new labels and LLM was able to suggest broader category names instead of the preassigned more specific topics.

c) Cluster 3: Mergers, Acquisitions, and Stock Performance (Unity/Ironsource):

- **LLM Response:** *"The 'text' column predominantly discusses financial transactions, mergers, and stock performance, primarily related to Unity and Ironsource. However, some entries labeled as 'trading' may be more appropriately classified under 'stock' due to their focus on stock performance and market reactions. Misclassified identifiers include: 840, 847, 849. Suggested new topic labels in order: 1. Mergers & Acquisitions 2. Stock Performance 3. Market Analysis"*
- **Top Word Frequencies:** unity (16), ironsource (14), u (14), merger (8), software (6), business (6), agreement (5), shares (5), finance (4), investing (4)
- **Top N-grams:** unity software (6), ironsource http (5), merger agreement (5), agreement ironsource (4), agreement ironsource http (4), merger agreement ironsource (4), merger agreement ironsource http (4), announces merger (2), announces merger agreement (2), announces merger agreement ironsource (2)

TABLE VII
FREQUENCY OF LLM SUGGESTED LABELS FOR TWITTER FINANCIAL NEWS CLUSTER 3 (10 RUNS)

Suggested Label	**Priority Points**
Mergers & Acquisitions	28
Stock Performance	20
Investment Analysis	4
Five other topics	1-3

Frequent terms such as 'unity', 'ironsource', and 'merger agreement' clearly point to merger and acquisition activity. The LLM's suggested change of topic from preassigned label "trading" to new labels "Mergers & Acquisitions" and "Stock Performance" is supported by recurring N-grams such as "unity software" and "merger agreement ironsource." The dominance of these new label suggestions across multiple runs (Table VIII) underscores the cluster's precision along these new broader topic areas.

These examples demonstrate the consistent ability of the LLM to process impure text clusters and suggest new topic labels better anchored in the recurrent high-frequency terms and N-grams of the documents. It is also shown that the suggested new topics are stable across multiple LLM runs and provide labels that better capture the thematic coherence within these clusters. By also pinpointing precise misclassifications and suggesting more granular event-specific categories, the LLM demonstrates that the label entropy-based clustering methodology not only grouping semantically related content, but also refining it into more accurate and contextually relevant labels. This establishes a robust foundation for improved content categorization and analysis.

V. DISCUSSION

Our approach demonstrates promising capabilities for discovering new topics and addressing mislabeled items in short-text collections. We also demonstrate potential integration points for language models in text-classification pipelines, which is an area of considerable interest given the expansion in LLM capabilities and application areas. Our impure cluster identification technique suggests directions for targeting LLM tasks to relevant sets of items to improve its performance on tasks. Our work builds primarily on a novel adaptation of density-based clustering methods and LLMs capability to identify shared and coherent themes running across a set of short texts.

Limitations and Future Work There are several limitations to our approach that could be addressed in future work. First, our approach focuses on short-texts, such as tweets, and does not consider longer text documents, such as articles. We expect our general approach - embedding, followed by identifying impure clusters and utilizing an LLM for relabeling - to extend to collections of longer text documents. However, with token limits on LLM inputs, the specific pipeline would need to be adapted to accommodate longer text items. In addition, longer texts may present computational challenges that would need to be addressed. Second, we do not consider how our approach would scale to more items. We target a common use case involving collections with thousands of items, which could be gathered by selection for refinement and testing of existing classification models. However, there may be cases where pipelines need to scale to larger item collections. The scalability of this technique would require additional testing. Third, we focus on datasets with a single label for each text and do not consider multi-class datasets. Multi-class labeling presents additional complexity to the identification of impure clusters. We intend to address this challenge in future work. Finally, our approach sets the stage for iterative refinement of label taxonomies to improve classification performance and integrate evolving themes. However, we do not fully develop such a pipeline here. Such an approach would require additional steps in our pipeline, and may need to involve user input to steer label refinement and model retraining.

