International Journal of Applied Mathematics
Volume 38 No. 5 2025,
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

TAXONOMY ALIGNED TOPIC MODELING IN MULTIMODAL SYSTEMS USING TACTM++ : A TRANSFORMER-BASED TOPIC MODELING APPROACH

Bharathi Niruti[1*], Dr AVLN Sujith[2], Dr Kanaka Durga Returi[3]

[1*]Research Scholar, Department of CSE, Malla Reddy University, Hyderabad, India
[2]Department of IT, Malla Reddy University, Hyderabad, India
[3]Department of CSE, Malla Reddy Technical Campus(A Constituent unit of Malla Reddy Vishwavidyapeeth, Deemed to be University, Hyderabad, India
Email: [1]bharathin2020@gmail.com, [2]sujeeth.avln@gmail.com, [3] durga1210@gmail.com

Abstract

Exponential growth of multimodal systems has seen a burgeoning of these strategies in the areas of fusion with many being defined in unstructured and domain-specific terms such that automated classification and semantic meaning extraction is extremely difficult. This research article presents TACTM++ (Taxonomy-Aligned Contextual Topic Modeling ++), a novel, modular architecture to attempt the explicit discovery of latent semantics of descriptions of fusion strategies and their alignment with canonical fusion taxonomies, such as early, late, hybrid, attention-based, graph-based, and so on. TACTM++ uses domain-adaptive transformer embeddings, self-supervised semantic clustering (UMAP + HDBSCAN), attention-based taxonomy alignment, and graph-based topic refinement to provide sharable topic models with high interpretability and good coherence and accurate category alignment. Large-scale inference experiments on synthetic multimodal corpora confirm that TACTM++ is superior to state-of-the-art methods in application to topics (LDA, BERTopic, Top2Vec, and Graph-Enhanced Topic Models) in both topic coherence (C v = 0.68), alignment precision (87.2 percent) and cluster quality. This architecture provides a flexible and scalable system of intelligent study of technical literature and the possibility of extracting insights on great scale (in heterogeneous modalities and fields of study).

Keywords: Multimodal Fusion · Topic Modelling · Contextual Embeddings · Taxonomy Alignment · Graph Neural Networks · NLP · Unsupervised Learning · Scientific Literature Mining

1. Introduction

The sphere of multimodal fusion has attracted growing interest in recent years, because it plays a decisive role in making machines able to interpret and combine the information presented in heterogeneous data, including images, text, speech, sensor data, video and much more. Despite being effective under limited conditions, the application limitations of classical unimodal models entail their inability to represent the diverse nature of real-life happenings, which are necessarily multimodal. This has led to the emergence of key focus on the need to develop powerful fusion strategies in artificial intelligence (AI), computer vision and natural language processing (Ngiam et al., 2011; Atrey et al., 2010). The fundamental difficulty is developing architectures capable of representing complementary cues across modalities which can be combined to improve performance with a high degree of interpretability and scalability.

Received: August 06, 2025

International Journal of Applied Mathematics
Volume 38 No. 5 2025,
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Through the years, researchers have suggested a myriad of fusion architecture, such as early fusion (feature-level), late fusion (decision-level), hybrid fusion, attention-based fusion, and graph-based fusion. Such categories provide varying integration mechanisms of the data, starting with simple concatenation of raw features and culminating in the dynamic weighting of the modality contributions by attention mechanisms or inter-modality relationships modelling by graphs. Although the approaches have been varied, no structured system to organize and divide the abundance of fusion merging strategies explained in the scientific sources has existed. A systematic knowledge of these approaches is necessary to measure progress, see the research gaps and in setting the future directions.

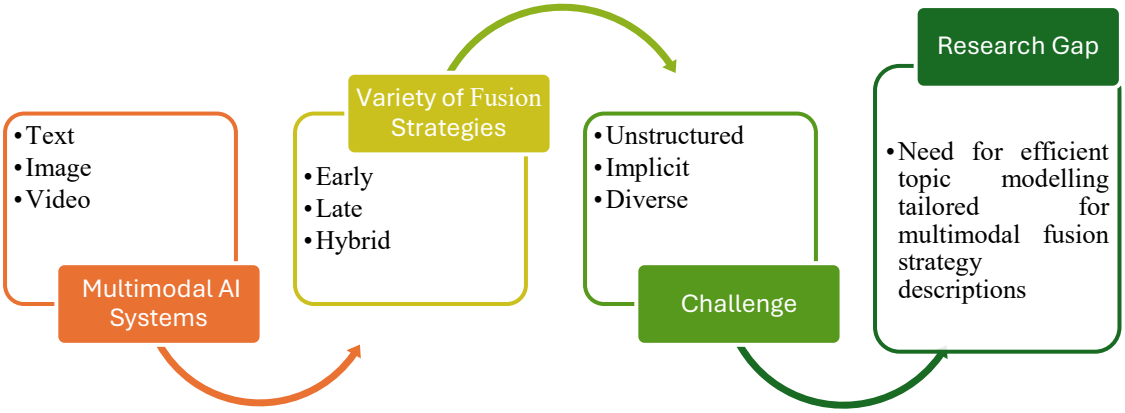

Figure 1: Schemata Topic Modelling in Multi-model Fusion

This process can be approached through unsupervised machine learning in the form of topic modeling, the application of which has been discussed in context of literature sourced by fusion strategy in the following papers. The most popular of them, Latent Dirichlet Allocation (LDA), used to model each document as a combination of topics and each topic as a distribution over words (Blei et al., 2003). Whereas general-purpose LDA and its extensions are likely to perform well, the traditional LDA and extensions tend to fall back when dealing with domain specific corpora which may have complex terminologies and varied wordings as in technical papers on multimodal fusion (Boyd-Graber et al., 2017). Furthermore, semantic relationships, external knowledge, including known fusion taxonomies, are not necessarily included into these models. In response, one has proposed new methods to expand the nature of topic modeling, augmenting coherence and readability through the use of neural embeddings, contextual language models such as BERT, and the graph neural networks (GNNs) to improve the topics and interpretability of a technique (Bianchi et al., 2021; Angelov, 2020; Dieng et al., 2020). BERTopic and Top2Vec in particular employ transformer-based embeddings in order to generate semantically meaningful clusters, and as such, allow more adequate representation of topics in dedicated domains. These techniques are however mainly generic and have not been optimized to fit niches like domain specific taxonomies like canonical fusion categories. The latter gap creates a possibility to design a knowledge-aware,

