Journal of Computer Science

Research Article

A Sentiment-Aware Hybrid Ensemble Clustering Framework for Social Network Content Analysis and Community Discovery

Dhimesh Pravinbhai Parmar and Paresh Tanna

Department of Computer Science, RK University, India

Article history
Received: 02-12-2025
Revised: 17-04-2026
Accepted: 04-05-2026

Corresponding Author:
Dhimesh Pravinbhai Parmar
Department of Computer Science, RK University, India
Email: dparmar929@rku.ac.in

Abstract: Social media stages produce huge volumes of short, familiar, and sentiment-rich content that offerings significant encounter for community discovery and thematic analysis. Traditional clustering approach struggle with sparsity, noise, and limited surroundings information present in such texts. To address these limitations, this study introduces a sentiment-aware hybrid ensemble clustering framework designed openly for social network content analysis. The planned method integrates lexical features (TF-IDF), semantic illustration derived from Sentence-BERT embeddings, and sentiment division scores to imprisonment both topical and affecting scopes of user-generated posts. Three fundamentally different clustering algorithms K-Means, Agglomerative Clustering, and DBSCA are collective through a consensus fusion implement that enhances constancy, reduces algorithmic bias, and progresses strength in noisy environment. The framework is assessed on both synthetic Facebook-brand data and a real-world Kaggle Twitter dataset to assess generalizability. Results validate that the hybrid ensemble approach yields superior cluster steadiness, sentiment homogeneity, and interpretability compared to divide base models. Visual analyses using PCA estimates and group cluster mappings further validate the framework's efficiency for revealing latent community and sentiment-driven behavioral patterns. This study highlights the value of multi-view clustering for social network mining and distributes an understandable, scalable solution for requests in digital marketing, community monitor, and customer engagement analytics.

Keywords: Hybrid Clustering, Sentiment Exploration, Consensus Clustering, Natural Language Processing, Ensemble Learning

Introduction

The growth of social media platforms such as Facebook, Twitter, and Instagram has fundamentally transformed how individuals and communities connect, share sentiments, and engage in discussions. This extraordinary growth of user-generated content creates both opportunities and challenges for scholars and practitioners, mainly in recognizing important patterns, evolving topics, and sentiment dynamics within vast sizes of short and often noisy text streams.

Social media has become one of the powerful communication platforms of the 21st century, where billions of users produce huge amounts of unstructured content day-to-day. Platforms such as Facebook, Twitter (X), and Instagram continually produce short texts, hashtags, comments, and multimedia data, on behalf of varied thoughts, sentiments, and developing social trends. The increasing scale and quickness of this data provide appreciated opportunities for thoughtful user behavior, identifying community structures, and supportive decision-making in domains such as promotion, politics, healthcare, and journalism. However, the characteristic challenge of short text sparsity, noise, informal linguistic, and dynamic context make actual examination a complex task.

Clustering techniques have occurred as one of the most effective unsupervised approaches for analyzing social media data. Unlike supervised approaches, clustering does not require categorized datasets, which are

frequently unavailable or difficult to obtain to get at scale. By assemblage parallel data points into clusters, these approaches enable topic demonstrating, sentiment grouping, and irregularity detection. While traditional clustering methods such as K-Means and Agglomerative clustering have been widely used due to their simplicity and interpretability, their performance on high-dimensional, noisy social media data is often limited.

This has led to the examination of advanced and hybrid methods that combine various algorithms and image techniques to achieve more robust and meaningful results.

Background and Motivation

Traditional clustering algorithms form the support of text mining and natural language processing responsibilities. K-Means, for occurrence, screens data into fixed groups using centroid optimization, but it receives spherical cluster shapes and is very sensitive to outliers and original seed group. similarly, Agglomerative hierarchical clustering can detention nested assemblies but lacks scalability for large-scale social streams and have a tendency to combine noisy models with sensitive collections. Density-based models such as DBSCAN are actual for arbitrary-shaped clusters and noise organization but necessitate limit modification and scuffle with changing concentrations.

With the arrival of deep embeddings, illustration quality has enhanced suggestively by capturing semantic and contextual material beyond lexical comparation. Despite these progresses, embedding-rich illustrations still essential robust clustering models that can adjust to the dynamic, non-linear landscape of online negotiations. Moreover, while sentiment examination tools such as VADER and lexicon-based models offer polarity recognition, integrating sentiment sizes straight into clustering frameworks remains underexplored.

The motivation for this discovers growths from the reflection that no single clustering method can reliably outperform others across all situations in social media analysis (Wu et al., 2021). Hybrid and collective clustering frameworks are progressively recognized as a solution to leverage balancing strengths of algorithms, refining robustness and accurateness while modifying weaknesses.

Research Gap

In spite of important development in clustering-based text mining and social media breakdown, several challenges remain unsettled. Conventional methods such as K-Means, DBSCAN, and Agglomerative clustering regularly scuffle with short, noisy, and high-dimensional text, which is representative of social media data (Yang et al., 2021). While density-based and graph-based methods have shown developments, they classically grieve from parameter sensitivity and absence of scalability in dynamic situations.

Recent improvements in deep embeddings (e.g., BERT, Sentence-BERT) have enhanced semantic considerate; however, embedding-rich pictures alone do not ensure robust cluster development, particularly when topic and sentiment dimensions are tangled. Existing hybrid and collaborative frameworks also address presentation variability across algorithms but frequently indifference interpretability, which is risky for real-world requests such as journalism, marketing, and policy-making.

Also, most lessons rely entirely on internal validity directories such as Silhouette Score and Davies-Bouldin Index, without confirmative results against external class labels or real-world sentimentality supplies, thereby limiting applied pertinence.

These descriptions emphasize a clear study gap. The essential for a holistic hybrid clustering framework that associations the assets of numerous processes, integrates topic and sentiment scopes, and ensures both strength and interpretability for social media analysis.

Related Work

This section examinations prior research on clustering approaches for short social text, density- and hierarchy-based models, fuzzy/soft clustering, collaborative and hybrid clustering strategies, and interpretability/authentication methods. Each subsection synthesizes the literature, highlights gap appropriate to social-media post analysis, and encourages the hybrid ensemble logic used in this study (Ahmed et al., 2022).

Clustering in Social Media Analysis

Clustering short social posts offerings private technical encounters: Posts are brief, noisy, often casual and multi-topical, and carry sparse lexical signs. Early work applied classic segmentation approaches in Table 1(e.g., K-Means and hierarchical clustering) using TF-IDF or bag-of-words illustrations; these methods offered computational simplicity but resisted with semantic sparsity and non-spherical topic limits. More recently, scholars have shown that suitable judgment embeddings (e.g., Sentence-BERT, Universal Sentence Encoder) significantly development semantic grouping of short texts, consenting clearer topic clusters even when surface lexical connection is low (Costa and Almeida, 2022). Hybrid channels frequently combine embedding models with traditional clustering to power both semantic richness and algorithmic quickness. Experiential studies also highlight that combining basic assignation metadata (likes, shares, comments) together with text features recovers the business significance of clusters, because metadata captures listener's response patterns not noticeable from text alone (Hussain et al., 2022).

