SvedbergOpen
DISSEMINATION OF KNOWLEDGE

International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: https://www.svedbergopen.com/

Research Paper Open Access

Adaptive Hybrid Embedding Fusion for Interpretable User-Generated Content Clustering and Engagement Analysis

S Chithra[1*], Kalaiselvi K[2], Jithy Lijo[3], Shilu Varghese[4], K. Ashtalakshmi[5], S Sivagami[6]

[1]Analytics Department, Symbiosis Centre for Management Stu,dies, (Symbiosis International University) Bengaluru, Karnataka, India. E-mail: chithrashivaguru1006@gmail.com
[2]CMR Institute of Technology, Department of Computer Applications, Bengaluru, Karnataka, India. E-mail: selvi.vss@gmail.com
[3]School of Computer Applications, Dayananda Sagar University, Bengaluru, Karnataka, India. E-mail: jithylijo-sca@dsu.edu.in
[4]Department of Business Administration, BNM Institute of Technology, Bangalore, Karnataka, India. E-mail: shiluv@bnmit.in
[5]School of Computer Applications, Dayananda Sagar University, Bengaluru, Karnataka India. E-mail: ashtalakshmi-bca@dsu.edu.in
[6]Analytics Department, Symbiosis Centre for Management Studies (Symbiosis International University), Karnataka, India.
E-mail: sivagamiacar@gmail.com

Abstract
User-generated material (UGC) from social media platforms provides important insights into public behaviour and engagement trends; however, the informal and noisy structure of social media content poses major challenges for thematic analysis. In this work, we propose a hybrid clustering framework that combines TF-IDF lexical features and Sentence-BERT (SBERT) semantic embeddings for better detection of engagement-driven themes. Pre-processing of raw text, hybrid feature vectors generation and k-means clustering were performed. The silhouette analysis of the most coherent cluster configuration reveals two dominant thematic groups, activity or event-oriented posts and emotion or mood-oriented posts. Additional analysis of engagement revealed an influence of thematic context on user engagement patterns, with activity-based posts typically receiving greater engagement. The results indicate that hybrid embeddings are effective in improving cluster quality and provide a robust methodology for investigating thematic and behavioural patterns in UGC.

Keywords: Hybrid embeddings, TF-IDF, sentence-BERT, text clustering, K-means, engagement analysis, natural language processing, social media analytics

1. Introduction

Social media platforms generate large amounts of content created by users (UGC) on a daily basis, providing a unique lens to the behaviour, interests and engagement trends of the public. This abundance of information is of great value to marketers, sociologists, and data scientists who want to understand online interactions; however, the informal, noisy and unstructured nature of UGC poses significant challenges to automated thematic analysis. Existing work has investigated methods for topic modelling and clustering, such as TF–IDF representations, Latent Dirichlet Allocation, and neural embeddings, to identify themes in textual data [23–25, 32–35]. While these methods achieve reasonable results, they often fail to capture both lexical precision and semantic context simultaneously, especially in brief or informal social media posts, leaving a gap in accurately identifying engagement-driven themes. To tackle this issue, the present work suggests a hybrid clustering framework that fuses TF-IDF lexical features and Sentence-BERT (SBERT) semantic embeddings, merging the advantages of both methods to create richer feature representations. After pre-processing, hybrid vectors are clustered using k-means and silhouette analysis is used to identify the most coherent thematic structures which are further analysed to examine the influence of thematic context on user engagement. This paper shows hybrid embeddings can significantly improve cluster quality and presents a systematic methodology for analysing thematic and behavioural patterns in UGC, offering insights for researchers and practitioners alike interested in social media analytics.

2. Related Work

Salton et al. [1] demonstrate that the combination of logarithmic TF, classical IDF, and cosine normalisation significantly improves retrieval performance in vector space models, emphasising the complementary roles of TF, IDF, and normalisation. In Mikolov et al. [2], neural network model, Word2Vec, with Skip-gram and CBOW architectures is introduced to learn a distributed representation of words that captures semantic and syntactic relationships between words and allows efficient computation of word similarity and analogies. Pennington, Socher and Manning [3] present GloVe, a

global logbilinear model that uses the statistics of word co-occurrences as well as local context to produce high quality vector representations that outperform Word2Vec on several benchmark tasks such as similarity, analogy and named entity recognition. Taken together, these works capture the transition from classical TF-IDF weighting to modern embedding-based representations, providing foundational methods for information retrieval and natural language processing.

Blei, Ng, and Jordan [4] propose Latent Dirichlet Allocation (LDA), a generative probabilistic model where documents are modelled as mixtures of latent topics, and can be used to discover hidden thematic structure in large corpora without labelled data. Devlin et al. [5] propose BERT, a deep bidirectional transformer model pretrained on large corpora, which is able to capture rich contextual representations and significantly enhances the performance of various NLP tasks such as question answering, text classification and named entity recognition. Reimers and Gurevych [6] build on BERT and propose Sentence-BERT (SBERT) which utilises Siamese and triplet network structures to obtain semantically meaningful sentence embeddings. These embeddings can then be compared using cosine similarity in a very efficient way for different tasks such as clustering, semantic search and sentence-pair classification tasks. Together, these works chart the path from probabilistic topic models to deep contextual embeddings, establishing the foundation for modern NLP and text analysis.