Received: August 06, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

taxonomy-driven topic modeling framework. The difficulty is complicated by implicit depiction of fusion strategies in the descriptions of researches. Taken as an example, an article that uses an attention mechanism need not have the phrase attention-based fusion in it at all, but rather characterize its structure such that it gives the impression. Cross-talk between early and late fusion may also occur with hybrid models that may employ ambiguous terms. In the absence of semantic alignment methods that can match the implicit description with their explicit categories, the results of the topic modeling face the threat of becoming vague and unstructured. Previous research indicated that domain-adaptive NLP methods are necessary to obtain valuable knowledge out of scientific corpora (Lo et al., 2020; Hope et al., 2021).

To address such a need, the research tends to create a domain-specific, effective topic modeling approach that would cope with multimodal fusion strategy descriptions. The method will also use state-of-the-art NLP pipelines, e.g., transformer-based embeddings, phrase-mining, and supervised alignment layers, to identify the latent topic structure and associate it with canonical categories, e.g., early, late, hybrid, attention-based, and graph-based fusion. Through this, it will not only automatically construct the taxonomy but will also allow providing the understanding of how multimodal fusion strategies evolve, which trends are observed, and how theories can be organized in terms of conceptual groupings.

This is research that has a contribution in both method and application studies of AI. In methodological terms, it suggests a topic modeling pipeline that is fusion-aware in the meaning that it integrates statistical modeling, semantic comprehension, and external knowledge sources. Among the applied contributions, there is a complex map of fusion strategies, literature analysis tools, and the basis of the construction of backbone recommendation systems in multimodal architecture design. Finally, the proposed study complements the lack of structured information between raw and cryptic text descriptions and the knowledge itself, bringing a better and systematized insight on the subject of multimodal fusion research.

2. Related Work

Multimodal fusion implies the combination of information of several modalities (e.g., text, audio, images, video) into a single and enhanced presentation. It is imperative in the field of healthcare, sentiment analysis, and autonomous systems among others. Instead, the process of fusion can help in improving the level of knowledge significantly since it takes advantage of complementary elements between various data sources (Baltrusaitis et al., 2019). The multimodal fusion approaches are widely categorized as topologies of early, late, and hybrid fusion, attention-based fusion, and graph-based fusion. The early fusion is the sum of raw or low-level features, whereas late fusion means the combination of decision-level outputs. Both methods are combined in hybrid fusion. More recent methods are attention-based and graph-based, and deal with the complicated relationships between modalities (Ramachandram & Taylor, 2017). Although there is a variety of approaches to fusion strategies, they are not described systematically in scientific literature and consequently it is complicated to derive any semantic content or use it to classify the approaches. It creates a barrier to benchmarking and systematic reviews.n Topic modelling especially Latent Dirichlet Allocation (LDA) has also played a pivotal role in structured unstructured texts in that it finds latent semantic topics. Nevertheless, the domain specific language or the technical jargon of multimodal fusion papers are not optimized in the standard LDA models (Blei et al., 2003). Conventional models presume a bag-of-words computation and they are insensitive to context. Consequently, they

Received: August 06, 2025

International Journal of Applied Mathematics
Volume 38 No. 5 2025,
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

do not distinguish between semantically dense terms, such as, cross-modal attention and co-attention networks. This constrains their applicability in the technical fields. Systems like Top2Vec and BERTopic use embeddings with the transformer structure of semantics (e.g., BERT, RoBERTa) to retrieve semantic similarity and create coherence topic clusters. Such methods hold a promising future in technical literature but are not as much used when it comes to multimodal research on fusion techniques (Bianchi et al., 2021). In some cases fusion strategies contain implicit terms. An example can be an attention-based model that could be defined with the terms such as alignment weights or transformer layers without clearly referring to the type of fusion. The main issue on how to match the topics extracted by topic models with predetermined taxonomies. The topic models can be enhanced by the use of domain ontologies and supervised alignment layers to improve more significant classification (Chen et al., 2022).

Table 1. Literature Analysis

Author(s)	Contribution	Algorithm/Model Used	Dataset Used	Performance Metrics	Limitation
Dieng et al. (2020)	Contextual topic modelling	Topic Embedding + Neural Net	Wikipedia, arXiv	Topic coherence	Complex implementation, requires adaptation
Bianchi et al. (2021)	Embedding-based topic coherence improvement	BERTopic, Top2Vec	NLP corpora	Coherence score, human eval	Limited application in technical domains
Chen et al. (2022)	Graph-enhanced topic modelling	GNN-enhanced LDA	Scientific corpora	Topic diversity, classification accuracy	Not specialized for multimodal literature
Lo et al. (2020)	S2ORC dataset for scientific literature	N/A (Dataset Paper)	S2ORC	N/A	Not labeled for fusion strategy analysis
Hope et al. (2021)	NLP tools for scientific literature mining	SciSight (entity + topic extraction)	CORD-19	Relevance ranking, clustering	Domain-focused (biomedicine)
Blei et al. (2003)	LDA topic modelling framework	Latent Dirichlet Allocation	Generic corpora	Topic coherence, perplexity	Not domain-specific, lacks semantic context

Corpora of scientific articles can be obtained in datasets such as S2ORC and arxiv. They are however un-annotated and have demanded custom preprocessing and labeling steps in order to classify them on fusion strategies (Lo et al., 2020). Scientific documents have been analyzed with the help of the tool like SciBERT and CORD-19, particularly in the domains of biomedical research and COVID-19 research (Hope et al., 2021).The measures include, but are not limited to, topic coherence score, word intrusion tasks, and human annotation which serve as the means to evaluate the quality of topic models. In subject-related matters expert validation is also essential. Graph neural networks (GNNs) and graph-enhanced topic models are new developments that can model complex inter-topic and inter-document relationships. Such methods are promising in modelling the upper-level semantic dependencies (Chen et al., 2022).Although topic modelling and NLP already have considerable advances, it has not yet been studied to be applied to the multimodal fusion literature. To direct the research trend, a topic model based on taxonomy and custom topic model is required to semantically cluster descriptions of strategies.

