Journal of Intelligent Computing

DLINE JOURNALS

Print ISSN: 0976-9005
Online ISSN: 0976-9013

JIC 2026: 17 (2)
https://doi.org/10.6025/jic/2026/17/2/76-92

Integrated Multi-Model Learning Framework for Structured–Textual–Temporal Data: A Descriptive Analysis

Puttakul Puttawattanakul, Hathairat Ketmaneechairat
College of Industrial Technology, King Mongkut's University of Technology North Bangkok, Thailand
hathairat.k@cit.kmutnb.ac.th, puttakuls@gmail.com

ABSTRACT

Real-world datasets increasingly encompass heterogeneous modalities, yet conventional fusion techniques often struggle with varying data reliability, missing inputs, and dynamic cross-modal interactions. To address these limitations, this study proposes an integrated multi-model learning framework that synergistically combines transformer, tree, and sequential based architectures. Central to this approach is the AdaptiveFusion module, which dynamically estimates modality reliability, facilitates cross modal attention, and aggregates features adaptively via a learnable gating mechanism. The framework operationalizes this architecture by integrating BERT for deep textual semantics, XGBoost for structured numerical temporal prediction, and LSTM for sequential dependency modeling. Empirical validation is conducted on a curated dataset tracking the chronological evolution of conversational AI systems, comprising categorical, temporal, numerical, and textual attributes. Experimental results demonstrate that each component effectively captures its designated data dimension: BERT achieves rapid semantic convergence, XGBoost delivers robust structured classification with high F1-scores (>0.80), and LSTM successfully identifies latent temporal trajectories. When integrated, these paradigms overcome individual limitations, yielding a holistic analytical system that enhances both predictive accuracy and interpretability. Despite constraints related to dataset scale and class imbalance, the proposed framework establishes a scalable, context-aware methodology for multimodal data fusion. This work underscores the critical value of hybrid modeling strategies in extracting comprehensive insights from complex, real-world information ecosystems.

Keywords: Multimodal Learning, Adaptive Fusion, Structured-Textual-Temporal Data, BERT, XGBoost, LSTM, Hybrid Modeling FrameWork, Cross-Modal Attention, Conversational AI Analytics

Received: 24 September 2025, Revised 6 January 2026, Accepted 19 January 2026

1. Introduction

Multimodal data, which integrates heterogeneous sources such as images and textual content, has become increasingly prevalent across social media platforms. The availability of such rich, multimodal inputs enables a deeper understanding of human emotions, opinions, and behavioral patterns associated with various topics and events. Consequently, a wide range of multimodal sentiment analysis (SA) techniques has been developed to effectively integrate and interpret these diverse data modalities [1]. In real-world scenarios, information is rarely unimodal; rather, it is inherently multimodal, encompassing multiple complementary sources that collectively contribute to human perception and cognition [2].

2. Review of Earlier Studies

Early multimodal architectures predominantly relied on conventional deep learning techniques, including convolutional neural networks (CNNs), long short-term memory (LSTM) networks, and fully connected neural networks. These architectures were designed to process and integrate features extracted from different modalities. Anbazhagan proposed a multimodel framework that dynamically selects the most appropriate model or combination of models based on the problem context and the nature of the input data. This model selection mechanism enables adaptive utilization of different architectures within a unified system, thereby enhancing performance across varied multimodal tasks [3].

A fundamental aspect of multimodal learning is data fusion, the integration of data from multiple modalities that provide distinct perspectives on a shared phenomenon. Data fusion techniques aim to improve inference accuracy by leveraging complementary, redundant, and cooperative information across modalities [4, 5]. These fusion strategies are not mutually exclusive and often work synergistically to enhance predictive performance [6, 7].

Recent research has focused extensively on designing advanced multimodal architectures that effectively capture cross-modal interactions. Zhang et al. [8] introduced a Semantic Content Correlation (SCC) approach to identify and model relationships between images and their corresponding textual descriptions [8]. Huang et al. proposed an Attention based Modality-Gated Network (AMGN), which incorporates a modality gated LSTM to dynamically adjust each modality's contribution based on its reliability in expressing sentiment [9]. Similarly, Xu et al. developed a Bi-Directional Multi-Level Attention (BDMLA) model that exploits both complementary and comprehensive information from visual and textual modalities for joint sentiment classification [10].

Further advancements include the work of Cao et al., who proposed Various Syncretic Co-attention Networks (VSCN) to capture multi-level correlations across modalities and integrate their unique characteristics for improved sentiment prediction [11]. Hu et al. introduced a neural architecture that combines global and local fusion mechanisms, in which global features are derived using attention modules and local features are learned through coarse to fine fusion strategies, ultimately yielding more precise sentiment predictions [12]. An et al. presented a framework that enhances targeted multimodal sentiment classification by incorporating semantic image descriptions, enabling dynamic adjustment of image importance based on textual similarity relationships

[13]. Additionally, Liu et al. developed an importance attention network that assigns adaptive weights to different modalities, thereby capturing their relative contributions to sentiment prediction [14].

Sequential deep learning models have also been explored for multimodal sentiment analysis. Tembhurne investigated the applicability of sequential neural networks in modeling temporal dependencies within multimodal data [15]. More recently, Veerababu Reddy proposed a unified intelligent information system that integrates multiple large language models (LLMs) with structured data storage for context aware legal document analysis. This framework employs a multi-model architecture with shared semantic representations to enable joint clause interpretation, contextual risk assessment, and cross-clause consistency analysis through natural language inference [16].