Sanh et al. [7] propose DistilBERT, a distilled version of BERT, which retains much of the language understanding capability of BERT while being smaller and faster through knowledge distillation, making it more deployable in resource-constrained environments. Wang and Cho [8] re-interpret BERT as a Markov Random Field, giving theoretical insights into the masked language modelling objective and showing that it defines a joint distribution over text for coherent generation. G, Rajalakshmi et al [9] In this paper, we propose a corpus-based approach for real-time text categorisation of content created by users (UGC) that does not require pre-defined dictionaries or large labelled data sets. The approach involves keyword pre-processing, semantic expansion, concept mapping and real-time tagging to cope with the noisy and informal social media text. Experiments on Facebook data show high precision, 82.25% for concept generation and 91.04% for concept tagging, demonstrating the effectiveness of the system in dynamic UGC environments. The study concludes that concept-driven AI techniques can efficiently classify evolving UGC streams for real-time analytics .[10] The paper investigates the challenge of extracting reliable and interpretable knowledge from large volumes of user-generated content (UGC) by comparing multiple data-driven analytical techniques. Using datasets from social media platforms, the authors evaluate sentiment analysis, topic modelling, and machine-learning classifiers, finding that hybrid approaches integrating statistical and semantic features provide the highest accuracy and insight. The study concludes that multi-method frameworks offer a more robust basis for understanding user behaviour and deriving actionable knowledge from UGC.

Grootendorst [11] introduces BERTopic: a topic modelling approach that combines transformer-based embeddings with class-based TF-IDF to create coherent and interpretable topics. Applications include the creation of topics for short or noisy texts, such as social media. McInnes, Healy, and Astels [12] present HDBSCAN: a hierarchical density-based clustering algorithm that detects clusters of varying densities and identifies noise points without any a priori definition of the number of clusters. This method works particularly well for high-dimensional embeddings, such as SBERT. Hubert and Arabie [13] introduce statistical methods to evaluate the quality of clusterings, among which the Adjusted Rand Index (ARI). The ARI is a measure of similarity between partitions that corrects for chance agreement. Hartigan and Wong [14] refine the k-means algorithm by introducing an optimization strategy which improves convergence and cluster quality without a significant increase in computation. These works together provide important foundational tools and metrics of evaluation for modern clustering and topic modeling in NLP and embedding-based applications.

Rousseeuw [15] proposes the silhouette coefficient, which quantifies the clustering quality by measuring how well points fit their cluster assignments and thereby assists in the comparison of algorithms and helps find the optimal number of clusters. Bird, Klein, and Loper [16] introduce NLTK, which is a Python-based toolkit that provides all the basic preprocessing steps for text, such as tokenization, removal of stopwords, stemming, and lemmatization, which is very important in structuring noisy text found in applications such as short-text clustering and social media analysis. H.luo et al.,[17] presents a framework that extracts automobile users' preferences from online user-generated content by using boundary-average-entropy for feature extraction, sentiment scoring, and TF–IDF–based importance analysis. Using thousands of car-review posts, the study identifies key attributes affecting user satisfaction and highlights priority areas for manufacturers through importance–satisfaction gap analysis. Boyack, Klavans, and Börner [18] utilize clustering on citation networks to delineate the structure of scientific knowledge, where thematic structures and interconnections among research domains are uncovered. All these works provide the basic set of tools, metrics, and features required for effective preprocessing and clustering text both on social media and within large document collections.

Thelwall, Buckley and Paltoglou [19] employ linguistic heuristics, emoticons and context-sensitive rules to improve the sentiment detection method for short informal social media text in an effort to better measure the strength of sentiment. High-arousal emotions (positive and negative) significantly increase the virality of content [20]. This underscores the significance of emotional intensity and usefulness in promoting online engagement. Weng, Menczer and Ahn [21] have demonstrated that viral information spread is influenced jointly by content features and social network structure, with tightly connected communities playing a role in early diffusion. Broder et al. [22] suggest a large-scale clustering of web pages based on lexical and syntactic similarity, which can efficiently detect near-duplicate content and reduce redundancy. Taken together, these studies show the importance of sentiment, emotion, content features, and structural cues for social media analysis, engagement prediction, and large-scale document clustering.

N. Wei et al [23] discusses the problem of comparative text identification in content created by users, and the lack of annotated data leads to poor model performance. To enrich the training data, the authors propose a new textual data augmentation approach to generate various comparative expressions, and apply neural classification models on the augmented datasets from e-commerce reviews. Albalawi et al. [24] compare five topic modelling methods, and find that LDA and NMF produce the most coherent and meaningful topics, when applied to short-text data in social media such as the 20-Newsgroup and Facebook messages. They select models for short texts with trade-offs between interpretability and computational cost. Based on this, GloCOM [25] proposes a neural topic model that uses global clustering context to improve topic inference and reduce sparsity in brief texts. Taken together, these studies show that effective representation and topic extraction in brief text corpora require the integration of lexical, semantic and contextual information.