VI. CONCLUSION

We introduce a novel entropy-based clustering algorithm that operates in the embedding space of text, where clusters are defined by continuous regions of semantic similarity. Instead of relying solely on density, our method highlights areas with diverse labels, exposing "impure regions" that signal uncertainty or emerging topics. Within clusters, semantic coherence is preserved, while entropy identifies boundaries where mislabeling or overlap occurs. LLMs are then applied to interpret these regions, generating meaningful summaries, relabeling suggestions, and new category proposals. Experiments

on clean and noisy datasets confirmed that this combination of entropy-driven clustering and LLM reasoning reliably uncovers ambiguities, refines taxonomies, and strengthens the adaptability of text classification systems.

REFERENCES

[1] T. T. Aurpa, M. S. Ahmed, M. M. Rahman, and M. G. Moazzam. Instructnet: A novel approach for multi-label instruction classification through advanced deep learning. *Plos one*, 19(10):e0311161, 2024.

[2] M. Bessrour, Z. Elouedi, and E. Lefèvre. E-dbscan: An evidential version of the dbscan method. In *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*, pp. 3073–3080, 2020. https://doi.org/10.1109/SSCI47803.2020.9308578 doi: 10.1109/SSCI47803.2020.9308578

[3] D. Birant and A. Kut. St-dbscan: An algorithm for clustering spatial–temporal data. *Data & knowledge engineering*, 60(1):208–221, 2007.

[4] D. M. Blei. Probabilistic topic models. *Communications of the ACM*, 55(4):77–84, 2012.

[5] Y. Chae and T. Davidson. Large language models for text classification: From zero-shot learning to fine-tuning. *Open Science Foundation*, 10, 2023.

[6] R. Chalapathy and S. Chawla. Deep learning for anomaly detection: A survey. *arXiv preprint arXiv:1901.03407*, 2019.

[7] T. Choudhary. Political bias in large language models: A comparative analysis of chatgpt-4, perplexity, google gemini, and claude. *IEEE Access*, 13:11341–11379, 2025. https://doi.org/10.1109/ACCESS.2024.3523764 doi: 10.1109/ACCESS.2024.3523764

[8] R. Churchill and L. Singh. The evolution of topic modeling. *ACM Computing Surveys*, 54, 01 2022. https://doi.org/10.1145/3507900 doi: 10.1145/3507900

[9] R. M. Cronin, D. Fabbri, J. C. Denny, S. T. Rosenbloom, and G. P. Jackson. A comparison of rule-based and machine learning approaches for classifying patient portal messages. *International journal of medical informatics*, 105:110–120, 2017.

[10] L. de Marcos and A. Domínguez-Díaz. Llm-based topic modeling for dark web qa forums: A comparative analysis with traditional methods. *IEEE Access*, 13:67159–67169, 2025. https://doi.org/10.1109/ACCESS.2025.3560543 doi: 10.1109/ACCESS.2025.3560543

[11] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, and T. Solorio, eds., *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota, June 2019. https://doi.org/10.18653/v1/N19-1423 doi: 10.18653/v1/N19-1423

[12] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, vol. 96, pp. 226–231, 1996.

[13] F. Gilardi, M. Alizadeh, and M. Kubli. Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120, 2023.

[14] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, 2017.

[15] S. Har-Peled, D. Roth, and D. Zimak. Constraint classification for multiclass classification and ranking. *Advances in neural information processing systems*, 15, 2002.

[16] Y. He, J. Lei, Z. Qin, and K. Ren. Dream: Combating concept drift with explanatory detection and adaptation in malware classification, 2024.

[17] E. Keogh and A. Mueen. *Curse of Dimensionality*, pp. 314–315. Springer US, Boston, MA, 2017. https://doi.org/10.1007/978-1-4899-7687-1$_1$92doi : 10100797814899768 71_192

[18] T. Kulesza, M. Burnett, W.-K. Wong, and S. Stumpf. Principles of explanatory debugging to personalize interactive machine learning. *Proceedings of the 20th International Conference on Intelligent User Interfaces*, 2015.