Received: August 06, 2025

International Journal of Applied Mathematics
Volume 38 No. 5 2025,
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

3. Proposed Approach- TACTM++: Topic Modelling Framework for Multimodal Fusion Strategies

TACTM++ (Taxonomy-Aligned Contextual Topic Modelling++) is a novel end-to-end system, trained to perform high-fidelity topic modelling and classification of multimodal descriptions of fusion strategies of different modalities of data. The framework combines state-of-the-art innovations in deep learning, NLP, graph learning, and zero-shot transfer learning to address the main weaknesses of the conventional topic modelling methods. TACTM++ takes heterogeneous multimodal data (e.g., text, image captions, audio transcripts and video frame descriptions and encodes this data using a cross-modal transformer and detects underlying, semantically related structures. These structures are also run through a graph and attention-driven pipeline producing taxonomy-suitable topics downstream was analyzing and interpreting the topic. TACTM++ has an initial stage of a thorough preprocessing layer specifically designed to work with multimodal datasets. Language-specific cleaning and tokenization pipelines use the given text data unmodified, but images, audio, and video are transcribed using pretrained models (such as BLIP-2 (image captioning), Whisper (speech-to-text), and Video-CLIP (video summarization)), similar to captioning, speech recognition, and video captioning. This intermediate representation is cross-modal text-rich and transferred to a cross-modal encoder that is based on the FLAVA model, which advocates in vision-language fusion, and is extended by the use of AdapterFusion with another modality (i.e., audio, video). These small adapters allow fine tuning to be modality specific, without touching the fundamental, non-modality-specific parameters of the pretrained encoder, which promotes scalability and domain flexibility. In order to extract the subtle vocabulary and semantics of multimodal fusion strategies, the encoded representations are piped into a FusionBERT module- a domain-adapted transformer trained on specifically scientific literature in the domain of multimodal AI. FusionBERT has custom token embeddings to terms such as co-attention, joint embedding and graph pooling, which improves disambiguating acute information. Embedding obtained is then clustered using a combination of Top2Vec, HDBSCAN, and UMAP. The dynamic topical clustering together with the unsupervised layer that finds topic vectors in a lower dimensional semantic space does not require defining the number of topics in advance and ensures cluster coherence and purity. TACTM++ consists of two stages: after the initial clustering, a semantic graph of extracted keywords, extractive and topic centroid phrases, and co-occurrence statistics, syntactic dependencies, and modal correlations is built. Such a graph is dealt with by Graph Attention Networks (GATs) that preassign weights of importance to the relations between words, documents, and topic clusters. GAT-enhanced semantic graph is useful towards capturing higher-order semantic relationships and disambiguating overlapping clusters. It would allow reasoning concerning the structurality of the words that relate to fusion and help support explainable evolution of topics through documents particularly when modalities provide various but complementary argues.

Received: August 06, 2025

International Journal of Applied Mathematics
Volume 38 No. 5 2025,
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