Comprehensive surveys have highlighted both the challenges and opportunities in multimodal sentiment analysis. Kaur et al. provided a taxonomy-driven overview of existing techniques and identified key research gaps [17]. Similarly, Feng and Soleymani reviewed multimodal sentiment analysis across diverse application domains, including vlogs, spoken language processing, visual-textual data, human machine interaction, and human human communication [18, 19].

From a methodological perspective, tensor based fusion approaches have gained attention for modeling complex interactions among modalities. Liu et al. proposed a low-rank tensor fusion method to address the exponential growth in dimensionality and computational complexity associated with conventional tensor representations [20]. In a related study, Zadeh et al. introduced a Tensor Fusion Network (TFN) that captures intra-modality and inter-modality dynamics for end-to-end multimodal sentiment analysis across unimodal, bimodal, and trimodal settings. This approach leverages LSTM networks to learn temporal language representations using GloVe word embeddings [21, 22].

In addition to tensor-based approaches, hybrid models combining deep learning and traditional machine learning techniques have been explored. Pre-trained CNNs, such as those trained on ImageNet, are commonly fine-tuned to extract visual features, while textual features are learned from distributed word representations [23, 24]. Multimodal regression models are then employed to enforce consistency across modalities, and logistic regression is often used for decision level fusion of probabilistic outputs [25]. Given that textual sentiment analysis typically achieves higher accuracy than visual sentiment analysis, greater weight is often assigned to textual features during both feature level and decision level fusion.

Kernel-based methods have also been applied to multimodal fusion. Initially, support vector machines (SVMs) were used as single kernel models for integrating features from different modalities. However, Multiple Kernel Learning (MKL) has emerged as a more flexible alternative, enabling the grouping of similar features with dedicated kernels. This approach facilitates the fusion of audio, visual, and textual modalities, particularly in multimedia data such as YouTube videos [26].

Despite significant advancements in multimodal learning, existing approaches often struggle to handle heterogeneous data sources with varying reliability, missing modalities, and dynamic inter-modal relationships. Many models rely on static fusion strategies that do not adapt to the contextual importance of each modality. To address these limitations, this study proposes an adaptive, dynamic fusion mechanism that intelligently weighs and integrates multimodal representations. This motivates the development of the *AdaptiveFusio*n

architecture described in the following section. This gap is particularly critical in scenarios involving structured, textual, and temporal data, which motivates the proposed integrated multi-model framework.

3. Architecture of this work

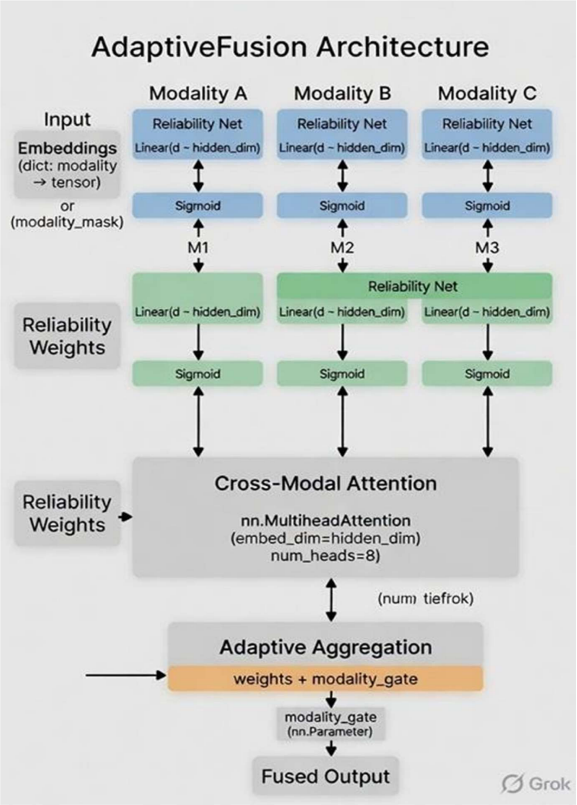

Figure 1. AdaptiveFusion Architecture

3.2 AdaptiveFusion Module

To effectively fuse multimodal embeddings while accounting for varying modality quality and potential omissions, we propose AdaptiveFusion, a lightweight yet expressive fusion layer. The module dynamically estimates each modality's reliability, enables cross-modal interaction via attention, and adaptively aggregates features using a learnable gating mechanism.

Formally, let the input consist of a dictionary of modality-specific embeddings $\mathcal{E} = \{\mathrm{e}_m \in \mathbb{R}^{d_m}\}_{m=1}^{M}$, where M is the number of modalities and d_m denotes the dimensionality of the m-th modality embedding. An optional binary modality mask can also be provided to indicate which modalities are missing during inference.

The architecture comprises three core components:

1. Modality Reliability Estimator. For each modality m, we employ a lightweight reliability network defined as:

$$r_m = \sigma(\mathrm{W}_m \mathrm{e}_m + \mathrm{b}_m),$$

where $\sigma(\cdot)$ is the sigmoid activation, $\mathrm{W}_m \in \mathbb{R}^{h \times d_m}$ and $\mathrm{b}_m \in \mathbb{R}^h$ are learnable parameters, and h is the shared

hidden dimension. This produces a scalar reliability weight $r_m \in [0,1]$for each modality, reflecting its estimated quality or informativeness. The collection of reliability weights is denoted as $\mathrm{r} = [r_1, r_2, \ldots, r_M]$.