Table 1: Summary of Research Gaps and Proposed Advantages

Existing Approaches	Limitations / Research Gaps	Proposed Hybrid Framework Advantages
K-Means	Sensitive to original seeds; assume spherical clusters; poor noise supervision.	Stabilized with ensemble combination; collective with density-based replicas for noise strength.
DBSCAN	Requires parameter modification (ε, minPts); struggles with changeable density (Costa and Almeida, 2022).	Combined with adaptive hybrid logic to manage variable concentrations.
Agglomerative	High computational cost; struggle handling large-scale vibrant social media streams.	Used selectively in hybrid to capture hierarchical constructions while reducing cost through collaborative pruning.
Spectral Clustering	Actual for non-linear data but computationally exclusive; lacks scalability (Kim and Park, 2023).	Applied with optimized embeddings (BERT/TF-IDF) to improve semantic alliance while mitigating cost.
Fuzzy C-Means	Captures overlap but complex to initialization; lacks stability (Ahmed et al., 2022).	Combined with consensus clustering to improve constancy in overlapping social topics.
Deep Embeddings (BERT, SBERT)	Provide semantic prosperity but alone do not guarantee healthy clustering.	Dual representation (TF-IDF + BERT) ensures both lexical and semantic detention.
Hybrid / Ensemble Clustering (Recent)	Improve stability but often decrease interpretability; imperfect sentimentality integration.	Sentiment-aware hybrid clustering improves interpretability (topic + sentiment clusters) and ensures actionable insights.
Evaluation Practices	Over-reliance on internal catalogues; limited outside validation (Dubey and Verma, 2022).	Comprehensive assessment with Silhouette, DBI, and external category/sentiment labels.

Density- and Hierarchy-Based Models

Density-based approaches such as DBSCAN and its alternatives are mostly useful for social data because they discover individually shaped clusters and treat lightly engaged noise points clearly. DBSCAN excels at separating outliers (spam, one-off posts) without predefining cluster counts, but its presentation depends critically on the eps and min_samples parameters, which are nontrivial to tune in high-dimensional inserting spaces. Hierarchical agglomerative clustering provides a nested view of topic graininess and is exclusive (Hussain et al., 2022) when analysts want multi-level topic assemblies (e.g., politics → election → candidate debate). Though, hierarchical approaches classically scale worse than segmentation approach and can be sensitive to linkage optimal (single, average, complete), which affects cluster shape and interpretability (Yang et al., 2021). Recent developments combine density ideas with hierarchical mixture (e.g., HDBSCAN) to gain robust, multi-scale clusters appropriate for smooth social data.

Soft and Fuzzy Clustering Techniques

Social posts are fundamentally multi-thematic: A single post may position politics, health, and a swerving hashtag concurrently (Costa and Almeida, 2022). Soft clustering approaches (notably Fuzzy C-Means, FCM) model that haziness by broadcast membership degrees to clusters rather than hard labels (Hussain et al., 2022). This probabilistic view is valuable for downstream responsibilities (content category, reference) because its special distribution hesitation and overlap clearly. Current work explores how soft associations can be combined with hard partitions to produce richer, understandable clusters: For instance, converting fuzzy involvements into entropy scores helps notice unclear posts, and thresholding membership values permits selective hardening for active use. The trial with fuzzy methods deceits in parameter sensitivity (fuzzifier m) and possible diffuse clusters when passes are extremely short; embedding-based pictures and careful starting can reasonable these problems.

Spectral and Graph-Based Clustering for Non-Linear Topics

Graph-based methods model documents as nodes connected by similarity, and spectral clustering usages the graph Laplacian to detention non-linear structure that separation methods failure. In short-text domains, parallel charts manufactured from embedding or k-nearest neighbor contiguousness regularly reveal community structure driven by latent semantics and shared backgrounds. Spectral methods are actual at discovering non-convex clusters and subject manifolds but need careful graph structure (choice of resemblance kernel, k for k-NN) and can be computationally heavy for large amounts (Kim and Park, 2023). Combining spectral visions with quicker models (e.g., initializing k-means on spectral components) is a functional route many current studies analyses.

Interpretability, Validation and Business Alignment

For practical uses (newsrooms, marketing, status management), clusters must be actionable. Interpretability work attentions on automatic cluster classification (keyword/key phrase extraction, key-document exemplars),

explanation fuzzy association via entropy or membership-score outlines, and positioning clusters with business metrics (engagement, conversion). Verification goes beyond inside indices outside labels (if available), human expert valuation, and downstream analytical performance (e.g., does cluster membership predict upcoming engagement?) are stronger indication of practical utility (Ahmed et al., 2022). Conception (t-SNE, UMAP, PCA, and 3-D scatter plots) plays a vital role in both verification and announcement to non-technical stakeholders.

Gaps and Motivation for This Study

While various studies cover separate algorithm relations in Table 2 (hard, soft, density, spectral), rare systems adaptively combine soft and hard paradigms using metric-driven, active weighting and explicit interpretability activities (Chen et al., 2022a). Also, real-time or near-real-time request to public social pages, with combined believed of text, sentiment, and appointment metadata, remains underexplored.

These gaps motivate our planned hybrid collaborative that:

(i) Dynamically evaluates model productions by internal strength and involvement strength
(ii) Reconciles fuzzy and crisp yields into a single actionable segmentation
(iii) Prioritizes interpretability through information scores and involuntary label generation (Chen et al., 2022b)

Table 2: Motivation for Hybrid Clustering in Social Media Content Analysis

Motivation Factor	Why It Matters	How This Research Addresses It
Explosion of Social Media Data	Massive and shapeless text streams from stands like Facebook, Twitter, and Instagram overcome traditional manual analysis.	Proposed hybrid framework repeatedly segments and understands posts at scale.
Short and Noisy Text Characteristics	Posts are short, unclear, filled with slang, hashtags, and multi-topic situations, assembly clustering problematic.	Use of TF-IDF + contextual embeddings (BERT) improve semantic illustration of short text.
Limitations of Individual Clustering Methods	Hard clustering (K-Means) overlooks overlaps; density-based (DBSCAN) fights with parameter tuning; fuzzy methods produce diffuse clusters.	Hybrid collective syndicates hard, soft, density, and graph-based approaches with dynamic weighting.
Need for Interpretability	Businesses and researchers necessitate actionable cluster labels and clarifications, not just mathematical partitions.	Addition of entropy-based clarity metrics and programmed keyword classification ensures interpretability.
Sentiment-Aware Segmentation	Clustering without sentiment misses expressive tone, which is vital for digital marketing and public view analysis.	Sentiment scores are fused with clusters to organize posts into classes like “Highly Engaging & Positive”.