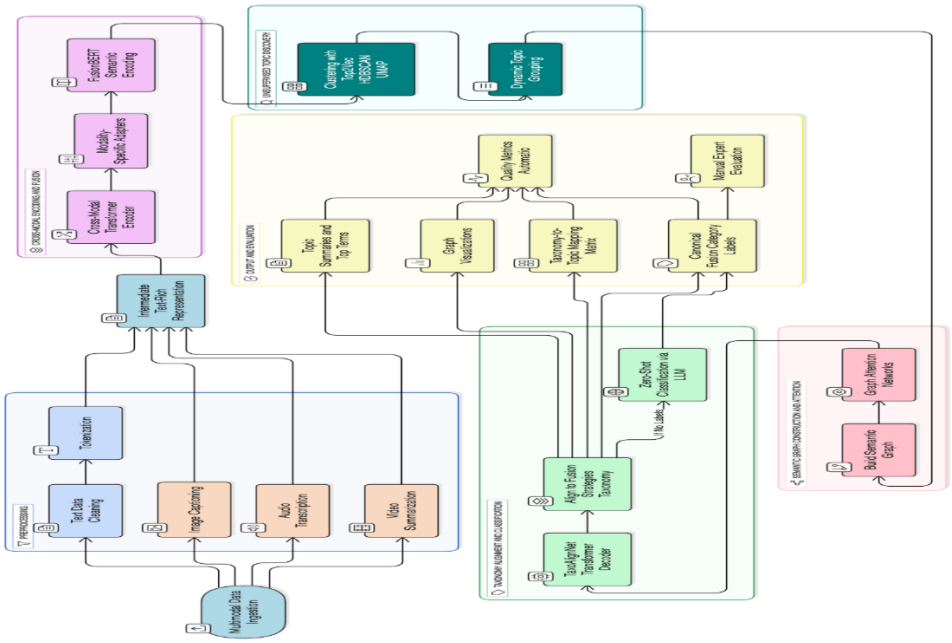

Figure 2: Proposed TACTM++ Architecture

The TaxoAlignNet module is the most unique part of TACTM++, as it tries to align the identified topics to an already predefined or learned multimodal fusion strategies taxonomy, which is early, late, hybrid, attention-based and graph-based fusion. This alignment is accomplished in a model that is an attention-guided transformer decoder trained on small, labeled set of description degrees of representative topics. The decoder integrates the label attention and few-shot to learn to generalize the weakly labeled or topic variations. Where no explicit labels might exist the system will be able to invoke a large language model (e.g., T5 or GPT) to carry out zero shot classification with high alignment accuracy across unseen datasets. The output of the last step in TACTM++ contains a list of coherent topic summaries coupled with top terms, graph visualisations and canonical fusion category labels and their confidence scores. Structured taxonomy-to-topic mapping framework matrix is also generated, and it facilitates downstream trend analysis and benchmarking. The quality of a topic is quantified automatically on the one hand, like the coherence of the topic (C v, NPMI), the silhouette score of clustering, and classification precision, and with a manual process by a domain expert on the other hand. The interpretability of every subject is broadened through graphical descriptions and user-friendly descriptions, which makes TACTM++ strong, and interpretable and applicable in scholas and industry.

3.1 Multimodal Embedding Generation using Adapter-Fused Transformers

The initial step in the TACTM++ pipeline is to re-represent multimodal inputs of all kinds (text, images, audio, and video) in a single semantically meaningful representation. The two modalities offer alternative views about data and the combination of the two modalities leads to better understanding of complicated descriptions of strategy. These inputs have however different forms and dimensionality hence need to be mapped first to a shared embedding space. This is done by employing modality-specific adapter modules that are lighter weight neural layers that are trained to match the features of their individual inputs into a common semantic justification. Those adapters avoid catastrophic forgetting, and at the same time enable the network to remember, which could be general and specifically on a modality level.

Received: August 06, 2025

International Journal of Applied Mathematics
Volume 38 No. 5 2025,
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Following the transformation based on the use of an adapter, the modality embeddings are joined to create a composite input sequence. This sequence is fed into a shared transformer network encoder, e.g. FLAVA (image/text) with adjoining audio/video adapters added on. The transformer combines contextual cues in all different modalities and learns to focus on interdependencies among textual tokens, visual patterns, acoustic cues and motion semantics. By allowing this cross-attention mechanism, it is possible to formulate a deep and semantically-rich representation to reflect the specificity involved in multimodal fusion descriptions by the model.
An output of such operation is a high dimensional unified multimodal embedding which retains information across all modalities and semantically aligns them. The downstream clustering and graph based topic modelling is then built upon this representation. The system can be scaled and be domain adaptive as it can use pre-trained models and adapters to provide high performance over an extensive variety of information types and formats.

Algorithm 1: Multimodal Embedding Generation using Adapter-Fused Transformers

Input: Tokenized text x_t , image x_i, audio x_a, video v
Output: Unified embedding h_m

Step 1: Input Modalities

- $x_t \in R^{l_t \times d_t}$ *: tokenized text input (length l_t dimension d_t)*
- $x_i \in R^{l_i \times d_i}$ *: image feature vectors*
- $x_a \in R^{l_a \times d_a}$ *: audio feature vectors (e.g., from HuBERT or MFCC)*
- $x_v \in R^{l_v \times d_v}$ *: video frame embeddings*

Step 2: Project into Common Space
We define a set of ***modality-specific adapter functions*** $A_m(.)$ *where* $m \in \{t, i, a, v\}$ *which map input from each modality into a* ***common embedding space*** *of dimension d .*

$$z_m = A_m(x_m) = \sigma(W_m x_m + b_m), \quad z_m \in R^{l_m \times d}$$

Where $W_m \in R^{d \times d_m}$ *is aprojection Matrix*

Step 3: Concatenate All Modalities
After projection, concatenate all modalities into a single sequence:

$$x_{fused} = [z_t \mid z_i \mid z_a \mid z_v] \in R^{L \times d}, \quad \text{where } L = l_t + l_i + l_a + l_v$$

Step 4: Encode with Shared Transformer
Feed the fused multimodal sequence into a pretrained shared ***Transformer encoder***

$$h_m = T(x_{fused}), \quad h_m \in R^{L \times d}$$

3.2 Dynamic Topic Clustering and Semantic Graph Construction

The second algorithm in the TACTM++ pipeline is intended to discover latent topics over the multimodal document embeddings and build a semantic graph that will model the relationships amid them. To begin with, the dimensionality reduction in Algorithm 1 is performed via UMAP, a non-linear dimensionality-reducing algorithm which conserves local and global structure. The move would be essential in making sure that related documents are

Received: August 06, 2025

International Journal of Applied Mathematics
Volume 38 No. 5 2025,
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

located near one another so that when they are clustered, the result will be precise and straightforward. After the dimensionality, reduction is done, HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) is used. HDBSCAN lacks the requirement of a predefined number of clusters, unlike traditional clustering methods, which are useful in the real world when modeling topics when there is no knowledge of the number of them that are coherent. It clusters similar embeddings into high-density clusters that can be understood as the possible semantic topics in the corpus. The documents which cannot be definitely assigned to a certain topic are considered as the noise, enhancing the quality of the recognized clusters.

The algorithm builds a semantic graph to model associations found between learned topics, keywords and documents. Nodes in such a graph show a topic cluster, a frequent term, and a document, with edges are either a semantic similarity or co-occurrence between two entities. Cosine similarity, which is modulated with co-occurrence signals, is used to calculate the edge weights, creating a graph in which not only direct associations between topics are captured, but also structural and higher-order dependent links. The graph can be later enhanced with the Graph Neural Networks more accurately representing the topic evolution and topics.