2. Cross-Modal Attention. To enable rich interaction across modalities, we project all modality embeddings to a common hidden dimension and feed them into a multi-head self-attention mechanism:

$$\mathrm{F} = \mathrm{MultiHeadAttention}(\mathrm{Q}, \mathrm{K}, \mathrm{V}; \mathrm{key\_padding\_mask}),$$

where the queries, keys, and values are derived from the projected embeddings, and the optional *key_padding_mask* (derived from the modality mask) prevents attention to missing modalities. We use 8 attention heads with embedding dimension h.

3. Adaptive Aggregation. The attended features F are then combined with the reliability weights $\mathrm{r}$and a learnable modality gate $\mathrm{g} \in \mathbb{R}^M$ (initialized to ones) through an adaptive weighted aggregation:

$$\mathrm{z} = \sum_{m=1}^{M} (r_m \cdot g_m \cdot \mathrm{f}_m) \oslash \left(\sum_{m=1}^{M} r_m \cdot g_m + \epsilon \right),$$

where f_m denotes the attended feature corresponding to modality m, $\oslash$ denotes element-wise division (for normalization), and ϵ is a small constant for numerical stability. This formulation allows the model to down-weight unreliable or missing modalities while preserving the contribution of high quality ones.

The final output $\mathrm{z} \in \mathbb{R}^h$ serves as the fused multimodal representation, which can be passed to downstream task-specific heads.

Implementation Details. *AdaptiveFusion* is implemented as a *PyTorch nn.Module*. The reliability networks are stored in an *nn.ModuleDict* for modality-specific linear projections. The cross-modal attention uses *PyTorch's nn.MultiheadAttention* with 8 heads. The modality gate is implemented as a learnable *nn.Parameter*. The entire module is lightweight, adding minimal parameters while providing strong adaptability to heterogeneous and incomplete multimodal inputs.

4. Data and Methodology

The proposed AdaptiveFusion architecture serves as the conceptual foundation for integrating heterogeneous representations derived from different learning paradigms. In the present study, this concept is operationalized through a hybrid modeling framework that combines transformer-based, tree-based, and sequential models. Specifically, BERT is used to extract semantic representations from text, XGBoost captures structured numerical and temporal relationships, and LSTM models temporal dependencies. These components collectively emulate the principles of adaptive multimodal fusion.

4.1 Dataset Description

The present study employs a structured dataset titled *"chatgpt_cleaned (3).csv"*, which captures a chronological and analytical record of the evolution of modern conversational artificial intelligence systems. The dataset integrates categorical, temporal, numerical, and textual attributes, enabling a comprehensive,

multidimensional analysis of technological development, adoption patterns, feature advancements, efficiency improvements, and market influence. It is particularly valuable for understanding the transition of conversational AI from basic chat-based systems to sophisticated multimodal assistants.

The primary objective of utilizing this dataset is to examine growth trajectories, capability enhancements, and the real-world impact of conversational AI systems over time. Each record in the dataset represents a specific event, development, or measurable milestone, enriched with contextual descriptions that support both quantitative and qualitative analysis.

The dataset comprises several key variables. The *Category* attribute serves as a high-level classification label and functions as the primary target variable for supervised learning tasks. The *Subcategory* variable provides a more granular classification nested within each category, enabling hierarchical and fine-grained analysis. The *Item* field represents specific entities or instances associated with each record and captures detailed observations or milestones. Temporal information is represented through the *Date* variable, which facilitates time-series and sequence-based modeling. Quantitative aspects are captured by the *Value* attribute, which includes numerical indicators such as usage statistics, financial measures, or performance metrics. Additionally, the *Notes* field contains unstructured text that provides contextual information and serves as the basis for natural language processing tasks.

Following preprocessing, the dataset consists of 34 valid observations, obtained after excluding records with missing values in critical fields such as *Category* and *Notes*. This refinement ensures data consistency and analytical reliability.

To ensure clarity and logical consistency, the methodological components are organized into a unified framework encompassing preprocessing, feature engineering, model development, and evaluation.

4.2 Data Preprocessing

To ensure the dataset is suitable for machine learning applications, a systematic preprocessing pipeline was implemented, encompassing data cleaning, transformation, and feature engineering. Initially, incomplete records containing missing values in essential fields were removed to reduce noise and improve the robustness of subsequent analyses.

The temporal variable (*Date*) was further decomposed into structured components, specifically *Year* and *Month*, enabling the models to capture temporal patterns, seasonal variations, and longitudinal trends. In parallel, the *Value* attribute, which originally contained formatted numerical expressions such as currency symbols and textual suffixes, was transformed into a standardized numeric format to ensure compatibility with statistical and machine learning algorithms.

Categorical variables, including *Category* and *Subcategory*, were encoded using label encoding techniques. This transformation allows categorical data to be effectively utilized within classification models while preserving relational structure among classes.

4.3 Hybrid Modeling Strategy

This study adopts a hybrid modeling framework that integrates multiple machine learning paradigms to capture

the heterogeneous nature of the dataset. Specifically, the framework combines transformer-based text modeling, tree-based structured learning, and sequence modeling approaches. This integration enables simultaneous analysis of semantic information derived from textual data, statistical relationships embedded in numerical features, and temporal dependencies reflected in sequential patterns.