[19] A. Lahiri, S. Shukla, B. Stear, T. M. Ahooyi, K. Beigel, E. Margolskee, and D. Taylor. Benchmarking transformer embedding models for biomedical terminology standardization. *Machine Learning with Applications*, p. 100683, 2025.

[20] C. Li, Y. Ge, J. Mao, D. Li, and Y. Shan. Taggpt: Large language models are zero-shot multimodal taggers, 2023.

[21] D. Li, Z. L. Zhu, J. van de Loo, A. Masip Gomez, V. Yadav, G. Tsatsaronis, and Z. Afzal. Enhancing extreme multi-label text classification: Addressing challenges in model, data, and evaluation. In M. Wang and I. Zitouni, eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pp. 313–321. Association for Computational Linguistics, Singapore, Dec. 2023. https://doi.org/10.18653/v1/2023.emnlp-industry.30 doi: 10.18653/v1/2023.emnlp-industry.30

[22] S.-S. Li. An improved dbscan algorithm based on the neighbor similarity and fast nearest neighbor query. *IEEE Access*, 8:47468–47476, 2020. https://doi.org/10.1109/ACCESS.2020.2972034 doi: 10.1109/ACCESS.2020.2972034

[23] C. Lin, M. Mausam, and D. Weld. Re-active learning: Active learning with relabeling. In *Proceedings of the AAAI conference on artificial intelligence*, vol. 30, 2016.

[24] P. Lindstrom, S. J. Delany, and B. Mac Namee. Handling concept drift in a text data stream constrained by high labelling cost. In *FLAIRS*, 2010.

[25] J. D. Menke, H. Kilicoglu, and N. R. Smalheiser. Publication type tagging using transformer models and multi-label classification. *medRxiv*, 2025.

[26] OpenAI. Gpt-4o-mini. https://openai.com/, 2025. Large language model.

[27] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, et al. Improving language understanding by generative pre-training. 2018.

[28] N. Reimers and I. Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019.

[29] M. E. Roberts, B. M. Stewart, D. Tingley, C. Lucas, J. Leder-Luis, S. K. Gadarian, B. Albertson, and D. G. Rand. Structural topic models for open-ended survey responses. *American journal of political science*, 58(4):1064–1082, 2014.

[30] A. Rosenberg and J. Hirschberg. V-measure: A conditional entropy-based external cluster evaluation measure. In *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)*, pp. 410–420, 2007.

[31] L. Ruff, R. Vandermeulen, N. Goernitz, L. Deecke, S. A. Siddiqui, A. Binder, E. Müller, and M. Kloft. Deep one-class classification. In *International conference on machine learning*, pp. 4393–4402. PMLR, 2018.

[32] C. E. Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948.

[33] A. V. Solatorio. Gistembed: Guided in-sample selection of training negatives for text embedding fine-tuning, 2024.

[34] R. Tang, C. Zhu, B. Chen, W. Zhang, M. Zhu, X. Dai, and H. Guo. Llm4tag: Automatic tagging system for information retrieval via large language models. *arXiv preprint arXiv:2502.13481*, 2025.

[35] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[36] R. Xu and D. Wunsch. Survey of clustering algorithms. *IEEE Transactions on neural networks*, 16(3):645–678, 2005.

[37] C. Yin and Z. Zhang. A study of sentence similarity based on the all-minilm-l6-v2 model with "same semantics, different structure" after fine tuning. In *2024 2nd International Conference on Image, Algorithms and Artificial Intelligence (ICIAAI 2024)*, pp. 677–684. Atlantis Press, 2024.

[38] X. Zhang, J. Zhao, and Y. LeCun. Character-level convolutional networks for text classification. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, eds., *Advances in Neural Information Processing Systems*, vol. 28. Curran Associates, Inc., 2015.

[39] X. Zhang and S. Zhou. Woa-dbscan: Application of whale optimization algorithm in dbscan parameter adaption. *IEEE Access*, 11:91861–91878, 2023. https://doi.org/10.1109/ACCESS.2023.3307412 doi: 10.1109/ACCESS.2023.3307412