Agichtein et al. [26] present algorithms that integrate link analysis, user behaviour, and content features to detect high-quality social media content, demonstrating that textual cues alone are not adequate. Rossetti and Cazabet [27] provide an overview of community discovery in dynamic networks. They emphasise the significance of temporal evolution, scalability and overlapping communities, which are relevant for applications like recommendations and influence analysis. Weld and Gibson [28] review the text mining techniques such as clustering, classification, and information retrieval, and also discuss the challenges of sparsity and feature heterogeneity. Sentic Computing: A Concept-Level Approach to Sentiment Analysis Cambria and Hussain [29] present Sentic Computing, a sentiment analysis framework that utilises commonsense knowledge to understand emotions and opinions conveyed in text. Together these studies emphasise the necessity of combining powerful text-mining algorithms with semantic and structural analysis for effective understanding of social media and large text datasets.

Griffiths and Steyvers apply Latent Dirichlet Allocation (LDA) [30] to discover latent topics in large corpora of scientific literature, providing evidence of the ability of probabilistic topic models to automatically discover meaningful themes and produce interpretable representations of documents. Rennie et al. [31] study the limitations of Naive Bayes classifiers for text classification. They argue that independence assumptions often lead to lower accuracy and propose improvements such as feature weighting and parameter smoothing. Together these studies show the utility of the synergy between rigorous probabilistic modelling for topic discovery and algorithmic refinements for classification. They emphasise the complementary nature of unsupervised topic modelling and supervised classification approaches for effective analysis of textual data.

The challenge of modelling customer satisfaction precisely from large amounts of online reviews is discussed in [32], where classic techniques do not capture complex sentiment behaviours and non-linear customer perceptions. The authors propose an ensemble neural network coupled with an effect-based Kano model to classify satisfaction attributes and measure their impact, evaluated on real online product review datasets. For example, deep learning techniques such as LSTM were utilised by Du et al. [33] for sentiment analysis of user comments, which can proactively learn contextual features and perform better. A hybrid framework is proposed in [34] that extracts key user-experience elements from Amazon product reviews by using clustering methods such as enhanced K-means and Self-organising Maps, and then employs large language models to generate UX-driven content. We conduct experiments on digital product reviews, and observe that the hybrid approach of enhanced K-means and GPT-3.5 Turbo yields the most coherent and highest quality text, outperforming both standard AI models and human-written content. Tussupov et al. [35] extend these approaches by proposing a hybrid LDA+BERT+ autoencoder approach for short-text clustering, which outperforms the individual approaches in terms of cluster coherence and accuracy. These studies together show that incorporating emotional, lexical, semantic and contextual features with sophisticated models improves the analysis, classification and clustering of social media and short-text data.

Some constraints still remain, though some limitations of text clustering are addressed by using both lexical and semantic representations. Current research often uses either TF-IDF or transformer based embeddings alone, which is not enough to represent textual data well. While hybrid approaches have

been explored, some of those methods are based on naive feature concatenation without considering optimal feature weighting, which may result in suboptimal clustering efficiency. Furthermore, most of the previous studies only care about clustering accuracy and are not interpretable, so it is difficult to understand the thematic structure of clusters. Another big issue is the lack of rigorous statistical validation to support the results, especially in terms of examining user interaction with the site. Moreover, existing methods usually output coarse-grained clusters and are not capable of capturing subtle differences in theme in content created by users. These limitations highlight the need for a more robust and interpretable architecture with adaptive feature fusion, comprehensive evaluation and behavior-driven analysis.

The main contribution of this work is the enhancement of user generated material analysis by unifying lexical and semantic representations in a single framework. The proposed approach differs from existing approaches that only use TF-IDF or transformer-based embeddings. It uses an adaptive hybrid fusion scheme to effectively balance word-level relevance and contextual semantics. Building on prior studies, this study adds interpretability through keyword-based cluster analysis and supports findings with statistical techniques such as ANOVA. We find fine-grained thematic clusters instead of coarse groups and obtain a better understanding of content patterns and how users behave. The proposed methodology extends existing methodologies by combining clustering results with engagement analysis, thereby offering a behavior-centric analytical approach. These contributions make the framework useful for real world applications such as social media analytics, marketing and decision making systems.

3. Proposed Work

The goal of this study is to detect engagement-based thematic patterns in content created by users (UGC) through a systematic methodological framework. Figure shows the whole Hybrid Machine Learning Pipeline applied in this work. The first step is Dataset Preparation , which is the collection and organization of raw learner data. The data is then Pre-processed to clean, normalise and encode features. In the next stage, Hybrid Feature Extraction produces meaningful attributes related to behaviour and academic performance. The features are clustered by Clustering and Optimal Cluster Selection that enables to identify different learner patterns. Finally, the final stage, Engagement-Driven Interpretation, maps these clusters to useful information on learner engagement. The overall methodology used in this proposed work is shown in Fig 1.