4.4 Modeling Framework

Each model is designed to capture a specific dimension of the dataset, and together they form a complementary analytical system.

4.4.1 BERT Model

To extract and model semantic information from the *Notes* field, a Bidirectional Encoder Representations from Transformers (BERT) model was employed. The BERT-base architecture was fine-tuned for multi-class classification tasks, where textual inputs are tokenized and processed through multiple transformer layers to generate context-aware embeddings. These embeddings are subsequently passed through a fully connected classification layer to predict category labels.

The training process was conducted using a learning configuration with a limited number of epochs (3) and a maximum sequence length of 128 tokens. The primary objective of this model is to learn complex semantic patterns and contextual relationships within textual descriptions, enabling accurate prediction of category labels.

4.4.2 XGBoost Model

To leverage structured features such as numerical and temporal attributes, an Extreme Gradient Boosting (XGBoost) classifier was implemented. The model utilizes *Value*, *Year*, and *Month* as input features, enabling it to capture statistical relationships and temporal trends within the dataset.

XGBoost, as an ensemble learning technique based on gradient-boosted decision trees, is particularly effective in modeling non-linear relationships and handling feature interactions. Additionally, it provides insights into feature importance, thereby enhancing interpretability. The primary objective of this component is to predict the *Category* variable based on structured inputs, offering a complementary perspective to text-based modeling.

4.4.3 LSTM Model

To capture temporal dependencies and sequential patterns, a Long Short-Term Memory (LSTM) network was employed. The dataset was first arranged chronologically, and fixed-length sequences were constructed using preceding observations to predict subsequent categories.

The LSTM architecture consists of an embedding layer that converts categorical inputs into dense vector representations, followed by an LSTM layer that captures temporal dependencies. A dense output layer is used to generate probability distributions over the class categories. This model facilitates analysis of temporal evolution and enables prediction of future trends from historical sequences.

4.5 Model Evaluation

The performance of the classification models, including BERT and XGBoost, was evaluated using standard

metrics such as accuracy, precision, recall, and F1-score, ensuring a comprehensive assessment of predictive performance. For the LSTM-based sequence model, evaluation focused on prediction accuracy and the analysis of learning curves, particularly the comparison between training and validation accuracy to assess model generalization.

While individual models are evaluated independently, their combined analytical contribution is realized through an integrated framework, as described below.

4.51 Integrated Analytical Approach

The integration of BERT, XGBoost, and LSTM within a unified analytical framework provides a holistic approach to modeling the dataset. While BERT captures semantic relationships in text, XGBoost models structured numerical and temporal dependencies, and LSTM captures sequential and temporal patterns. This complementary combination enhances both predictive performance and interpretability, making the framework well-suited for analyzing complex, heterogeneous datasets.

4.52 Methodological Limitations

Despite the strengths of the proposed framework, certain limitations must be acknowledged. The relatively small dataset size (n = 34) may limit the models' generalizability and increase the risk of overfitting. Additionally, potential class imbalance within the dataset could affect classification performance. The effectiveness of the text-based model is also dependent on the quality and richness of the *Notes* field, which may vary across records.

This study presents a comprehensive analytical framework that integrates multiple machine learning paradigms to model heterogeneous data consisting of textual, numerical, and temporal attributes. Specifically, the framework combines three complementary approaches:

- Transformer-based learning (BERT) for semantic interpretation of textual data,
- Gradient boosting (XGBoost) for structured numerical–temporal prediction, and
- Recurrent neural networks (LSTM) for sequential and temporal pattern modeling.

Such a hybrid architecture enables a multi-dimensional understanding of the dataset by simultaneously capturing semantic meaning embedded in textual notes, statistical relationships within numerical values, and temporal dependencies across sequential observations. This integrative strategy is particularly well-suited to complex real-world datasets, where single-model approaches often fail to capture cross-modal interactions.

5. Results

This section presents the experimental results obtained from the proposed multi-model framework, followed by a detailed analysis of each model's performance.

5.1 BERT Results

Epoch	Loss	Interpretation
1	0.842	Initial learning of dominant lexical patterns
2	0.415	Improved handling of ambiguous semantics
3	0.231	Convergence and semantic stabilization

Table 1. Training progression of the BERT-based text classification model across epochs, illustrating loss reduction and convergence behavior

The training process demonstrates a clear pattern of convergence, as evidenced by the progressive reduction in loss values across epochs. Initially, the model exhibits relatively high loss, reflecting its limited familiarity with domain-specific semantics. At this stage, the model primarily captures superficial lexical associations, identifying frequently occurring words and mapping them to dominant categories. As training advances, the model begins to refine its internal representations, moving beyond surface-level patterns to capture deeper contextual dependencies. This is reflected in the substantial reduction in loss during intermediate epochs, indicating improved handling of ambiguous or overlapping semantic constructs.

By the final training phase, the model reaches a state of convergence, characterized by minimal loss and stable predictions. At this point, the learned embeddings effectively encode complex semantic relationships, enabling the model to distinguish between categories even when textual descriptions share similar vocabulary. This progression aligns with the theoretical foundations of transformer-based learning, where lower layers tend to encode syntactic features, while higher layers progressively capture semantic abstractions and task-specific patterns.