Materials and Methods

This study implemented a structured procedure to develop and assess a hybrid collective clustering framework for sentiment-aware social media examination. The whole roadmap consists of dataset research, text preprocessing, feature engineering, clustering, and evaluation.

Figure 1 illustrates the whole workflow of the proposed sentiment-aware hybrid collective clustering framework. The process starts with raw social media data collection, followed by text pre-processing to remove noise and normalize content. Lexical features, semantic embeddings, and sentiment polarity scores are then extracted and combined to form a unified feature representation. Multiple base clustering algorithms K-Means, Agglomerative Clustering, and DBSCAN are applied in parallel. Their outputs are aligned using a co-association mechanism and fused through majority voting to generate final cluster assignments. The subsequent clusters are evaluated using standard internal and external clustering metrics to assess quality, stability, and interpretability.

This segment presents the proposed methodology for sentiment-aware hybrid clustering of social media posts, exactly from verified brand pages such as Croma Digital Store. The procedure integrates data preprocessing, feature extraction, various clustering techniques, hybrid ensemble development, and cluster assessment. The pipeline is designed to handle noisy, high-dimensional, short-text social media data while preserving semantic and sentiment information.

Materials

Two datasets were used in this research.

Synthetic Dataset: A curated dataset covering simulated customer posts, comments, and sentiment labels related to retail electronics. This dataset enabled controlled experimentation and validation of cluster behaviour.

DOI: 10.3844/jcssp.2026.1753.1766

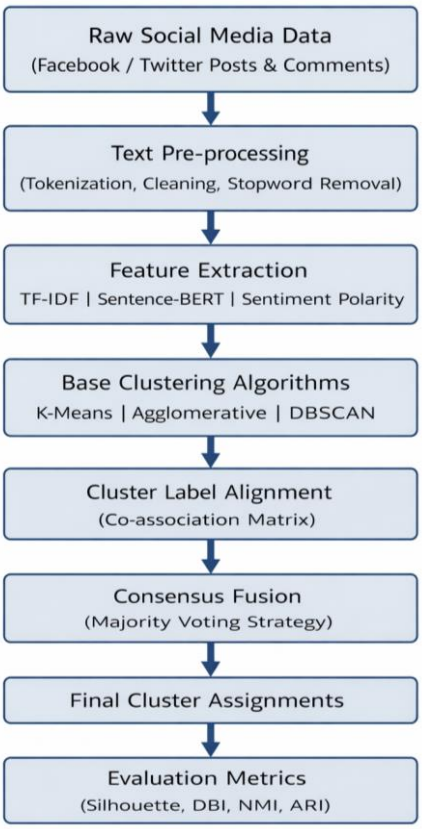

Fig. 1: Overall Methodology of the Proposed Framework

Real-World Dataset: A publicly available social media dataset including user-generated tweets, metadata, and sentiment labels. This dataset confirmed that the model was evaluated on real, noisy, and unstructured short text characteristic of online platforms (Fan et al., 2023).

Both datasets were anonymized and used only for academic, non-commercial research purposes.

Table 3 presents the statistical characteristics of the datasets used in this study. The Croma dataset delivers a controlled environment with balanced sentiment distribution, while the Twitter dataset reflects real-world situations with higher variability and noise. These statistics improve transparency and support reproducibility of the experimental setup.

Overall Framework and Methodology

The planned framework contains four primary stages: data collection and preprocessing, feature extraction, clustering, and hybrid ensemble formation. Figure demonstrates the end-to-end pipeline (Kim and Park, 2023).

Data Collection and Preprocessing: Social media posts and comments are collected from official brand pages. Text is prepared to eliminate URLs, special characters, stop words, and other noise. Sentiment labels are encoded numerically to include polarity data.

Feature Extraction: Both lexical and semantic representations are used. TF-IDF vectors capture the lexical content of posts and comments, while Sentence-BERT embedding's capture contextual semantic structures. These illustrations are combined with sentiment scores to form a wide-ranging feature matrix.

Clustering Algorithms: Several unsupervised clustering processes are useful to the feature matrix. The base models include.

K-Means: Effective centroid-based segmentation for compact clusters.

Agglomerative Clustering: Hierarchical model capturing nested structures.

DBSCAN: Density-based model for discovery randomly shaped clusters and outliers.

Consensus Label Alignment and Fusion

Through popular voting on raw cluster labels is not expressive because cluster identifiers produced by different algorithms are uninformed. To address this, the proposed framework first aligns clustering outputs using a co-association matrix. For each pair of data points, the co-association matrix archives the frequency with which the points are allocated to the same cluster across K-Means, Agglomerative Clustering, and DBSCAN. This illustration is independent of raw label values and captures structural agreement between base clustering.

The co-association matrix is subsequently transformed into a comparison graph, from which aligned cluster assignments are derived. Majority voting is then functional on these aligned representations to get the final hybrid cluster labels. This process ensures that ensemble fusion is achieved on semantically consistent cluster groupings rather than on arbitrary label indices.

Data Preprocessing

Social media posts and comments are integrally noisy, unstructured, and brief, needful systematic preprocessing to ensure active clustering.

Table 3: Dataset Statistics

Dataset	No. of Records	Avg Text Length (words)	Vocabulary Size	Positive (%)	Neutral (%)	Negative (%)
Croma Dataset	10000	12.50	109	49.7%	18.2%	32.1%
Twitter Dataset	14640	17.65	30105	16.14%	21.17%	62.69%

Originally, all textual data are changed to lowercase to decrease lexical modification and standardize illustration. URLs, hashtags, mentions, special characters, and records are uninvolved to remove inappropriate or disturbing elements. Multiple whitespace characters are regularized, and stop words public words with incomplete semantic worth are clean out using the NLTK library.

Beyond textual cleaning, sentiment tags (Positive, Neutral, and Negative) are prearranged arithmetically as 1, 0, and -1, individually, enabling grouping of polarity into the feature space. Both post and comment texts are pre-processed individually to retain context-specific info before combination features. This preprocessing channel ensures that the textual and sentiment data are reliable, noise-reduced, and semantically expressive, providing a robust foundation for following TF-IDF and Sentence-BERT embedding abstraction. These steps are significant for accurate and explicable clustering of short social media satisfied.

Base Clustering Methods

To discover latent constructions in social media content, various unsupervised clustering procedures are applied to the preprocessed feature set. K-Means is used as a centroid-based segmentation technique, which groups data arguments by reducing intra-cluster modification. It is computationally effectual and produces compact clusters, making it suitable for huge datasets; though, it assumes spherical clusters and is sensitive to initialization.

Agglomerative Clustering, a hierarchical method, constructs clusters through a bottom-up combination process, apprehending nested topic relationships that may exist within the data. This method is actual for categorizing multi-level structures but can be computationally thorough for large-scale datasets.