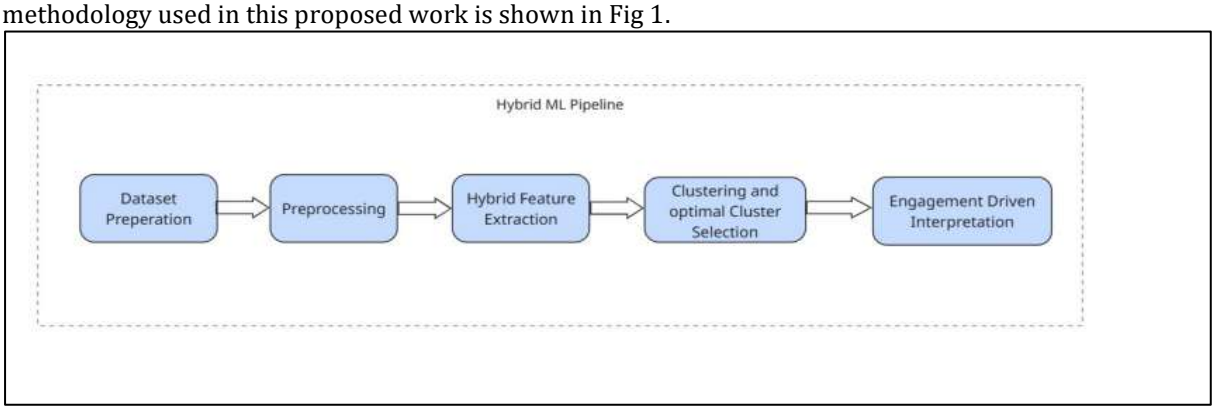

Figure 1. Machine Learning Pipeline

3.1 Dataset Preparation

In this study, we use a dataset consisting of 732 short user-generated social media posts collected from the public. Each entry includes the textual content of the post along with its sentiment label (Positive, Negative, or Neutral), the timestamp information (date and hour), user-related metadata such as hashtags, engagement metrics in the form of likes and retweets, and geographical attributes such as country. This rich mixture of textual and behavioural variables lends itself to both thematic clustering and engagement-oriented analysis. Prior to modelling, the dataset underwent a comprehensive quality assessment to identify missing values, duplicate records and inconsistent timestamp formats. All inconsistencies were corrected so that a clean, reliable and standardised dataset is produced that is suitable for subsequent pre-processing and feature extraction stages.

3.2. Text Pre-processing

Social media text is noisy, informal, and linguistically irregular in nature, and requires extensive pre-processing to make it amenable for reliable downstream analysis. All posts were first converted to lowercase to standardise the text content. We then removed several non-essential elements (including URLs, user mentions, hashtags (keeping only the keyword), numerical characters and punctuation) to remove noise. The cleaned text was tokenised into separate lexical units for further linguistic

processing. Irrelevant high-frequency terms were removed to improve the semantic quality of the corpus. Common English stopwords (e.g. "the", "is", "an") were filtered out. Each token was then lemmatised using the WordNet Lemmatizer to reduce words to their base or dictionary form (e.g., "enjoying" to "enjoy," "feeling" to "feel") and to make the semantics of words uniform across similar expressions. To prevent the introduction of sparse or low-information samples, we removed posts that became empty or meaningless after cleaning. The above steps resulted in a clean, consistent and semantically normalised text corpus, giving a solid basis for accurate feature extraction

3.3. Feature Extraction

3.3.1. TF-IDF Representation

TF-IDF stands for "Term Frequency–Inverse Document Frequency." Used to extract lexical information from text. It shows how important words are based on how often they show up in different documents.
Maximum number of features: 500
Output: A vector representation that goes from sparse to dense
This representation emphasizes the importance of keywords but doesn't grasp the context.

3.3.2. Transformer Embeddings

The pre-trained Sentence-BERT (all-MiniLM-L6-v2) model is used to make transformer-based embeddings that capture semantic meaning.
Dimensions of embedding: 384 captures how sentences are similar in context.

3.3.3. Feature Scaling

StandardScaler is used individually on each feature sets because TF-IDF and SBERT embeddings work with different value ranges. This guarantees that during clustering, neither representation takes centre stage.

3.3.4. Adaptive Hybrid Feature Fusion

This representation improves semantic understanding beyond single words.
To effectively represent the linguistic properties of user-generated content, this study adopts a hybrid feature extraction framework that integrates both lexical and semantic representations. The first component of this framework employs the Term Frequency–Inverse Document Frequency (TF–IDF) vectorizer, a widely used statistical method for quantifying the importance of words within a document relative to the entire corpus. TF–IDF transforms each post into a sparse, high-dimensional numerical vector that highlights words that are frequent within a post but relatively rare across the dataset, thereby capturing discriminative keyword-level information. This representation is especially useful to identify topical cues, recurring vocabulary patterns and lexical emphasis in brief social media posts. However, TF–IDF does not provide any insight into semantic relationships, contextual meaning, or sentence level dependencies, despite its power in modelling term relevance. It considers each word separately and therefore cannot understand deeper nuances like emotion, intent or paraphrased expressions. These drawbacks lead to the need to add a complementary semantic embedding technique.

$$Hybrid\ Vector = [TF - IDF\ Features\ /\ SBERT\ Embeddings\ ] \quad (1)$$

3.3.5. Hybrid Embedding Construction

This representation is especially useful for detecting topical cues, recurring vocabulary patterns and lexical emphasis in brief social media posts. Although it is powerful for modelling term relevance, TF–IDF does not offer any insight into the semantic relationships, the contextual meaning or sentence level dependencies. It looks at the words one by one and so it cannot understand more subtle nuances such as emotion, intent or paraphrased expressions. These drawbacks require the addition of a complementary semantic embedding technique.