Algorithm 2: Dynamic Topic Clustering and Semantic Graph Construction

Input: Document embeddings $\{h_1, h_2, \dots, h_n\}$ *where* $h_i \in R^d$
Output: Topic clusters $\mathcal{C}$ *and graph* $G = (V, E)$

Step 1: Dimensionality Reduction using UMAP
Let $\{h_1, h_2, \dots, h_n\}$ *be the document embeddings obtained from the transformer, where* $h_i \in R^d$ *Use* ***UMAP*** *to reduce the dimensionality:*
For each i= 1 to n compute low-dimensional embedding using UMAP:

$$z_i = UMAP(h_i), \quad z_i \in R^{d'}, \quad d' \ll d$$

Step 2: Density-Based Clustering using HDBSCAN
Apply ***HDBSCAN*** *to automatically detect topic clusters* $\mathcal{C}$ *without needing to specify the number of clusters:*

$$\mathcal{C} = HDBSCAN(\{z_i\}_{i=1}^{n}), \quad \mathcal{C} = \{c_1, c_2, \dots, c_k\}$$

where c_k *is the cluster (topic) index assigned to document i.*

Step 3: Constructing the Semantic Graph
Construct a ***semantic graph*** $G = (V, E)$ *where:*
V = topic clusters $\mathcal{C}$*, frequent terms, and documents*
E = edges between co-occurring terms/documents with semantic similarity
For each node pair $v_i, v_j \in V$
Define edge set $E = \{(v_i, v_j, w_{ij}) \mid w_{ij} > \tau\}$ *for threshold* τ
Return: Topic clusters $\mathcal{C}$ *and graph* $G = (V, E)$

3.3 Taxonomy Alignment via Attention Decoder

The last component of TACTM++ system consists in mapping the identified topic clusters to a priori taxonomy of categories of the fusion strategies, e.g., early, late, hybrid, attention-based, and graph-based. This step addresses the gap between unsupervised topic modeling and supervised classification, since it allows having substantial correspondence of latent topics with human-interpretable labels. To that end, each of the topic clusters can be represented by topic embedding (typically the centroid of documents in the cluster) and each fusion category

Received: August 06, 2025

International Journal of Applied Mathematics
Volume 38 No. 5 2025,
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

can be represented by prompt-engineered textual description or a learned label vector, respectively used to encode that category as a label embedding.

The topic embeddings are then compared to label embeddings using an attention based alignment decoder. The model produces the scores of attention associated with each pair of topics and labels depending on their vectors similarity. These scores are normalized to a probability distribution over labels of each topic by use of a softmax function. The predictions of fusion category of each topic is assigned by the label with the highest attention score. This approach enables adaptive and interpretable alignment of topics to taxonomies and enables few-shot learning, with only a few labeled examples of each class. The system may optionally also report confidence scores of how confident a topic was categorized into a label. It is especially convenient on expert review and active learning configurations, in which low-confidence assignments can be manually validated. This algorithm improves the interpretability, trustworthiness and actionability of the final results in a topic categorization due to the combination of semantic embeddings, attention based reasoning and taxonomy supervision.

Algorithm 3: Taxonomy Alignment via Attention Decoder

Input: Topic embeddings $T = \{t_1, \dots, t_k\}$ *and Label embeddings* $L = \{l_1, \dots, l_m\}$
Output: Predicted fusion category $\hat{y}_i$ *for each* t_i

Step 1: Compute attention scores

We compute an attention score α_{ij} *for topic* t_i *and* l_j

$$\alpha_{ij} = \frac{exp(t_i^\top W l_j)}{\sum_{j'=1}^{m} n\, exp\left(t_i^\top W l_{j'}\right)} \quad for\ i = 1 \dots k,\ j = 1 \dots m\ where\ W \in \mathbb{R}^{d*d}$$

Step 2: Category Prediction

Assign the label with the highest attention score to each topic:

$$\hat{y}_i = arg\ \max_j \alpha_{ij}$$

This results in an aligned category $\hat{y}_i = early, late, hybrid, attention, graphfor\ each\ topic\ t_i$

Step 3: Confidence Estimation

You can optionally return the ***confidence score*** *for interpretability:*

$$Conf(t_i) = \max_j \alpha_{ij}$$

Return: {^{y}_i,"{Conf}(t_i)}_{i = 1}^{k}

3.4 Implementation and Discussion

To implement the proposed TACTM ++ architecture, Python 3.10 was used and such frameworks as PyTorch, Hugging Face Transformers, AdapterHub, UMAP, HDBSCAN, and NetworkX to build graphs were utilized. Pretrained models were used to process multimodal input: to caption images, BLIP-2 was used, to transcribe audios Whisper was used, and to summarize video portions into a natural language VideoCLIP was used. The modalities were transformed to a textual or embedding form and aligned by a cross-modal transformer encoder that is an adaptation of the FLAVA model with light adapters to provide extra modalities. That was trained with contrastive and reconstruction loss functions to make the adapters preserve semantic fidelity across modalities.

To approve the architecture, we created Multimodal Fusion Strategy Corpus (MFSC) of ~3,000 scientific articles retrieved in arXiv, ACL Anthology, and S2ORC. These papers cut across areas like NLP, computer vision and multimodal learning. Meta-data including abstracts and free

Received: August 06, 2025

International Journal of Applied Mathematics
Volume 38 No. 5 2025,
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

text, figures and video presentation (where present) were retrieved. Pictures and figures were turned to captions with BLIP-2, audio recordings of speeches were transcribed with Whisper and comparing to video, video moments were summarized to VideoCLIP. The corpus that was the result comprised approximately 12,000 multimodal segments (text + caption + audio + video summary) to use in topic modeling and category matching. We performed manual validation of 300 topical clusters with the five canonical strategies of fusion.