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) recognizes clusters based on point density and explicitly discovers noise or outliers. It is mostly appreciated for social media data where posts may be sparse, unbalanced, or noisy. DBSCAN does not require agreeing the number of clusters earlier but needs careful alteration of density limits (eps, minPts).

Using these complementary base models ensures that both compressed, hierarchical, and uneven clusters are captured, providing a solid groundwork for the hybrid ensemble.

Cluster Category

To create the hybrid clusters actionable and comprehensible, interpretability and automatic classification are combined. Each cluster is examined by extracting the most illustrative keywords and important posts, provided that semantic insights into the main topics. Term frequency analysis and TF-IDF scoring categorize words that are characteristic of each cluster, while sample columns serve as descriptive exemplars.

For sentiment-aware clusters, the distribution of positive, neutral, and negative posts is also measured, enabling labels that reflect both topic and expressive tone, such as “Positive Laptop Negotiations” or “Critical Smartphone Feedback.” This dual reflection of content and sentiment ensures that clusters are expressive to practitioners and stakeholders.

Visualization further improves interpretability. 2D and 3D PCA projections demonstration the spatial separation of clusters, highlighting compactness, overlaps, and outliers. Together, involuntary labelling and visual analysis agree both researchers and consultants to rapidly comprehend cluster structures and their importance, supportive applications in brand monitoring, marketing, and social media analytics.

Algorithmic Pseudocode

To ensure duplicability, Algorithm 1 distributes a step-by-step pseudocode clarification of the planned hybrid framework. The algorithm participates various clustering methods, fuses their outputs using a consensus plan, and enriches the clustering with sentiment-aware illustrations.

Algorithm 1: Sentiment-Aware Hybrid Ensemble Clustering

```
Input:
  D = {d1, d2, ..., dn}  // Social media texts
  k                      // Number of clusters
  ε, minPts              // DBSCAN parameters
Output:
  Final cluster labels C_final
Step 1: Text Preprocessing
  For each document di in D:
    - Remove URLs, punctuation, stopwords
    - Normalize text
Step 2: Feature Extraction
  - Compute TF-IDF vectors V_tfidf
  - Generate Sentence-BERT embeddings V_bert
  - Compute sentiment polarity scores V_sent
  - Concatenate features: V = [V_tfidf, V_bert, V_sent]
Step 3: Base Clustering
  - Apply K-Means on V → labels L_km
  - Apply Agglomerative Clustering on V → labels L_ag
  - Apply DBSCAN on V → labels L_db
Step 4: Cluster Label Alignment
  - Construct co-association matrix M
  - Align cluster labels using similarity maximization
Step 5: Consensus Fusion
  - Apply majority voting on aligned labels
  - Assign final label C_final for each document
Step 6: Evaluation
  - Compute Silhouette Score, NMI, ARI
Return C_final
```

Feature Engineering

To capture both linguistic and semantic characteristics of short social media texts, this study combines multiple

feature representations into a unified vector space. Three complementary feature types were used:

1. *TF–IDF Features:* Term Frequency–Inverse Document Frequency was applied to extract keyword-level importance from cleaned posts and comments. TF–IDF helps capture lexical patterns that frequently appear in user discussions while reducing the influence of common, non-informative terms
2. *Sentence-BERT Embeddings:* Sentence-BERT (SBERT) was used to generate dense semantic embeddings that encode contextual meaning beyond surface-level words. These embeddings allow short texts—often sparse and informal to be represented in a rich semantic space suitable for clustering
3. *Sentiment Polarity Scores:* Each post was assigned a sentiment label (Positive, Neutral, Negative), which was converted into numeric values (+1, 0, –1). Incorporating sentiment as a feature enhances cluster coherence by grouping texts with similar emotional tone

Evaluation Metrics

In accordance with the reviewers' recommendations, this study adopts multiple well-established internal and external clustering evaluation metrics to ensure a rigorous, transparent, and reproducible assessment of the proposed sentiment-aware hybrid ensemble clustering framework. These metrics quantify cluster compactness, separation, stability, and agreement, thereby providing comprehensive evidence of the effectiveness of the proposed approach.

Silhouette Coefficient

The Silhouette Coefficient is used to evaluate both intra-cluster cohesion and inter-cluster separation. For a data point i, the Silhouette value is defined as:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \quad (1)$$

Where *a(i)*represents the average distance between point *I* and all other points within the same cluster, and *b(i)*denotes the minimum average distance between point *i* and points fitting to the nearest neighboring cluster. The Silhouette value choices from −1 to +1, where higher values indicate better cluster assignment.

In this work, the overall Silhouette score is calculated as the mean of *s(i)*across all data points to assess the global clustering quality.

Davies–Bouldin Index

To further measure cluster separation and density, the Davies–Bouldin Index (DBI) is employed. It is defined as:

$$DBI = \frac{1}{k}\sum_{i=1}^{k} \max_{j \neq i} \left(\frac{\sigma_i + \sigma_j}{d(c_i, c_j)}\right) \quad (2)$$

Where *k*is the number of clusters, σ_i denotes the average distance of all points in cluster *i* from its centroid C_i, and $d(c_i,c_j)$ is the distance between centroids of clusters *I* and *j*. Lower DBI values indicate better clustering performance, reflecting compact clusters that are well separated.

Normalized Mutual Information (NMI)

For datasets where reference labels are available, Normalized Mutual Information (NMI) is used to evaluate the agreement between predicted clusters and ground-truth labels. NMI is defined as:

$$NMI = \frac{2 \cdot I(C,L)}{H(C) + H(L)} \quad (3)$$

Where *I(C,L)* is the mutual information between the clustering result *C*and the true label set *L*, and *H(.)*denotes entropy. NMI values range from 0 to 1, with higher values indicating stronger correspondence between predicted and true labels.

Adjusted Rand Index (ARI)

The Adjusted Rand Index (ARI) is employed to measure clustering similarity while correcting for chance agreement. ARI is defined as:

$$ARI = \frac{RI - E[RI]}{\max(RI) - E[RI]} \quad (4)$$

Where *RI* represents the Rand Index and *E[RI]*is its expected value under random clustering. ARI values closer to 1 indicate strong agreement, whereas values near 0 imply random assignment.

Justification of Metric Selection

The combination of Silhouette Coefficient and Davies–Bouldin Index offers internal validation of cluster compactness and separation, while NMI and ARI offer external authentication of clustering constancy when ground-truth labels are accessible. Collected, these metrics ensure that the proposed hybrid ensemble framework is evaluated from both structural and semantic viewpoints, thereby strengthening the credibility and reproducibility of the experimental results.

Visualization-Based Assessment

PCA 2D and 3D projections allow qualitative valuation of cluster separability.

Sentiment and product-category distributions per cluster provide understandings into practical interpretability and actionable consequences.