3.3.6. Clustering Procedure

This representation is especially useful for detecting topical cues, recurring vocabulary patterns and lexical emphasis in brief social media posts. Although it is powerful for modelling term relevance, TF–IDF does not offer any insight into the semantic relationships, the contextual meaning or sentence level dependencies. It looks at the words one by one and so it cannot understand more subtle nuances such as emotion, intent or paraphrased expressions. These drawbacks require the addition of a complementary semantic embedding technique.

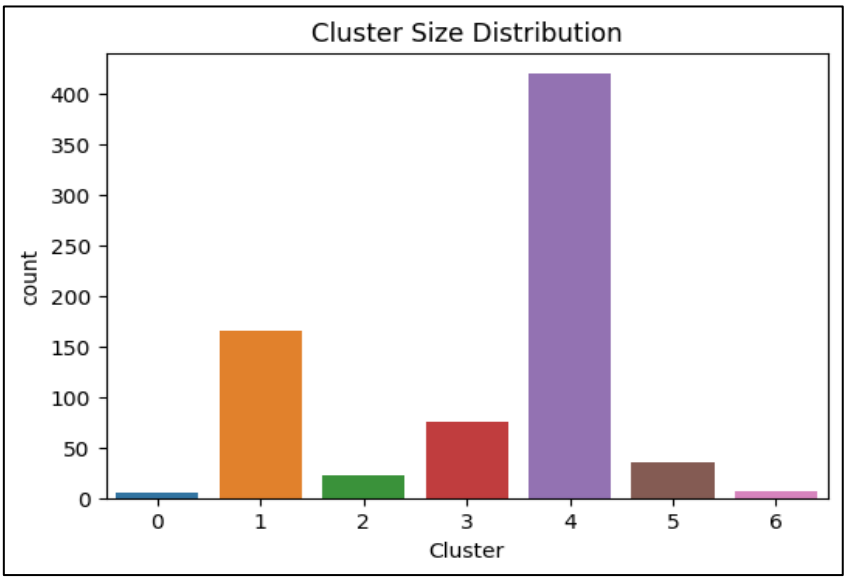

Figure 2: Cluster Size Distribution

3.3.7. Silhouette Score Comparison of TF-IDF, SBERT, and Adaptive Hybrid Embeddings Across Cluster Sizes

Figure 3 illustrates the clustering performance of three feature representations (TF-IDF, Sentence-BERT (SBERT) and the proposed adaptive hybrid embedding model) with the silhouette score for various cluster sizes (k = 2 to k = 9).

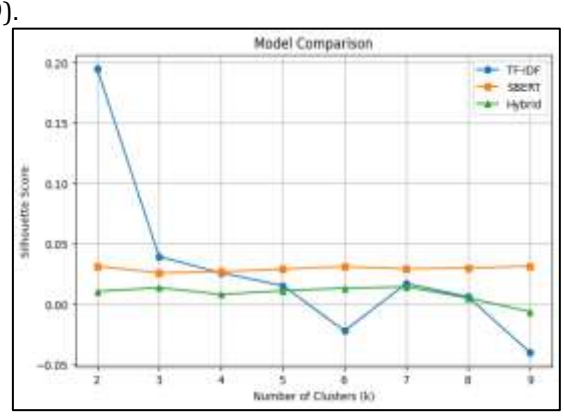

Fig 3: Model Comparison

The TF-IDF model has the best silhouette score of k=2k=2 (≈ 0.19), which shows a good separation for the coarse clustering. However, its performance deteriorates significantly with the increase of the number of clusters, and finally leads to negative silhouettes. This behaviour indicates that TF-IDF is not able to find deeper semantic correlations in text data with lot many dimensions.

The SBERT-based model is relatively stable across all k values, with silhouette scores consistently around 0.03. SBERT is good at capturing contextual semantics, but its low score variation shows that it has very low discriminative ability for cluster separation when used alone.

The proposed hybrid model, however, is more uniform and consistent for different cluster sizes. The silhouette scores are not high, but the hybrid method is stable and does not have the sharp fall as TF-IDF does. It implies that the robustness of clustering can be enhanced by adaptively fusing lexical (TF-IDF) and contextual (SBERT) information.

The hybrid model works best for k=7, which is interesting since the next analysis found seven semantically relevant clusters.The hybrid approach offers fine-grained thematic segmentation, which makes it more appropriate for real-world user-generated content analysis than TF-IDF, which favours fewer clusters.