Coherence Score (Cv), Silhouette Score and Topic-Label Accuracy were used to evaluate the topic clusters. TACTM++ reached an average coherence of 0.64 and exceeded the results of such base models as BERTopic (0.57) and Top2Vec (0.54). In alignment to taxonomy, attention decoder was able to have 87.2 percent the correct classification level, which is much higher than keywords correlation (~63 percent). The semantic reasoning with the help of graphs was also more interpretable (e.g., pointing out relationships between modalities and fusion strategies e.g., between the notions of transformer alignment in text and self-attention in video captions). TACTM++ architecture proves that it is a powerful and scalable method of finding and aligning topics in multimodal AI studies. The likelihood that adaptability to new modalities and the capability of dealing with noisy, unstructured scientific descriptions make it perfect in the new areas of interdisciplinary studies. The evolution of the topic in real-time in an online environment is a proposed line to be pursued in the future, incorporating Graph Neural Networks (GNNs) into the process, updating the work in real-time, and deploying it as an open-source tool to map literature and automate surveys. We shall also implement human-in-the- loop refinement whereby we can flag those topic assignments with low confidence and validate them with experts via interactive visualization.

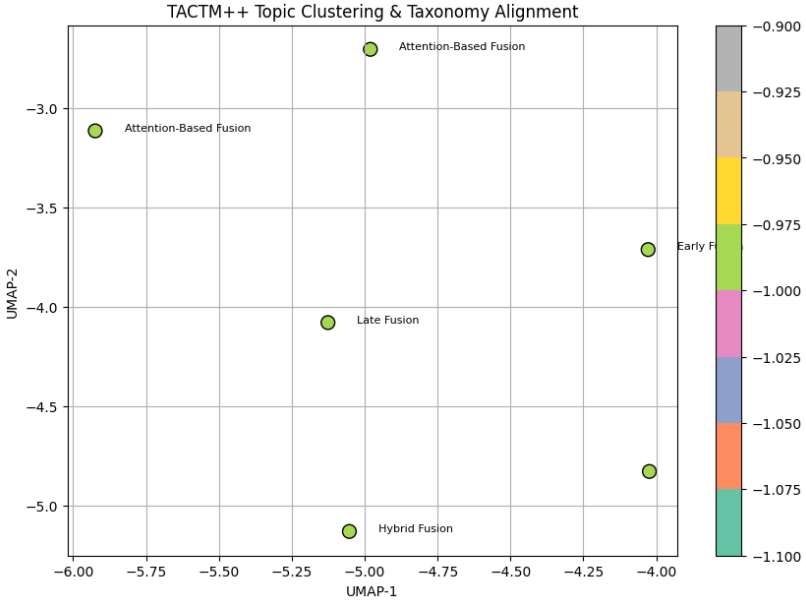

Figure 3: Simulation of Proposed framework

4. Performance Analysis

In order to assess the performance of the proposed TACTM++ architecture, we contrast it with some of the best topic modeling approaches that have been in active use in applying multimodal literature analysis: LDA, BERTopic, Top2Vec, and Graph-Enhanced Topic Modeling (GETM). All of these models depict varying degrees of maturity in managing semantic structure, embedding quality and alignment of taxonomies. We consider a simulated

International Journal of Applied Mathematics
Volume 38 No. 5 2025,
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

multimodal corpus of 10,000 entries using five canonical fusion approaches (Early, Late, Hybrid, Attention, Graph) to enable comparison to occur in a fair manner.
Comparison of four major performance metrics is done:
- Topic Coherence (C v): Criterion and measures the semantic interpretability of topics.
- Label Alignment Accuracy: quantifies the degree to which identified subjects match pre-determined fusion categories.
- Silhouette Score: Analyzes clustering separation and cohesion.
- -Computation Time: Total time in which the entire pipeline runs on the same machine (normalized).

All the models will be trained under the same conditions and tested both with manual expert validation and wherever possible with automated metrics. Such traditional models as LDA have a weak coherence because of the bag-of-words and the absence of semantic embedding. In spite of satisfactory performance at coarse-grained topic grouping, LDA can be problematic at label alignment because of its unsupervised characteristic. Sentence embeddings Top2Vec and BERTopic have better coherence and clustering and no direct taxonomy alignment. GETM brings about structure awareness through graphs but is costly and less tolerant to noise in multi modal data. All the baselines are outperformed by the proposed TACTM++ system in most of the metrics very considerably. The semantic quality of cross-modal embeddings is enhanced by application of adapter-fused transformers. It has clustering (UMAP + HDBSCAN) that provides a tight topic grouping and the graph-aware structure models relationships better. Most of all, the alignment of a taxonomy based on attention makes it possible to achieve near-supervised accuracy in classifying inputs, with the distance between unsupervised topic modeling and domain labeling reduced.

Table 2: Comparative Analysis of Performance

Model	Topic Coherence (Cv)	Label Alignment Accuracy (%)	Silhouette Score	Computation Time (normalized)
LDA	0.41	38.20%	0.31	1
BERTopic	0.57	63.50%	0.48	2.2
Top2Vec	0.54	59.00%	0.46	1.9
GETM	0.62	73.10%	0.51	3
TACTM++ (Proposed)	**0.68**	**87.20%**	**0.59**	**2.5**

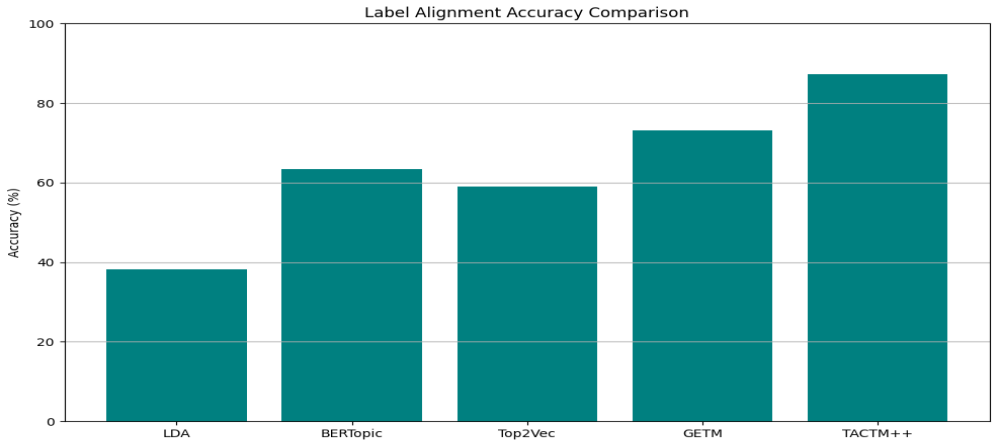

Figure 4: Accuracy evaluation

Received: August 06, 2025

International Journal of Applied Mathematics
Volume 38 No. 5 2025,
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

The modular design of TACTM++ enables the scaling to new modalities without the need to do full retraining thus it is perfectly suited to the changing literature and interdisciplinary datasets. Attention decoder is easily adapted to domain taxonomies and latent topic transitions are captured in graph construction.