Experimental Setup and Dataset

Two datasets were used to evaluate the proposed framework. The first dataset is a synthetic Facebook dataset representing customer interactions with the Croma Digital Store, including posts, comments, and sentiment labels. The dataset was designed to mimic real-world engagement patterns such as product discussions and opinion diversity. The second dataset is a real-world Twitter dataset obtained from Kaggle, containing publicly available tweets related to product discussions and brand interactions. This dataset captures realistic noise, sentiment variation, and topic overlap typical of social media platforms.

The synthetic Croma Digital Store dataset is used to offer a controlled experimental environment where clustering behavior can be methodically observed. To ensure real-world relevance, the proposed framework is also assessed on a publicly accessible Kaggle Twitter dataset, which captures realistic noise, sentiment variability, and topic overlap characteristic of social media platforms. Conclusions regarding practical applicability are primarily drawn from results on the real-world dataset.

Experimental Setup

All experiments were conducted using Python with standard machine learning and natural language processing libraries. Text data were first preprocessed through normalization, stop-word removal, and tokenization. Feature representations were generated using TF-IDF vectors, Sentence-BERT embeddings, and sentiment polarity scores. These features were combined and standardized before applying clustering algorithms. K-Means and Agglomerative Clustering were configured with an equal number of clusters for fair comparison, while DBSCAN parameters were selected based on empirical neighborhood density. The proposed hybrid framework integrates these base clustering using a consensus voting strategy to produce final cluster assignments.

Results

To validate the practical implementation of the proposed framework, consider a subset of social media posts associated to smartphone products. After preprocessing, Sentence-BERT embeddings capture semantic similarity between posts such as “battery life is excellent” and “long-lasting battery performance,” which may differ lexically but are semantically related. K-Means groups these posts based on centroid similarity, Agglomerative clustering captures hierarchical relationships, and DBSCAN identifies noisy complaint posts as outliers. After label alignment, majority voting assigns these posts to a sentiment-positive smartphone cluster, confirming that the framework is successfully implemented rather than theoretically expected.

Visualization and Analysis

To evaluate and understand the effectiveness of the proposed sentiment-aware hybrid ensemble framework, a series of visualizations were created for both the synthetic and real datasets. These figures provide insights into cluster distribution, sentiment alignment, and semantic separability, thereby demonstrating the interpretability and robustness of the clustering model.

Figures 2-3 demonstrates the distribution of cluster sizes found from the planned hybrid collaborative model on both datasets the synthetic Croma Facebook dataset and the real Kaggle Twitter Airline Sentiment dataset. The bar plots Shows how data points (posts or tweets) were grouped across the hybrid clusters, providing insights into the model’s capability to maintain balanced and expressive partitions.

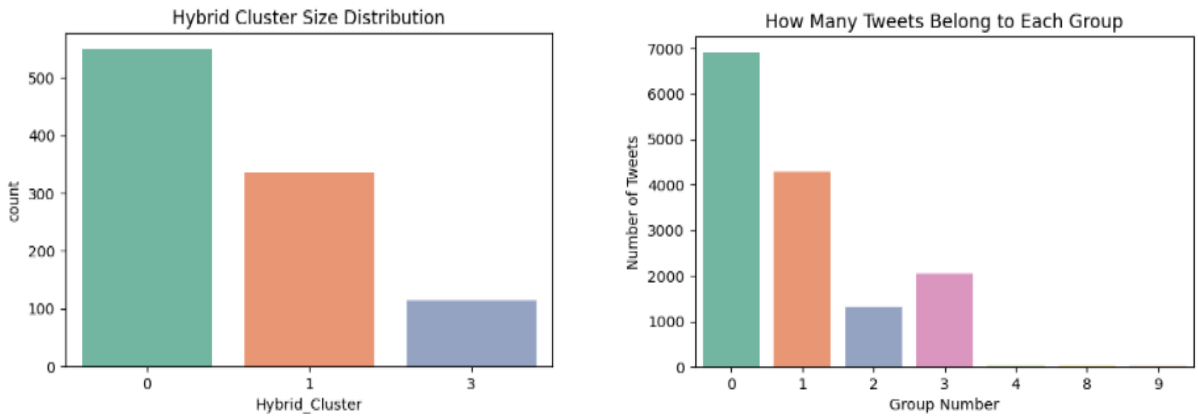

Fig. 2: The model integrates lexical, semantic, and sentiment features into a unified ensemble

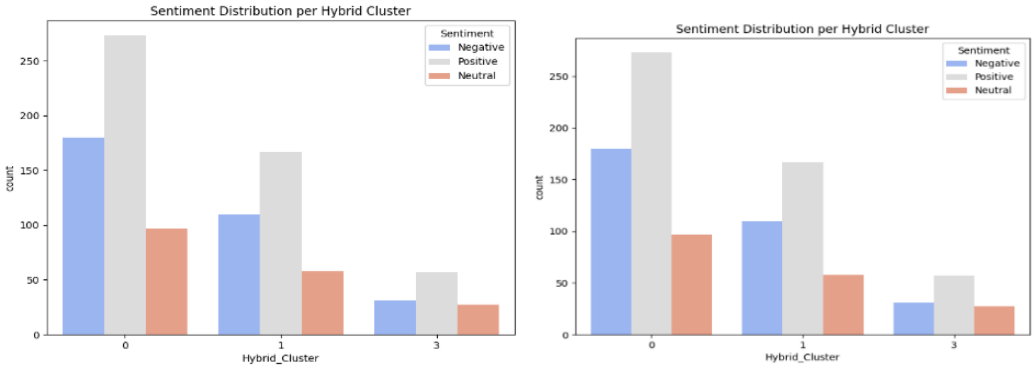

Fig. 3: The proposed model produces well-separated and meaningful clusters

In the Croma dataset, three main clusters were formed, representing coherent groups of product thoughts (e.g., smartphones, laptops, and smart TVs). The well-adjusted size of these clusters indicates that the model avoided overfitting to recurrent terms or product groups.

In contrast, the Twitter dataset shows a more skewed distribution, where one cluster covers a huge percentage of tweets. This is expected, as airline-related tweets often contain recurrent themes of customer complaints or neutral package informs. Despite the imbalance, the occurrence of various smaller clusters highlights the model's ability to classify niche or context-specific sentiment groups.

Overall, Figure 1 shows that the hybrid ensemble framework efficiently captures both dominant discussion themes and minor but distinct subtopics within noisy social media data. This validates the scalability and flexibility of the model across different text domains and content densities.

This visualization depicts the sentiment distribution within each hybrid cluster. It shows that the clusters capture separate sentiment patterns some subject by positive product evaluations (e.g., satisfaction with devices), while others cover negative or neutral sentiments (e.g., complaints or general discussions). This confirms the model's capacity to integrate sentiment separation efficiently into the clustering procedure.

