International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(10s) 2026 pp. 606-626
https://doi.org/10.70917/ijcisim-2026-3641
Article

AMTF-Net: An Adaptive Multi-Task Transformer Fusion Network for Joint Sentiment Analysis and Sarcasm Detection

M.L.S.N.S Lakshmi[1*], Sunita S. Padmannavar[2], M Siva Prasad[3], Udayalaxmi Aditya Teki[4], T. Hemalatha[5], Revathi Talari[6], Sampathirao Yoganandh[7]

[1*]Department of Electronics and Communication, Ramachandra College of Engineering, Vatluru, Eluru, Andhra Pradesh 534007, mlakshmi.0290@gmail.com.
[2]Department of MCA, KLS Gogte Institute of Technology, Belagavi, Karnataka, India, sspadmannavar@git.edu
[3]Dept. of CSE- AI & ML, Geethanjali College of Engineering and Technology, Keesara (M), Medchal (D), Telangana, mshivaprasad.cse@gcet.edu.in
[4]Department of Computer Applications, Aditya University, Surampalem, Andhra Pradesh-533437, India, udayalakshmiaditya@gmail.com
[5]Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Greenfields, Vaddeswaram, Guntur, Andhra Pradesh - 522 302, thonepudihemalatha@gmail.com
[6]Department of Computer Science and Engineering, Dr RVR NRI Institute of Technology Deemed to be University, Agiripalli, Pothavarappadu, Vijayawada, Andhra Pradesh -521212, revathi.chitti2@gmail.com
[7]Department of CSE (AI & ML, DS), ANITS, Vidsakapatnam, AP, India, syoganandh.csm@anits.edu.in

Abstract: With the widespread use of social media platforms, there is a massive amount of user-generated text which necessitates performing sentiment analysis, a natural language processing task. Nevertheless, correctly identifying the sentiment proves to be a difficult task because of the occurrence of sarcasm where the literal sentiment is opposite to the intended meaning. Prior research has either addressed sentiment analysis and sarcasm detection as separate tasks or developed static ensemble methods that do not exploit the semantic relationship between the two tasks. To overcome these drawbacks, this work presents AMTF-Net (Adaptive Multi-Task Transformer Fusion Network), a unified multi-task learning framework capable of jointly performing sentiment analysis and sarcasm detection. The framework extracts complementary contextual representations using BERT, GPT-2, and XLNet. The extracted contextual representations are fused through a MATF module that adaptively learns the contribution of each transformer for every input. The Shared Semantic Encoder (SSE) and Cross-Task Attention Interaction (CTAI) module refine the fused features to allow knowledge sharing between the two tasks. A Dynamic Joint Loss (DJL) function is used, which adaptively balances the optimization objectives of both tasks. The three public benchmark datasets employed were Twitter, Reddit, and HuffPost news headlines. The experimental results show that AMTF-Net has superior performance against individual transformers, conventional ensemble methods, and recent state-of-the-art methods. The proposed framework was evaluated on Twitter, Reddit, and HuffPost benchmark datasets for joint sentiment analysis and sarcasm detection. The proposed framework achieved an accuracy of 97.74 %, 98.12 % and 98.05 % respectively. Also, the Macro F1-scores were achieved at 98.12 %, 98.04 % and 97.82 % respectively for each dataset for sarcasm detection. These findings show that adaptive transformer fusion and cross-task semantic interaction enhance the robustness and accuracy of joint sentiment analysis and sarcasm detection.

Keywords: Sentiment analysis; Sarcasm detection; Multi-task learning; Transformer fusion; BERT; GPT-2; XLNet; Cross-task attention; Adaptive fusion; Natural language processing.

1. INTRODUCTION

The rapid growth of communication tools, particularly social media platforms such as Twitter, Reddit, and news forums has transformed human communication and created a large volume of user-generated text every day. The wealth of opinions, reviews, sentiments, and emerging trends that is generated on the social media is valuable for decision-making related to marketing, public policies and brand reputation management [1], [2]. In recent years, a large amount of subjective information has been generated. With the explosion of social media, we are overwhelmed by large volumes of unstructured data. Thus, automatic extraction of such subjective information from unstructured text has become a fundamental task in natural language processing (NLP) in modern times. Consequently, sentiment analysis has become one of the most widely researched tasks in NLP. Recent advances in deep learning [24] have enabled several researchers to develop effective models to classify the polarity (positive, negative, or neutral) of the given text (Sentiment analysis). Nevertheless, the usefulness of sentiment classifiers has been severely hampered by sarcasm, a complex linguistic phenomenon in which the intended meaning is opposite to that which is actually stated [3]. Sarcastic utterances, regulars on social networks to convey emphatic sentiments with the help of irony, hyperbole, and contextual mismatch, create a basic semantic inversion that severely confuses standard polarity classifiers. This intrinsic ambiguity gives rise to a critical bottleneck: effective understanding of sentiment relies on the correct detection of sarcasm, whereas effective detection of sarcasm relies on the sentiment grounded in context. Recently, transformer-based language models such as BERT [4], GPT-2 [5], and XLNet [6] have achieved state-of-the-art performance. Using self-attention mechanisms, these models have demonstrated remarkable performance on various NLP benchmarks like sentiment analysis and sarcasm detection. Most existing approaches, despite their individual merits, treat the two tasks independently, thus addressing them either in isolation or through oversimplified pipeline architectures that do not exploit their intrinsic interdependence [7]. At the same time, an increasing volume of research has investigated ensemble strategies that combine multiple transformer variants to exploit complementarity in representational capacities. However, traditional ensemble methods, e.g., majority-voting or stacking with a fixed weight, still commonly give static importance to each model [25], [26]. These methods do not consider the dynamic difference in feature relevance for different linguistic contexts or input instances. [8] Recognizing the connection between sentiment and sarcasm, several multi-task learning (MTL) frameworks have been proposed to facilitate parameter sharing [9], [10]. Although these methodologies have exhibited slight enhancements in relation to single-task baselines, they largely depend on basic parameter sharing (such as hard or soft sharing of encoder layers) without the addition of elaborate cross-task interaction components. Therefore, they cannot capture the asymmetric, bidirectional dependencies of the sarcastic sentiment reversal. Moreover, as task-specific loss weightings are usually static, this leads to sub-optimal equilibrium of the tasks which have different convergence dynamics and difficulty levels [11]. The difficulty of adaptively balancing task-specific gradients during training is still insufficiently tackled for sentiment and sarcasm co-prediction. These limitations, including the isolated modeling of sentiment and sarcasm, static fusion for multi-transformer ensembles, the absence of cross-task semantic interaction, and conventional loss balancing, represent significant research gaps. Inspired by these limitations, we propose AMTF-Net, a unified framework based on deep learning to address the above defects. In contrast to existing works, our framework employs an instance-adaptive fusion strategy that dynamically fuses contextual representations from BERT, GPT-2, and XLNet, effectively avoiding fixed ensemble weighting. Moreover, there is a cross-task interaction module in the architecture. This enables bidirectional information exchange between the two modules. In other words, the sentiment branch helps reduce noise. To guarantee stable and balanced optimization, a dynamic joint loss function is proposed to adaptively modulate task contributions during the training. This work makes four main contributions.