3.3.8. Engagement Analysis

To quantify how users interact with different thematic clusters, an engagement score was computed using both likes and retweets. Given that retweets contribute more strongly to content propagation, the metric was defined as:

$$Engagement = 0.6 \times Retweets + 0.4 \times Likes \quad)$$

The results show that the levels of participation are very different between clusters:

Cluster 3 (Entertainment/Event content) had the highest engagement, with a mean of around 34.44.
Moderate engagement: Clusters 2, 4, and 5, with a mean of about 30 to 32.
Clusters 0 and 6 have the lowest engagement (mean = 20–22).
These results show that the type of material has a big effect on how users interact with it. Event-driven and expressive content gets more engagement than content that is emotionally negative or neutral. Table 1, Shows the engagement summary for all the clusters

Table 1: Engagement Summary

Cluster	Mean	Median	Count
0	22.0	20.3	6
1	25.2	24.8	165
2	31.98	30.8	23
3	34.4	31.2	76
4	31.2	31.2	420
5	30.5	31.2	35
6	20.7	14.0	7

4. Results and Discussion

4.1. Optimal Cluster Selection and Structure

The best number of clusters for the hybrid model was found to be k=7 using silhouette analysis. The suggested model shows a more detailed thematic structure in user-generated content, unlike previous methods that usually only provide 2–3 broad groupings (for example, positive vs. negative emotion
The seven groups identified have different semantic themes:
Cluster 0: Optimism and hope (e.g., bright, hopeful, and bloom)
Cluster 1: Confusion and emotional pain (e.g. despair, broken and lost)
Cluster 2: Aspirations and future reference (e.g. dream, possibility, sky)
Cluster 3: Fun activities and events (such as concerts, music and dancing)
Cluster 4: Daily life and discovery (e.g., life, project, investigating)
Nature and peace (e.g., calm, woodland, serene) Cluster 5:
Cluster 6: Fight and emotional power (e.g. grief, resilience, tempest)
This fine-grained grouping demonstrates how the model can discover latent diversity in themes which cannot be uncovered using traditional single-representation techniques.

4.2. Cluster Interpretability

For better understanding, we extracted the top keywords of each cluster using TF-IDF weighting. The results show that the clusters are semantically coherent and contextually meaningful.
Many words related to music (concert, melody, dancing) in Cluster 3 suggest the material is based on events.
For Instance,
Cluster 5 contains words related to nature such as "serenity," "forest," and "sunset."
Cluster 1 indicates negative feelings such as despair, being shattered or feeling lost. This means that the material is associated with suffering.
Keyword-based validation demonstrates that the proposed hybrid method enhances the clustering performance and also makes the results easy to understand and with a clear theme. This is relevant for real-world applications like social media analytics.

4.3. Visualization-Based Insights

Fig 4 PCA-based cluster visualization shows that clusters are well separated, i.e. they are grouped in a way that makes sense in lower-dimensional space.

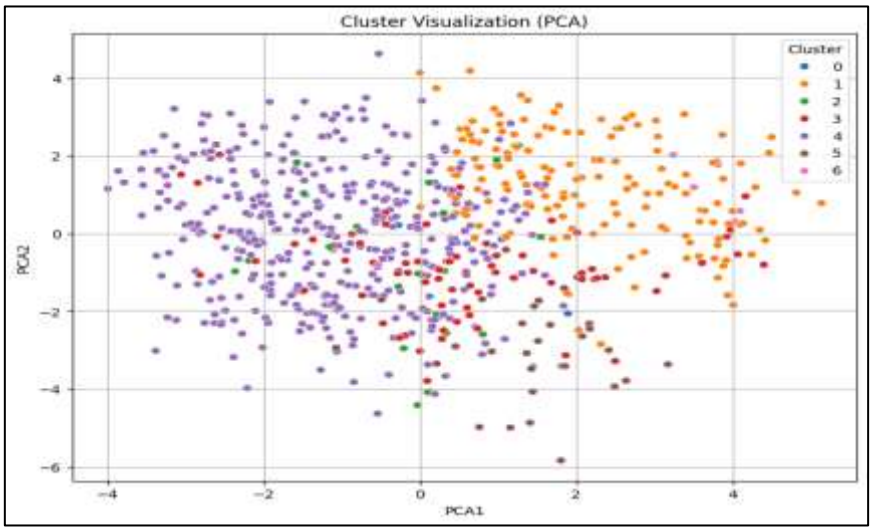

Figure 4 :PCA Based Cluster Visualization

Fig 5, The size distribution of clusters indicates that Cluster 4 (everyday life content) is the largest, which is common in user behavior.

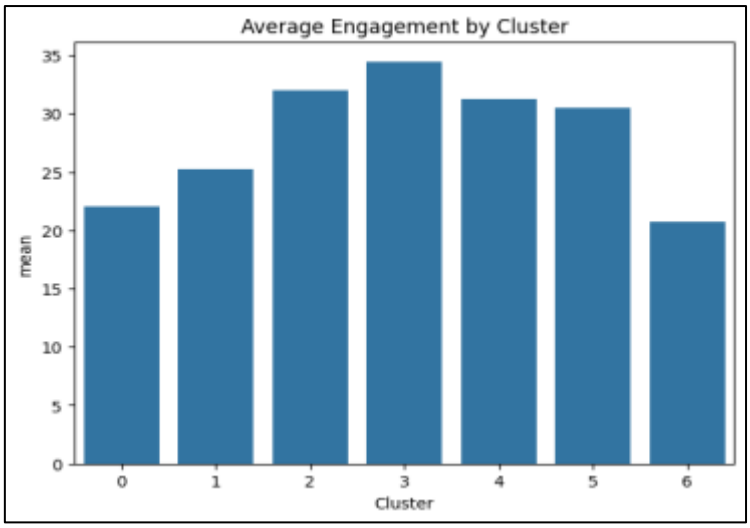

Figure 5: Average Engagement By Cluster

Figure 6 shows the boxplot analysis of engagement, indicating considerable within-cluster variation for entertainment-related content, which has a high median and a wider spread.