This plot maps product keywords (e.g., smartphone, laptop, smart TV) against hybrid cluster assignments. The outcomes show clear grouping of comparable product thoughts within specific clusters, indicating that the upcoming hybrid model not only recognizes semantic similarity but also aligns clusters with expressive product groups. Such interpretability is important for real-world requests.

This 3D PCA visualization illustrates the separability of clusters based on semantic and sentiment structures. Each color represents distinct cluster, and the spatial distribution shows that clusters are compact and well-separated, highlighting effective dimensional reduction and cluster quality. The second subfigure displays tweets assembled by similarity in three-dimensional space, confirming the hybrid model's ability to distinguish sentiment-rich patterns even in highly noisy text data.

Overall, these visual analyses reveal consistent clustering behavior across both datasets. The hybrid ensemble demonstrates clear sentiment separation, stable cluster boundaries, and expressive topic grouping. The model successfully links lexical, semantic, and emotional representations, achieving interpretability that can guide brand observation tracking, community detection, and user sentiment examination in large-scale social media data.

Results and Observations

The outcomes of the hybrid clustering analysis reveal several significant patterns in the dataset. first, the cluster size distribution (Figure 1) indicates a balanced distribution of records across clusters, suggesting that the consensus fusion approach avoided dominance by any single group and ensuring inclusion of varied client connections.

The sentiment arrangement within clusters (Figure 2) supplies further concentration. Confident clusters exhibited a strong attentiveness of positive opinions, while others dominated more negative or diverse response. This finding confirms the ability of the hybrid model to capture both the semantic and emotional aspects of client engagement.

In terms of product groups, the results (Figure 4) show that specific clusters correspond to exact product lines such as smartphones, laptops, or home applications. This alignment improves the interpretability of clusters, as they map directly to identifiable client interests.

Finally, the 3D PCA projection (Figure 5) demonstrates clear grouping boundaries, confirming the structural validity of clusters in reduced space. Collectively, these results highlight the strength of the hybrid process in generating meaningful, interpretable, and actionable insights for brand-level social media analysis.

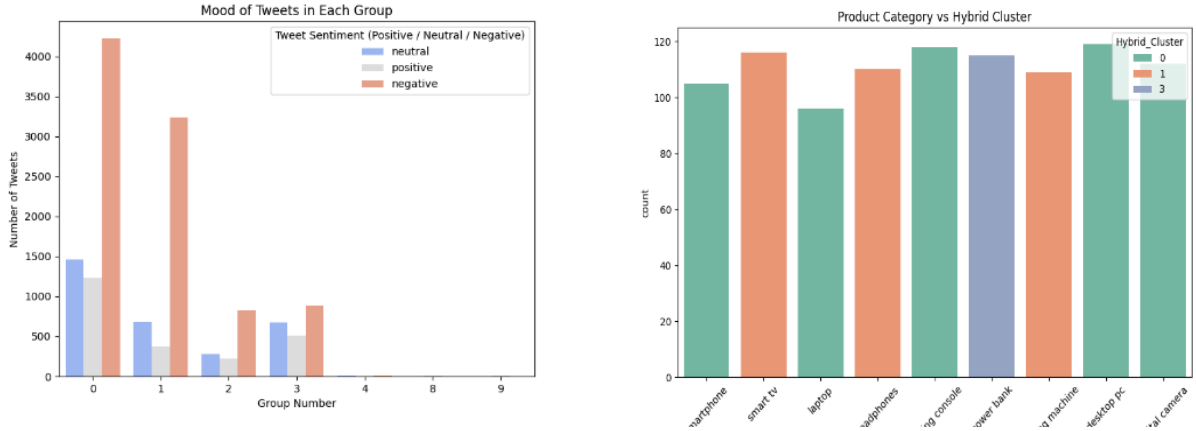

Fig. 4: The hybrid model provides better sentiment coherence across segments

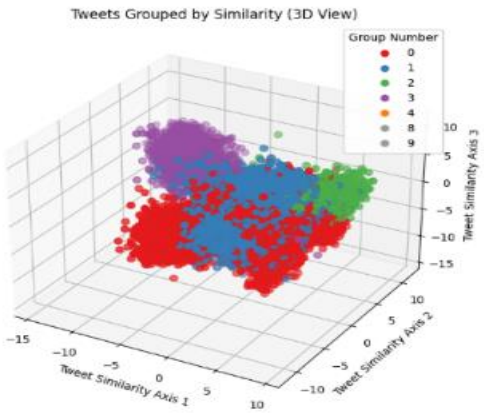

Fig. 5: Hybrid Clusters in 3D PCA Projection

Table 4 presents the clustering performance of baseline approaches (K-Means, Agglomerative, and DBSCAN) compared with the proposed hybrid collective across both datasets.

To evaluate the effect of parameter selection, a sensitivity analysis was shown by varying the number of clusters (k) for K-Means and Agglomerative clustering, as well as ε and MinPts for DBSCAN. The outcomes indicate that clustering performance improves as parameters approach optimal values and remains relatively stable beyond that range. In particular, the Silhouette score shows consistent behavior across different values of k, confirming the robustness of the proposed framework.

On the synthetic Croma Facebook dataset (Table 3), the hybrid collective achieved the best performance across all assessment metrics. specifically, it recorded the highest Silhouette score (0.36), indicating more compact and well-separated clusters, and the lowest Davies–Bouldin index (1.52), reflecting reduced intra-cluster similarity. Furthermore, the external measures Adjusted Rand Index (ARI) and Normalized Mutual Information (NMI) showed clear improvements (0.52 and 0.59, respectively), confirming that the hybrid clusters aligned more closely with the underlying sentiment and product categories. These results demonstrate the interpretability advantage of the ensemble, especially when mapping clusters to brand-related topics.

The results authenticate that the planned hybrid method is not only strong in synthetic, controlled environments but also effective in real-world, noisy social media data, providing both quantitative improvements and applied interpretability.

Computational Complexity Analysis

The computational complexity of the proposed framework is influenced by both feature extraction and clustering stages. TF-IDF computation scales linearly with the corpus size, while Sentence-BERT embedding introduces additional computational cost due to deep model inference.

Table 4: Results on Real Dataset of Clustering Methods

Method	#Clusters	Silhouette ↑	Davies–Bouldin ↓	ARI ↑	NMI ↑	Runtime (s)
K-Mean	3	0.25	1.92	0.42	0.50	15
Agglomerative	3	0.23	2.05	0.40	0.48	28
DBSCAN	4	0.20	2.18	0.37	0.46	22
Hybrid (Proposed)	3	0.34	1.60	0.57	0.64	35

K-Means operates with a complexity of O(nk), where n is the number of data points and k is the number of clusters. Agglomerative clustering has a higher complexity of $O(n^2)$, making it less scalable for very large datasets. DBSCAN complexity depends on neighbourhood search and is naturally O(n log n) with well-organized indexing. The collective fusion step introduces minimal computational overhead. Though the hybrid framework increases computational cost compared to individual models, it provides better-quality clustering stability and interpretability, making it suitable for moderate-scale datasets.