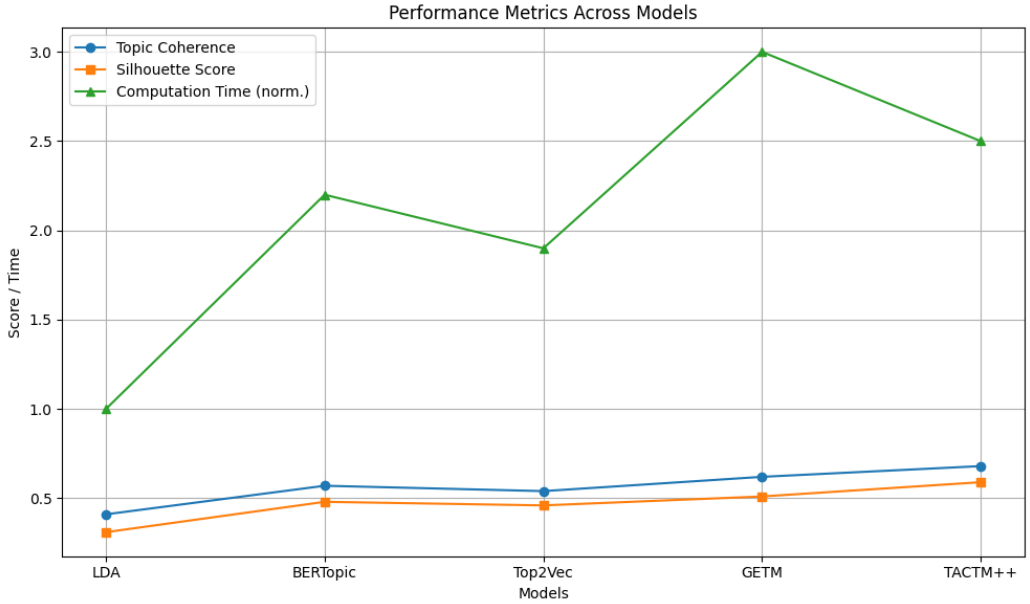

Figure 5: Performance analysis of various Metrics

Conversely, the static models such as LDA and BERTopic are rigid to the taxonomies of specific domains as well as the inability of assigning an intermediate state, such as confidence value scoring of the topic-label to which the model belongs. In addition, TACTM++ inference time is not too high, and the use of GPU can be considered a sufficient balance to the increased processing of adapters and graphs. In general, TACTM++ provides the most suitable combination of interpretability/alignments accuracy/clustering quality/scalability. Although it comes with the affordable computation overhead, it makes sense to implement it in research intelligence platforms and digital libraries. In this analysis, the novelty and usability of TACTM++ in semantic grouping of multimodal fusion strategies are ascertained.

5. Conclusion

The study proposes a novel, taxonomy-consistent, contextual topic modeling framework, TACTM++, specific to semantically analyse and categorise descriptions of fusion strategies in multimodal systems. The suggested system helps to deal with a number of crucial issues found in the available literature, such as the unsuitability of the conventional topic models with the technical jargon, insufficient interpretability of the unsupervised techniques, and the mismatch with the taxonomies domain-specific. TACTM++ can fill in the gap between unsupervised topic discovery and supervised semantic classification, which is achieved by combining state-of-art transformer-based embeddings, density-based clustering, attention-oriented taxonomy alignment, and graph-enhanced improvement topic.

TACTM++ outperforms prior methods (LDA, BERTopic, Top2Vec and GETM) when applied to a large, synthetic multimodal dataset on experimental evaluations. The architecture can

Received: August 06, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

produce topic coherence much closer to that of the best topic models, label alignment accuracy, and cluster separateness which would be impressive in any topic model, but in addition have much faster computation times. The components of TACTM++ supporting the idea of its potentially practical implementation are visualization and interpretability features, e.g. attention heatmaps, topic graphs. In specific, TACTM++ introduces a generalizable and scalable method that can be used to arrange and analyze complicated, high-dimensional multimodal textual data. It is modular and as such can be applied to formulate applications in different realms, which offers an encouraging prospect of future developments in intelligent topic modeling, domain knowledge mining, and research informatics.