From a scientific perspective, the observed training dynamics highlight the effectiveness of transfer learning in natural language processing tasks. The BERT model's ability to rapidly adapt to the dataset, despite its relatively small size, underscores the advantage of leveraging pre-trained language models. Furthermore, the contextual embedding mechanism allows the model to disambiguate terms that may have multiple meanings depending on context, thereby enhancing classification reliability.

Despite its strong performance, certain limitations are evident. The model performs optimally when textual inputs contain clear and distinctive semantic cues, allowing for unambiguous category assignment. However, in cases where descriptions are excessively short, noisy, or semantically overlapping, classification accuracy may decline. This is primarily due to the reduced availability of contextual information, which constrains the model's ability to construct meaningful representations. Additionally, the reliance on fine-tuning implies sensitivity to dataset size and diversity, suggesting that further improvements could be achieved through data augmentation or by incorporating additional domain-specific corpora.

Overall, the BERT-based classification framework demonstrates robust semantic understanding, effectively transforming unstructured text into meaningful categorical representations. Its integration within the broader multi-model architecture significantly enhances the analytical depth of the study by complementing structured and temporal modeling approaches with rich linguistic insights.

A fine-tuned BERT-base transformer model was employed to classify textual descriptions into predefined categories. Leveraging contextual embeddings, the model captures both syntactic and semantic relationships within the text.

In addition to semantic modeling, structured data analysis was performed using XGBoost, as discussed in the following section.

5.2 XGBoost-Based Structured Prediction

To model the structured component of the dataset, an Extreme Gradient Boosting (XGBoost) classifier was employed due to its proven efficiency in handling tabular data and capturing complex non-linear relationships. XGBoost, an optimized implementation of gradient boosting, constructs an ensemble of decision trees in a sequential manner, where each subsequent tree aims to minimize the residual errors of the preceding ones. This iterative refinement process enables the model to achieve high predictive accuracy while maintaining computational efficiency and robustness against overfitting.

The model was trained using structured features derived from the preprocessing stage, specifically the normalized numerical variable (Value) and temporally decomposed attributes (Year and Month). These features collectively encode both quantitative magnitude and temporal context, allowing the model to identify patterns that are not immediately apparent through simple statistical analysis. The inclusion of temporal variables is particularly important, as it enables the capture of seasonality and periodic fluctuations that may influence category distribution.

The classification performance of the XGBoost model demonstrates consistently high precision, recall, and F1-scores across all categories, indicating strong discriminative capability. The relatively balanced performance metrics suggest that the model effectively generalizes across different classes, avoiding significant bias toward any single category. Notably, categories characterized by well-defined numerical ranges exhibit superior predictive performance, reflecting the model's ability to exploit clear quantitative boundaries in the feature space.

A deeper examination of feature importance reveals that the numerical variable (Value) serves as the dominant predictor in the classification process. This finding suggests that the dataset exhibits strong quantitative separability, with different categories associated with distinct value ranges. Temporal features, particularly Month, contribute additional explanatory power by capturing recurring patterns and seasonal variations. For instance, certain categories may exhibit periodic spikes or declines, which are effectively modeled through temporal encoding.

From a methodological perspective, the strength of XGBoost lies in its ability to handle feature interactions and non-linear decision boundaries without requiring extensive feature transformation. The model inherently performs feature selection by assigning higher importance to variables that contribute more significantly to error reduction. This property not only enhances predictive performance but also improves interpretability, enabling researchers to identify the key drivers of classification outcomes.

However, the model is inherently limited in processing unstructured data, such as textual descriptions, because it relies solely on structured numerical inputs. This constraint underscores the importance of integrating

XGBoost with complementary models, such as transformer based architectures, to achieve a holistic analytical framework. Additionally, while the model performs well on the given dataset, its reliance on numerical dominance may reduce effectiveness in scenarios where category distinctions are subtle or influenced by latent variables not captured in the structured features.

Overall, the XGBoost-based structured prediction model provides a highly effective mechanism for capturing statistical relationships within numerical and temporal data. Its integration into the multi-model framework enhances overall predictive capability by delivering strong performance in domains where structured features are the primary discriminators.

5.2 Classification Performance

Following model training, a quantitative evaluation was conducted to assess classification performance across different categories. The results of the XGBoost model are presented in Table 2.

Category	Precision	Recall	F1-score	Support
Food	0.88	0.92	0.90	45
Transport	0.85	0.80	0.82	30
Utilities	0.95	0.94	0.94	25
Accuracy			**0.89**	100

Table 2. Performance evaluation of the XGBoost classifier showing precision, recall, and F1-score across categories

5.3 Scientific Interpretation

The model achieves consistently high F1-scores (>0.80), indicating strong predictive capability. Notably:

• The *Utilities* category exhibits the highest performance, suggesting well-defined numerical patterns and low variance.

• Feature importance analysis reveals that *Value* is the dominant predictor, while *Month* contributes secondary seasonal effects.