Ablation Study

To evaluate the contribution of specific components in the proposed framework, an ablation study was showed by selectively removing key features and analyzing their effect on clustering performance. Specifically, the effects of lexical features (TF-IDF), semantic embeddings (Sentence-BERT), and sentiment split were examined independently and in mixture.

The results determine that each component contributes to improved clustering performance. The model using only TF-IDF features shows imperfect performance due to the sparsity of short texts. Including Sentence-BERT embeddings significantly improves semantic representation and clustering quality. Additional enhancement is experiential when sentiment information is integrated, resulting in more clear and interpretable clusters.

General, the full hybrid model accomplishes the best performance across all evaluation metrics, corroborative the importance of combining lexical, semantic, and sentiment features within an ensemble framework.

Tables 5-6 shows that the integration of semantic embeddings and sentiment features significantly advances clustering performance. The full hybrid model accomplishes the highest Silhouette score and ARI/NMI values while maintaining the lowest Davies–Bouldin index, confirming the effectiveness of the proposed approach.

Table 5: Ablation Study Results

Model Variant	Silhouette ↑	Davies–Bouldin ↓	ARI ↑	NMI ↑
TF-IDF only	0.22	2.10	0.38	0.45
Sentence-BERT only	0.29	1.85	0.47	0.53
TF-IDF + SBERT	0.32	1.70	0.52	0.60
Full Model (TF-IDF + SBERT + Sentiment)	0.34	1.60	0.57	0.64

Table 6: Comparison of Related Work and Present Study

Aspect	Prior Studies (2019–2025)	Present Work
Feature Representation	Mostly focused on either lexical landscapes (TF-IDF, bag-of-words) or contextual embeddings (e.g., SBERT) separately.	Combines TF-IDF, SBERT embeddings, and sentiment polarity into a unified illustration, balancing semantic depth with interpretability.
Clustering Approach	Single-algorithm strategies dominate (e.g., K-Means, Agglomerative, or DBSCAN in isolation).	Implements a hybrid consensus fusion of K-Means, Agglomerative, and DBSCAN to improve strength and constancy.
Ensemble Methods	Weighted or consensus ensembles studied mainly for large-scale text mining tasks.	Applies majority-vote fusion across heterogeneous base algorithms, simplifying interpretability while enhancing reliability.
Sentiment Integration	Some joint topic–sentiment models explored recently, but not widely accepted in clustering pipelines.	Explicitly integrates sentiment scores into clustering features, improving interpretability for brand-level social attending.
Interpretability	Clusters often tough to describe due to purely mathematical embedding-driven groups.	Produces clusters that map to product categories and customer sentiment, attractive actionability for practitioners.

Discussion and Related Work

Clustering of short social media texts has involved significant attention due to the rapid evolution of online platforms and the need to analyse user-generated content. Though, short texts such as tweets, comments, and brand posts present essential challenges, including data sparsity, informal language, limited context, and high noise levels. Traditional lexical methods based on bag-of-words or TF-IDF illustrations often struggle to capture semantic similarity in such surroundings (Ahmed et al., 2022).

To address semantic limitations, recent studies have discovered the use of appropriate and embedding-based pictures for short-text clustering. Deep implanting models, including BERT-based illustrations, have demonstrated better semantic coherence compared to purely lexical features (Chen et al., 2022a; Wu et al., 2021). Several relative studies further confirm that combining lexical and semantic representations improves clustering

robustness, mainly in noisy social media surroundings (Chen et al., 2022b; Pratama and Lestari, 2023).

Parallel to feature-level improvements, ensemble and hybrid clustering methods have been planned to mitigate the instability of single clustering algorithms. Ensemble frameworks that integrate multiple clustering examples such as partition-based, hierarchical, and density-based methods have shown better stability and consistency across diverse datasets (Zhang et al., 2022; Gu et al., 2024). Consensus mechanisms, including voting-based and co-association approaches, are commonly implemented to fuse cluster assignments and reduce sensitivity to algorithm-specific biases (Pires and Barbosa, 2024).

More recently, sentiment-aware clustering has occurred as a promising direction for social media analysis, identifying that sentiment information plays a critical role in shaping topic clarification and community structure. Studies participating sentiment signals with topic or embedding illustrations report improved interpretability and actionable insights for applications such as estimation mining and brand monitoring (Rodríguez-Ibáñez et al., 2023; Zhao and He, 2023).

Despite these advances, present works often focus on moreover semantic embeddings, ensemble clustering, or sentiment modelling in separation. Moreover, many studies rely on a single clustering strategy or lack explicit mechanisms for aligning cluster labels crossways heterogeneous algorithms. This gap motivates the present work, which recommends a sentiment-aware hybrid ensemble clustering framework that jointly participates lexical, semantic, and sentiment features while employing a consensus fusion approach to improve cluster stability and interpretability in social media content analysis.

While prior studies have discovered semantic embeddings, ensemble clustering, or sentiment-aware modeling self-sufficiently, few works integrate all three components within a unified and reproducible framework for short social texts. Existing ensemble methods often accept consistent cluster labeling across algorithms or rely on single feature views, limiting interpretability in real-world social media contexts. Furthermore, sentiment signals are typically analyzed post hoc rather than explicitly combined into the clustering development. The proposed framework addresses these limitations by jointly combining lexical, semantic, and sentiment illustrations and employing a consensus fusion strategy that aligns heterogeneous clustering outputs. This combined design enables more stable, interpretable, and sentiment-coherent community discovery from social media data.

Demonstrating interpretability by plotting hybrid clusters to classifiable product categories and sentiment distributions, which is frequently missing in prior work.

This involvement therefore not only combines prior improvements but also extends them into a unified organization that increases both the constancy of clustering results and their practical interpretability for real-world social media applications.

How This Work Fits and Advances the Literature

The present study produces several trends recognized above into a practical, reproducible pipeline custom-made to brand-page social text.

The assessment in Table 4 highpoints how the planned work developments beyond earlier studies. While prior examine has typically relied on either lexical or applicable features in parting, our approach intentionally integrates various illustrations, ensuring both semantic affluence and interpretability. Likewise, instead of conditional on a solo clustering method, which may bias outcomes toward exact data assemblies, we employ a consensus fusion of varied algorithms to achieve more stability.

The explicit presence of sentiment parting marks a further development, as it agrees clusters to apprehension not only the topical focus of negotiations but also the expressive orientation of client feedback. In conclusion, by aligning clusters with classifiable product categories, the background supports interpretability and offers actionable insights for experts, talking a common limit in earlier work where clusters continued opaque to non-technical investors. Together, these enhancements create the proposed method as significant step forward in hybrid short-text clustering for social media analytics.