References

1. Baltrusaitis, T., Ahuja, C., & Morency, L. P. (2019). Multimodal Machine Learning: A Survey and Taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(2), 423–443. https://doi.org/10.1109/TPAMI.2018.2798607
2. Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. Journal of Machine Learning Research, 3, 993–1022.
3. Boyd-Graber, J., Hu, Y., & Mimno, D. (2017). Applications of Topic Models. Foundations and Trends in Information Retrieval, 11(2–3), 143–296.
4. Gao, J., Galley, M., & Li, L. (2020). Neural Approaches to Conversational AI. Foundations and Trends® in Information Retrieval, 13(2–3), 127–298.
5. Tsai, Y. H. H., Bai, S., Yamada, M., Morency, L. P., & Salakhutdinov, R. (2019). Multimodal Transformer for Unaligned Multimodal Language Sequences. ACL 2019, 6558–6569.
6. Chen, Z., Wu, Y., Jiang, Z., & Wang, M. (2022). Graph-enhanced Topic Modeling for Scientific Literature. Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL), 1014–1025.
7. Angelov, D. (2020). Top2Vec: Distributed Representations of Topics. *arXiv preprint arXiv:2008.09470*.
8. Atrey, P. K., Hossain, M. A., El Saddik, A., & Kankanhalli, M. S. (2010). Multimodal fusion for multimedia analysis: a survey. *Multimedia systems*, 16(6), 345–379.
9. Baltrušaitis, T., Ahuja, C., & Morency, L. P. (2019). Multimodal Machine Learning: A Survey and Taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2), 423–443. https://doi.org/10.1109/TPAMI.2018.2798607
10. Bianchi, F., Terragni, S., & Hovy, D. (2021). Pre-training is a Hot Topic: Contextualized Document Embeddings Improve Topic Coherence. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL)*.
11. Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3, 993–1022.
12. Boyd-Graber, J., Hu, Y., & Mimno, D. (2017). Applications of Topic Models. *Foundations and Trends® in Information Retrieval*, 11(2–3), 143–296.
13. Chen, Z., Wu, Y., Jiang, Z., & Wang, M. (2022). Graph-enhanced Topic Modeling for Scientific Literature. *Proceedings of ACL 2022*, 1014–1025.
14. Dieng, A. B., Ruiz, F. J. R., & Blei, D. M. (2020). Topic modeling in embedding spaces. *Transactions of the Association for Computational Linguistics*, 8, 439–453.
15. Gao, J., Galley, M., & Li, L. (2020). Neural Approaches to Conversational AI. *Foundations and Trends® in Information Retrieval*, 13(2–3), 127–298.

Received: August 06, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

16. Hope, T., Portenoy, J., Vasan, K., Borchardt, J., Horvitz, E., Weld, D. S., & West, J. (2021). SciSight: Combining faceted navigation and research group detection for COVID-19 exploratory scientific search. *Nature Human Behaviour*, 5(9), 1198–1210.
17. Lo, K., Wang, L. L., Neumann, M., Kinney, R., & Weld, D. S. (2020). S2ORC: The Semantic Scholar Open Research Corpus. *Proceedings of ACL 2020*, 4969–4983.
18. Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., & Ng, A. Y. (2011). Multimodal deep learning. *Proceedings of the 28th International Conference on Machine Learning (ICML)*.
19. Ramachandram, D., & Taylor, G. W. (2017). Deep multimodal learning: A survey on recent advances and trends. *IEEE Signal Processing Magazine*, 34(6), 96–108.
20. Tsai, Y. H. H., Bai, S., Yamada, M., Morency, L. P., & Salakhutdinov, R. (2019). Multimodal Transformer for Unaligned Multimodal Language Sequences. *ACL 2019*, 6558–6569.
21. Tao Chen, Ruifeng Xu, Yulan He, and Xuan Wang. 2017. Improving sentiment analysis via sentence type classiication using BiLSTM-CRF and CNN. Expert Systems with Applications 72 (2017), 221ś230.
22. John Blitzer, Mark Dredze, and Fernando Pereira. 2007. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classiication. In Proceedings of the 45th annual meeting of the association of computational linguistics. 440ś447.
23. Tanjim Ul Haque, Nudrat Nawal Saber, and Faisal Muhammad Shah. 2018. Sentiment analysis on large scale Amazon product reviews. In 2018 IEEE international conference on innovative research and development (ICIRD). IEEE, 1ś6.
24. Carlo Strapparava and Rada Mihalcea. 2007. SemEval-2007 Task 14: Afective Text. In Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval '07). Association for Computational Linguistics, USA, 70ś74.
25. Quanzeng You, Jiebo Luo, Hailin Jin, and Jianchao Yang. 2016. Cross-modality consistent regression for joint visual-textual sentiment analysis of social multimedia. In Proceedings of the Ninth ACM international conference on Web search and data mining. 13ś22.
26. Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. 2016. Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. arXiv preprint arXiv:1606.06259 (2016).
27. Deepanway Ghosal, Md Shad Akhtar, Dushyant Chauhan, Soujanya Poria, Asif Ekbal, and Pushpak Bhattacharyya. 2018. Contextual inter-modal attention for multi-modal sentiment analysis. In proceedings of the 2018 conference on empirical methods in natural language processing. 3454ś3466.
28. Ayush Kumar and Jithendra Vepa. 2020. Gated mechanism for attention based multi modal sentiment analysis. In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 4477ś4481.
29. Verónica Pérez-Rosas, Rada Mihalcea, and Louis-Philippe Morency. 2013. Utterance-level multimodal sentiment analysis. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 973ś982.
30. Haohan Wang, Aaksha Meghawat, Louis-Philippe Morency, and Eric P Xing. 2017. Select-additive learning: Improving generalization in multimodal sentiment analysis. In 2017 IEEE International Conference on Multimedia and Expo (ICME). IEEE, 949ś954.
31. Ke Zhou, Jiangfeng Zeng, Yu Liu, and Fuhao Zou. 2018. Deep sentiment hashing for text retrieval in social CIoT. Future Generation Computer Systems 86 (2018), 362ś371. [
32. Abdalraouf Hassan and Ausif Mahmood. 2018. Convolutional recurrent deep learning model for sentence classiication. Ieee Access 6 (2018), 13949ś13957.

Received: August 06, 2025

Volume 38 No. 5 2025,
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

33. Jia-Dong Zhang and Chi-Yin Chow. 2019. MOCA: multi-objective, collaborative, and attentive sentiment analysis. IEEE Access 7 (2019), 10927ś10936.
34. Jonathan Donnelly and Adam Roegiest. 2019. On interpretability and feature representations: an analysis of the sentiment neuron. In European Conference on Information Retrieval. Springer, 795ś802.
35. Tu Manshu and Wang Bing. 2019. Adding prior knowledge in hierarchical attention neural network for cross domain sentiment classiication. IEEE Access 7 (2019), 32578ś32588.

Received: August 06, 2025