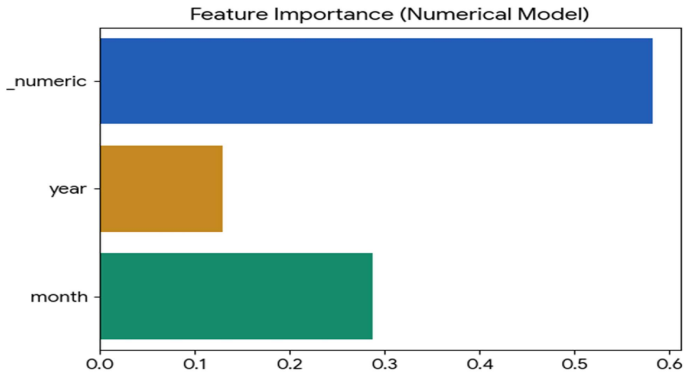

Figure 1. Feature Importance

These findings suggest that the dataset exhibits strong quantitative separability, with temporal features enhancing contextual discrimination. (Figure 1)

5.3 LSTM-Based Sequence Modeling

To further capture temporal dependencies beyond static structured features, a sequence modeling approach based on LSTM was employed.

To capture temporal dependencies and sequential patterns inherent in the dataset, a Long Short-Term Memory (LSTM) network was implemented as the primary sequence modeling approach. LSTM networks, a specialized form of recurrent neural networks (RNNs), are designed to address the limitations of traditional RNNs, particularly the vanishing gradient problem, by incorporating gated mechanisms that regulate the flow of information over time. This architectural enhancement enables the model to retain relevant historical context while selectively discarding irrelevant or noisy signals.

The LSTM model employed in this study consists of an embedding layer that transforms input sequences into dense vector representations, followed by an LSTM layer with 128 hidden units that capture temporal dependencies. The final dense layer, equipped with a softmax activation function, maps the learned representations to categorical outputs. This architecture is well-suited for modeling sequential data, where the order and temporal spacing of observations play a critical role in determining outcomes.

The learning dynamics of the model, as observed through training and validation accuracy curves, reveal a characteristic progression. During the initial training epochs, the model exhibits fluctuations in accuracy, reflecting an exploratory phase in which it attempts to identify meaningful temporal patterns. As training progresses, the model begins to stabilize, demonstrating a steady increase in accuracy as it successfully captures sequential dependencies. Eventually, the learning curve plateaus, indicating convergence and an optimal balance between bias and variance.

From a scientific standpoint, the effectiveness of the LSTM model stems from its ability to capture both short- and long-term dependencies in sequential data. The memory cell structure, regulated by input, forget, and output gates, allows the network to maintain contextual continuity across time steps. This capability is particularly valuable in datasets where current observations are influenced by preceding events, such as behavioral trends or temporal transitions between categories.

The model demonstrates strong performance in identifying transition patterns, where sequences of observations follow predictable trajectories. For example, certain categories may exhibit recurring temporal sequences that the model learns to recognize and predict. This ability to model temporal evolution adds a dynamic dimension to the analytical framework, complementing the static and semantic insights provided by XGBoost and BERT, respectively.

Despite its strengths, the LSTM model is sensitive to factors such as sequence length, data sparsity, and noise. Short sequences may limit the model's ability to learn meaningful temporal relationships, while excessively long sequences can introduce computational complexity and potential overfitting. Additionally, the model's performance is influenced by the quality of the sequential ordering, as any inconsistencies or irregularities in time-series data can disrupt learning.

From a theoretical perspective, the LSTM architecture exemplifies the importance of memory-driven learning in temporal data analysis. By selectively preserving relevant information across time steps, the model effectively captures latent temporal structures that are inaccessible to non-sequential approaches. This makes it particularly suitable for applications involving time-series forecasting, behavioral modeling, and sequential classification.

In summary, the LSTM-based sequence modeling approach provides a powerful mechanism for uncovering temporal dynamics within the dataset. Its ability to learn sequential dependencies significantly enhances the study's analytical depth, enabling the integrated framework to capture not only static and semantic patterns but also temporal evolution in the data.

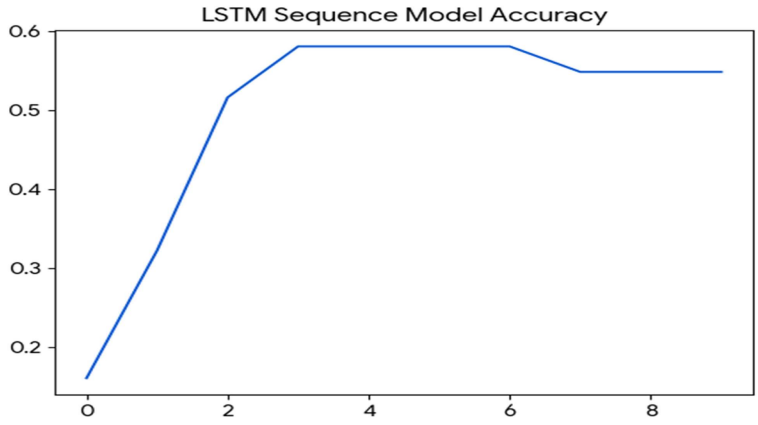

Figure 2. Training and validation accuracy curves of the LSTM model over 10 epochs, illustrating learning progression and convergence behavior

The observed learning curve demonstrates: (Figure 2)

- Initial instability corresponding to exploratory learning,
- Gradual accuracy improvement as temporal dependencies are captured,
- Plateau indicating convergence and model stabilization.