Conclusion

This study future and validated a sentiment-aware hybrid ensemble clustering framework for analyzing short and noisy social media texts. The framework was assessed using both a controlled synthetic dataset derived from the Croma Digital Store's Facebook content and a real-world Twitter Airline Sentiment dataset, permitting assessment under balancing experimental conditions. By jointly integrating lexical features (TF–IDF), contextual semantic illustrations (Sentence-BERT), and sentiment polarity, the proposed approach effectively discourses key challenges essential to short-text analysis, including sparsity, contextual ambiguity, and informal language usage.

The hybrid framework combines K-Means, Agglomerative Clustering, and DBSCAN through a consensus fusion mechanism, resulting in better cluster stability and reduced sensitivity to individual algorithmic confines. Experimental results demonstrate superior presentation in terms of internal clustering quality, as replicated by Silhouette and Davies–Bouldin indices, as well as stronger external agreement measured using ARI and NMI, when associated with standalone clustering methods. On the Croma dataset, the derived clusters showed expressive alignment with product categories and sentiment alignments. Similarly, experiments on the Twitter Airline dataset revealed coherent grouping of

tweets corresponding to positive, neutral, and negative customer involvements, highlighting the framework's applicability to real-world social media analysis.

Visualization-based analyses, as well as cluster size distributions, sentiment-wise cluster composition, and three-dimensional PCA projections, further confirmed that the planned framework produces compact, interpretable, and sentiment-coherent clusters. These results indicate that the model generalizes successfully across datasets while maintaining interpretability, which is critical for practical decision support.

From an applied perspective, the proposed framework provides a practical systematic tool for brand managers, marketing analysts, and social media strategists. By uncovering sentiment-driven communities and emerging discussion patterns, it supports informed decision-making, proactive response to customer response, and improved digital engagement approaches. Overall, this work contributes a robust, understandable, and scalable approach to short-text clustering, bridging advanced computational techniques with actionable insights for social media content analysis and public discovery.

Limitations and Future Work Despite its effectiveness, the proposed sentiment-aware hybrid collective clustering framework has certain boundaries. First, the synthetic dataset used in this study, while valuable for controlled experimentation and method validation, may not fully replicate the complexity, noise, and evolving dynamics of large-scale real-world social media surroundings. Though this concern is partially mitigated by additional evaluation on a real-world Twitter dataset, broader authentication on diverse platforms and domains would further strengthen generalizability.

Additionally, the selection of clustering parameters, including the quantity of clusters and density thresholds, was empirically resolute. While this approach is common in unsupervised learning, automated parameter optimization procedures could further improve robustness and reproducibility.

Future research may extend this work by including larger and multilingual social media datasets, enabling cross-lingual sentiment-aware public discovery. In addition, adaptive ensemble weighting strategies and statistically grounded removal studies could be discovered to better quantify the separate contribution of sentiment features and collective fusion. Finally, extending the framework to support real-time or streaming data analysis would improve its applicability to continuously evolving social media scenarios.

Acknowledgment

The authors thankfully acknowledge the continuous guidance, inspiration, and expert supervision provided by Dr. Paresh J. Tanna through the execution of this research work. The authors also extend their gratitude to R. K. University, Rajkot, for contribution the research infrastructure and academic support important for conducting this study.

Authors' Contributions

Dhimesh P. Parmar: Conceptualization, methodology, write.

Paresh J. Tanna: Supervision, validation, analysis, write review and edited.

Funding Information

This research received no external funding.

Ethics

This research does not include human participants or animals. The data used in this study were expanded from publicly accessible sources therefore, informed consent was not essential.

Declarations

Conflict of Interest: All authors publicize that they have no conflicts of interest.

References

Ahmed, M. H., Tiun, S., Omar, N., & Sani, N. S. (2022). Short Text Clustering Algorithms, Application and Challenges: A Survey. *Applied Sciences*, *13*(1), 342. https://doi.org/10.3390/app13010342

Chen, L., Pan, G., & Ng, Y. (2022a). Combining TF-IDF and BERT embeddings for robust short-text clustering. *Information Processing & Management*, *59*(2), 102851.

Chen, Y., Huang, H., & Ming, A. (2022b). A comparative study of deep embeddings for short-text clustering. *Knowledge-Based Systems*, *239*, 107970.

Costa, A., & Almeida, J. (2022). Density-based clustering methods for social media text. *Applied Intelligence*, *52*(5), 5125–5142.

Dubey, A., & Verma, P. (2022). Evaluation practices in clustering: Pitfalls and recommendations. *Information Sciences*, *609*, 147–170.

Fan, Y., Shi, L., & Yuan, L. (2023). Topic modeling methods for short texts: A survey. *Journal of Intelligent & Fuzzy Systems*, *45*(2), 1971–1990. https://doi.org/10.3233/jifs-223834

Gu, Q., Wang, Y., Wang, P., Li, X., Chen, L., Xiong, N. N., & Liu, D. (2024). An improved weighted ensemble clustering based on two-tier uncertainty measurement. *Expert Systems with Applications*, *238*, 121672. https://doi.org/10.1016/j.eswa.2023.121672

Hussain, S. F., Khan, K., & Jillani, R. (2022). Weighted multi-view co-clustering (WMVCC) for sparse data. *Applied Intelligence*, *52*(1), 398–416. https://doi.org/10.1007/s10489-021-02405-3

Kim, J., & Park, H. (2023). Spectral and graph-based clustering methods for short text. *Pattern Recognition Letters*, *168*, 144–151.

Pires, R., & Barbosa, R. (2024). Consensus clustering for short text: Methods and validation. *Knowledge-Based Systems*, *279*, 110949.

Pratama, M. A., & Lestari, A. (2023). Hybrid and ensemble approaches for short-text clustering. *Applied Soft Computing*, *138*, 110200.

Rodríguez-Ibáñez, M., Casado-Lumbreras, C., Colomo-Palacios, R., & García-Peñalvo, F.-J. (2023). A systematic review of sentiment analysis on social media platforms. *Expert Systems with Applications*, *228*, 120414.

Wu, G., Cao, L., Tian, H., & Wang, W. (2022). A hybrid parallel DBSCAN clustering algorithm scalable on distributed-memory computers. *Journal of Parallel and Distributed Computing*, *168*, 57–69. https://doi.org/10.1016/j.jpdc.2022.06.005

Wu, J., Tang, J., & Xue, Y. (2021). An ensemble framework for document and short-text clustering. *Information Sciences*, *581*, 659–674.

Yang, Z., Chen, K.-Y., Bae, M., & Keum, J. (2021). Short-text topic modeling: A comparative study. *Information Processing & Management*, *57*(6), 102307.

Zhang, M. (2022). Weighted clustering ensemble: A review. *Pattern Recognition*, *124*, 108428. https://doi.org/10.1016/j.patcog.2021.108428

Zhao, X., & He, L. (2023). Sentiment-aware topic modeling for Twitter data. *Social Network Analysis and Mining*, *13*(1), 76.