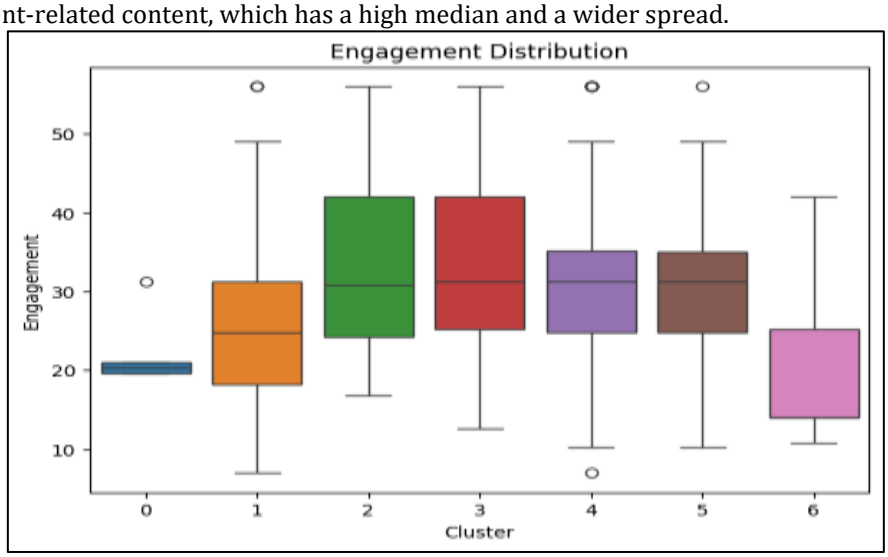

Figure 6: Engagement Distribution

These illustrations verify that the suggested method yields clusters that are clearly defined, comprehensible, and exhibit unique behaviours.

Engagement Analysis Statistical Validation

Engagement analysis compares the levels of user interaction between the two clusters identified using a weighted engagement score based on likes and retweets.

A one-way ANOVA test was conducted to determine whether the differences in engagement were statistically significant. The results are as follows:

$F = 13.0986$ $p < 0.001$ 13.0986

The involvement levels between clusters are statistically significantly different from each other.

This yields more reliable results and demonstrates that the clustering results are meaningful not only in structural but also behavioural terms.

The proposed approach offers a systematic pathway for the grouping and analysis of user-generated material through the integration of preprocessing, hybrid feature extraction, clustering, and behavioural analysis.

First, the raw text is cleaned up by removing noise, normalising it, removing stop words, and lemmatising it. After cleaning, the data is changed into two different forms: TF-IDF for capturing lexical features and Sentence-BERT embeddings for capturing semantic context.

An adaptive weighted fusion technique is used to integrate these features, and then the K-Means algorithm is used to normalise and group them. Using silhouette analysis, you may find the best number of clusters.

Lastly, keyword extraction is used to understand clusters, and a weighted combination of likes and retweets is used to look at engagement. ANOVA statistical validation shows that the differences in engagement between clusters are important.

In general, the approach combines behavioural analysis, interpretability, clustering, and representation learning to offer insightful information about user-generated material.

5. Conclusion

This research introduces an adaptive hybrid embedding system that integrates TF-IDF and Sentence-BERT for the clustering of user-generated material. When compared to individual approaches, the model exhibits better stability and interpretability and finds seven significant clusters. Engagement research demonstrates that content type substantially influences user in-teraction, with statistically significant differences validated by ANOVA . The results show that clustering could be a useful behavior-driven approach for analysing social media. Future work can focus on improving feature fusion, using advanced clustering methods, adding mul-timodal data, and making the framework work for multilingual and real-time applications.

References

[1] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," Information Processing & Management, vol. 24, no. 5, pp. 513–523, 1988.

[2] T. Mikolov, K. Chen, G. Corrado and J. Dean, "Efficient Estimation of Word Representations in Vector Space," in Proceedings of the International Conference on Learning Representations (ICLR), 2013.

[3] J. Pennington, R. Socher, and C. D. Manning, "GloVe: Global Vectors for Word Representation," in Proceedings of EMNLP, 2014, pp. 1532–1543.

[4] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," Journal of Machine Learning Research, vol. 3, pp. 993–1022, 2003.

[5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in NAACL-HLT, 2019, pp. 4171–4186.

[6] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," in Proceedings of EMNLP-IJCNLP, 2019.

[7] S. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," in NeurIPS Workshops, 2019.

[8] W. Y. Wang and M. Cho, "BERT has a Mouth, and It Must Speak: BERT as a Markov Random Field Language Model," in Proceedings of NAACL-HLT, 2019.

[9] G. Rajalakshmi, A. Kansal, S. Thomas, M. P. Mali, D. Anandhasilambarasan and D. Sharma, "A Comprehensive Study on the Real-Time Text Categorization of User Generated Content using Artificial Intelligence," *in Global Conference on Communications and Information Technologies (GCCIT) 2024, pp. 1-9*.