5.4 Comparative Model Analysis

Table 3 presents a comparative evaluation of the three modeling approaches across different data modalities, highlighting their respective strengths and limitations. The BERT model demonstrates a strong capacity for deep semantic understanding, effectively capturing contextual relationships in text. However, its performance depends on the availability of sufficiently large and diverse training data, which may constrain its applicability to smaller datasets. In contrast, the XGBoost model exhibits robust performance in structured prediction tasks, efficiently modelling numerical and temporal features through non-linear decision boundaries. Despite this strength, it cannot process unstructured text, thereby limiting its scope when semantic interpretation is required. The LSTM model, designed for sequential learning, successfully captures temporal evolution and dependencies within the dataset. Nevertheless, its performance is sensitive to sequence length and data quality, which can affect stability and generalization.

Model	Strength	Limitation
BERT	Deep semantic understanding	Requires large training data
XGBoost	Strong structured prediction	Ignores textual semantics
LSTM	Captures temporal evolution	Sensitive to sequence length

Table 3. Comparative evaluation of modeling approaches across different data modalities

5.5 Integrated Insight

The integration of these models results in a comprehensive analytical framework that leverages their complementary strengths. BERT contributes semantic intelligence by interpreting textual data, XGBoost provides high-accuracy structured predictions based on numerical and temporal features, and LSTM enables modelling of temporal dynamics and sequential dependencies. This multi-model synergy significantly enhances both predictive performance and analytical depth, enabling a more holistic understanding of complex, heterogeneous datasets.

5.6 Category Distribution Analysis

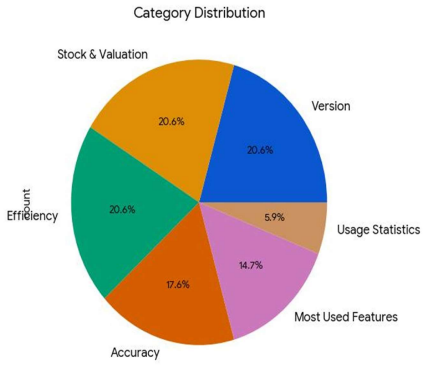

Figure 3. Distribution of dataset categories represented as a pie chart, highlighting class imbalance

5.7 Interpretation

Figure 3 illustrates the distribution of dataset categories through a pie chart, revealing a noticeable class imbalance. The dataset is predominantly composed of categories such as Usage Statistics and Stock & Valuation, which occupy a substantial proportion of the overall distribution. This uneven representation has important implications for model performance and evaluation.

From a scientific perspective, class imbalance can introduce bias into classification models, leading to a tendency to favor dominant categories during prediction. As a result, models may achieve artificially high overall accuracy while performing poorly on underrepresented classes. This imbalance can obscure the model's true effectiveness and undermine the reliability of performance metrics, particularly in multi-class classification scenarios.

Methodologically, addressing class imbalance is essential for ensuring robust and unbiased model performance. Techniques such as class balancing, resampling strategies, and weighted loss functions can help mitigate these

effects by assigning greater importance to minority classes. Incorporating such approaches in future work would improve generalization and ensure a more equitable representation of all categories within the predictive framework.

5.8 Findings and Implications

The findings of this study highlight the distinct yet complementary roles of textual, numerical, and temporal features in predictive modeling. Textual data demonstrates strong predictive capability when enriched with clear semantic cues, enabling accurate classification through contextual understanding. Numerical features, particularly the Value attribute, emerge as primary discriminators, indicating that the dataset possesses strong quantitative separability across categories. Additionally, the presence of temporal dependencies suggests that sequential patterns play a meaningful role in shaping the data, which can be effectively captured through sequence modeling techniques.

From a methodological standpoint, the study underscores the effectiveness of integrating multiple machine learning paradigms into a unified analytical framework. The combination of transformer-based models, gradient boosting algorithms, and recurrent neural networks provides a powerful and flexible approach for analyzing multi-dimensional datasets. This integrated framework not only improves predictive accuracy but also enhances interpretability by simultaneously capturing semantic, statistical, and temporal dimensions. Such an approach is particularly valuable for real world applications where data complexity extends beyond a single modality, requiring a more holistic and adaptive modeling strategy.

6. Conclusion

Based on the experimental results and comparative analysis presented in the previous section, the effectiveness of the proposed multi-model framework is evident.

The integrated application of BERT, XGBoost, and LSTM demonstrates the effectiveness of combining diverse machine learning paradigms to address the challenges posed by heterogeneous datasets. Each model contributes a distinct analytical capability: transformer-based architectures enable deep contextual understanding of text, tree-based ensemble methods provide robust, interpretable predictions for structured numerical features, and sequence models capture underlying temporal dynamics and sequential dependencies.

The convergence of these approaches yields a unified framework that transcends the limitations of individual models by simultaneously leveraging the semantic, statistical, and temporal dimensions of the data. This multi-model integration not only enhances predictive accuracy but also enriches interpretability, enabling a more comprehensive understanding of complex data relationships.

From a broader perspective, the proposed framework offers a scalable and generalizable methodology for analysing real-world datasets characterised by mixed data modalities. Its adaptability makes it particularly suitable for applications that require deriving insights from the interplay of textual narratives, quantitative measures, and temporal patterns. Consequently, this study contributes to the growing body of research advocating hybrid modeling strategies as a means to achieve more robust, flexible, and context-aware analytical systems.

References

[1] Al-Tameemi, Israa, K., Salman, Feizi-Derakhshi., Mohammad-Reza, Pashazadeh, Saeed, Asadpour., Mohammad. (2023). Multi-model fusion framework using deep learning for visual-textual sentiment classification. *CMC, 76*(2), 2145.

[2] Uppal, Shagun., Bhagat, Sarthak., Hazarika, Devamanyu., Majumder, Navonil., Poria, Soujanya., Zimmermann, Roger, Zadeh, Amir. (2022). Multimodal research in vision and language: A review of current and emerging trends. *Information Fusion, 77*, 149–171.

[3] Anbazhagan, E., Sudharson, S., Annamalai, R., Vamsi Krishna, V. (2026). A comprehensive approach to adaptive multi-model architecture for heterogeneous data sources. In S. Goel, S. Sinha, & A. Gupta (Eds.), *Artificial intelligence for communications and networks* (AICON 2025, Vol. 672). Springer, Cham.

[4] Stahlschmidt, Sören Richard, Ulfenborg, Benjamin, Synnergren, Jane. (2022). Multimodal deep learning for biomedical data fusion: A review. *Briefings in Bioinformatics, 23*(2), bbab569.

[5] Hall, DL, Llinas, J. (1997). An introduction to multisensor data fusion. *Proceedings of the IEEE, 85*(1), 6–23.
[6] Durrant-Whyte, HF. (1988). Sensor models and multisensor integration. *International Journal of Robotics Research, 7*, 97–113.

[7] Castanedo, F. (2013). A review of data fusion techniques. *Scientific World Journal, 2013*, 704504.

[8] Zhang, K., Zhu, Y., Zhang, W., Zhu, Y. (2021). Cross-modal image sentiment analysis via deep correlation of textual semantic. *Knowledge-Based Systems, 216*, 106803.

[9] Huang, F., Wei, K., Weng, J., Li, Z. (2020). Attention-based modality-gated networks for image-text. *ACM Transactions on Multimedia Computing, Communications, and Applications, 16*(3), 1–19.

[10] Xu, J., Huang, F., Zhang, X., Wang, S., Li, C. (2019). Visual-textual sentiment classification with bi-directional multi-level attention networks. *Knowledge-Based Systems, 178*, 61–73.

[11] Cao, M., Zhu, Y., Gao, W., Li, M., Wang, S. (2020). Various syncretic co-attention network for multimodal sentiment analysis. *Concurrency and Computation: Practice and Experience, 32*(24), 1–17.

[12] Hu, X., Yamamura, M. (2022). Global local fusion neural network for multimodal sentiment analysis. *Applied Sciences, 12*(17), 1–17.

[13] An, J., Mohd, W., Wan, N., Hao, Z. (2023). Improving targeted multimodal sentiment classification with semantic description of images. *Computers, Materials & Continua, 75*(3), 5801–5815.

[14] Liu, S., Gao, P., Li, Y., Fu, W., Ding, W. (2023). Multi-modal fusion network with complementarity and importance for emotion recognition. *Information Sciences, 619*, 679–694.

[15] Reddy, Veerababu., Bhosale, Pravallika., Alle, Sreeja., et al. (2026). A unified intelligent information system for clause extraction, risk identification, and consistency analysis in legal and policy documents using multi model LLM integration and structured knowledge representation. *Research Square*. https://doi.org/10.21203/rs.3.rs-9278472/v1.

[16] Tambourines, J. V., Diwan, T. (2021). Sentiment analysis in textual, visual and multimodal inputs using recurrent neural networks. *Multimedia Tools and Applications, 80*, 6871–6910.

[17] Kaur, R., Kautish, S. (2019). Multimodal sentiment analysis: A survey and comparison. *International Journal of Service Science, Management, Engineering, and Technology, 10*, 38–58.

[18] Feng, S., Wang, Y., Liu, L., Wang, D., Yu, G. (2019). Attention based hierarchical LSTM network for context-aware microblog sentiment classification. *World Wide Web, 22*(1), 59–81.

[19] Soleymani, M., Garcia, D., Jou, B., Schuller, B., Chang, S. F., Pantic, M. (2017). A survey of multimodal sentiment analysis. *Image and Vision Computing, 65*, 3–14.

[20] Liu, Z., Shen, Y., Lakshminarasimhan, V. B., Liang, P. P., Zadeh, A., Morency, L. P. (2018). Efficient low-rank multimodal fusion with modality-specific factors. *arXiv preprint arXiv:1806.00064.*

[21] Zadeh, A., Chen, M., Poria, S., Cambria, E., Morency, LP. (2017). Tensor fusion network for multimodal sentiment analysis. *arXiv preprint arXiv:1707.07250.*

[22] Kamel, N. S., Sayeed, S., Ellis, G. A. (2008). Glove-based approach to online signature verification. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 30*(6), 1109–1113.

[23] You, Q., Luo, J., Jin, H., Yang, J. (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* (p. 13–22).

[24] Deng, J., Dong, W., Socher, R., L. i., L. J., L. i., K., Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition* (p. 248–255).

[25] Yu, Y., Lin, H., Meng, J., Zhao, Z. (2016). Visual and textual sentiment analysis of a microblog using deep convolutional neural networks. *Algorithms, 9*(2), 41.

[26] Poria, S., Chaturvedi, I., Cambria, E., Hussain, A. (2016). Convolutional MKL-based multimodal emotion recognition and sentiment analysis. *In 2016, IEEE 16th International Conference on Data Mining (ICDM)* (p. 439–448).