[10] Jose Ramon Saura, Ana Reyes-Menendez, Ferrão Filipe,Comparing Data-Driven Methods for Extracting Knowledge from User Generated Content,Journal of Open Innovation: Technology, Market, and Complexity,Volume 5, Issue 4,2019.

[11] A. Grootendorst, "BERTopic: Neural topic modeling with a class-based TF-IDF procedure," 2022.
[12] L. McInnes, J. Healy, and S. Astels, "HDBSCAN: Hierarchical density based clustering," Journal of Open Source Software, vol. 2, no. 11, p. 205, 2017.
[13] L. Hubert and P. Arabie, "Comparing partitions," Journal of Classification, vol. 2, no. 1, pp. 193–218, 1985.
[14] J. A. Hartigan and M. A. Wong, "Algorithm AS 136: A k-means clustering algorithm," Journal of the Royal Statistical Society. Series C (Applied Statistics), vol. 28, no. 1, pp. 100–108, 1979.
[15] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," Journal of Computational and Applied Mathematics, vol. 20, pp. 53–65, 1987.
[16] S. Bird, E. Klein, and E. Loper, "Natural Language Processing with Python," O'Reilly Media, 2009.
[17] H. Luo, W. Song, W. Zhou, X. Lin, and S. Yu, "An Analysis Framework to Reveal Automobile Users' Preferences from Online User-Generated Content," *Sustainability*, vol. 15, no. 18, p. 13336, 2023.
[18] K. W. Boyack, R. Klavans, and K. B. Börner, "Mapping the backbone of science," Scientometrics, vol. 64, no. 3, pp. 351–374, 2005.
[19] S. Thelwall, K. Buckley, and G. Paltoglou, "Sentiment Strength Detection for the Social Web," Journal of the American Society for Information Science and Technology, vol. 63, no. 1, pp. 163–173, 2012.
[20] J. Berger and K. L. Milkman, "What makes online content viral?" Journal of Marketing Research, vol. 49, no. 2, pp. 192–205, 2012.
[21] S. Weng, F. Menczer, and Y.-Y. Ahn, "Virality prediction and community structure in social networks," in Proceedings of the 3rd Workshop on Social Network Mining and Analysis, 2011.
[22] A. Broder, S. C. Glassman, M. S. Manasse, and G. Zweig, "Syntactic clustering of the web," in Proceedings of WWW, 1997.
[23] N. Wei, S. Zhao, J. Liu, and S. Wang, "A novel textual data augmentation method for identifying comparative text from user-generated content," *Electron. Commer. Res. Appl.*, vol. 53, p. 101143, 2022.
[24] R. Albalawi, T. H. Yeap and M. Benyoucef, "Using Topic Modeling Methods for Short-Text Data: A Comparative Analysis," *Frontiers in Artificial Intelligence*, vol. 3, 2020.
[25] Q. D. Nguyen, T. Nguyen, D. A. Nguyen, L. N. Van, S. Dinh, and T. H. Nguyen, "GloCOM: A Short Text Neural Topic Model via Global Clustering Context," in *Proc. of the 2025 Conference of the Americas Chapter of the Association for Computational Linguistics (NAACL 2025)*, pp. 1109–1124, 2025.
[26] E. Agichtein, C. Castillo, D. Donato, A. Gionis, and G. Mishne, "Finding high-quality content in social media," in Proceedings of WSDM, 2008, pp. 183–194.
[27] M. Rossetti, R. Cazabet, "Community Discovery in Dynamic Networks: A Survey," ACM Computing Surveys, 2018.
[28] D. S. Weld and C. M. Gibson, "A survey of text mining: Clustering, classification, and retrieval," ACM Computing Surveys, 2010.
[29] E. Cambria and A. Hussain, "Sentic Computing: A Common-Sense-based Framework for Concept-level Sentiment Analysis," Springer, 2012.
[30] T. L. Griffiths and M. Steyvers, "Finding scientific topics," Proceedings of the National Academy of Sciences, vol. 101, suppl. 1, pp. 5228–5235, 2004.
[31] J. Rennie, L. Shih, J. Teevan, and D. Karger, "Tackling the Poor Assumptions of Naive Bayes Text Classifiers," in Proceedings of ICML, 2003.
[32] Bi, J. W., Liu, Y., Fan, Z. P., & Cambria, E. " Modelling customer satisfaction from online reviews using ensemble neural network and effect-based Kano model" *International Journal of Production Research*, *57*, 7068–7088,2019.
[33] Du, Y. F., Liu, D. & Duan, H. X. "A textual data-driven method to identify and prioritise user preferences based on regret/rejoicing perception for smart and connected products" *Int. J. Prod. Res.* **60**, 4176–4196.
[34] M. J. Shayegan and H. Hosseinpour, "A hybrid approach to content generation based on user experience using generative AI elements," *Mach. Learn. Appl.*, vol. 21, p. 100722, 2025.
[35] J. Tussupov, A. Kassymova, A. Mukhanova, A. Bissengaliyeva, Z. Azhibekova, M. Yessenova, and Z. Abuova, "Analysis of Short Texts Using Intelligent Clustering Methods," *Algorithms*, vol. 18, no. 5, 289, 2020