1. We suggest a new method of transformer fusion which can automatically weight importance of various pre-trained language models depending on the input context. This reflects the inadequacies of existing majority-voting and static ensemble methods.
2. We develop a cross-task attention interaction module with shared semantic encoder that explicitly models the bidirectional dependencies between sentiment classification and sarcasm detection to ensure robust feature transfer without losing task-specific outputs.
3. The introduction of this dynamic joint loss function helps to automatically balance and optimize the network on both tasks while reducing the instability during training and improving performance on the test set without the need for manual adjustment of hyper-parameters.
4. We perform comprehensive empirical evaluations on 3 publicly available datasets – Twitter, Reddit as well as the Headlines of HuffPost News. The experimental results confirm that our AMTF-Net has a consistent performance enhancement over its state-of-the-art individual transformers, traditional ensembles, and recent MTL baselines on various metrics.

The structure of the remainder of this paper is as follows. In Section 2, we critically review the related literature and identify the gaps in the literature that motivate our work. The design details of AMTF-Net are presented in section 3. Section 4 describes the experiment set up, datasets and implementation. Section 5 covers the empirical results and their discussion alongside ablation and sensitivity analyses. Finally, in Section 6, the paper concludes by summarizing the findings and suggesting suitable avenues for further research.

2. RELATED WORK

This section offers a critical synthesis of existing literature, structured in a way to highlight progressive research trajectories and more importantly, identify the gaps that our framework will fill. The existing works can be divided into the following categories: transformer-based single-task models, multi-tasking for affect analysis, feature fusion and cross-task interaction, and dynamic optimization strategies.

2.1 Transformer Models for Sarcasm and Sentiment

The introduction of the Transformer architecture [12] marked a turning point in NLP, leading to subsequent PLMs with unmatched contextual awareness. According to [4], the BERT implementation from Google established bidirectional representations by masking language modeling and next-sentence prediction tasks. Since then, BERT has been extensively fine-tuned for sentiment classification. The XLNet [6] architecture utilizes permutation language modeling to learn dependencies from both the left and right. This provides beneficial complementary effects in processing syntactically complicated constructions, such as those containing negation, which are often found in subtleties of sarcasm. Due to its capacity to model sequential incongruity in an autoregressive fashion, GPT-2 [5] has been found to perform well in sarcasm detection, despite being unidirectional [13].

Many researchers have experimented with applying ensemble methods to pool the power of these models. For example, stacking ensembles with XLNet, BERT, RoBERTa have been proposed for sarcasm identification which provide modest performance improvements [14]. The authors [15] applied majority-voting ensembles of BERT, RoBERTa, and ALBERT to a multi-class sentiment task. Although these ensemble methods allow architectures to compensate for each other, ultimately, they are limited by the fixed weighting schemes. It is assumed that any particular transformer's relative importance is the same across all inputs. However, this assumption is often unrealistic based on empirical evidence. Different linguistic constructions such as hyperbolic sarcasm vs subtle irony render some representations more important than others on the instance level. The gap in research for adaptive, input-dependent fusion remains largely unexplored.

2.2 Multi-task Learning for Joint Sentiment and Sarcasm Modeling

The strong relationship between sentiment and sarcasm has led to the design of MTL frameworks that go beyond single tasks. Majumder et al. [9] showed that joint training of sentiment and sarcasm classifiers yields a 3–4% improvement over separate training and that the improvement is due to the sharing of complementary features. El Mahdaouy et al. [10] then extended to Arabic social media text classification by integrating Bert-based encoder with a multi-task attention mechanism for language-specific tasks. Recent investigations into multimodal MTL settings have taken place that have classified sentiment, sarcasm, emotion, and humor from audio-visual-textual data [16], [17].

Context dependency and modality fusion is what these models aim to address. Though proven effective, MTL approaches mainly make use of simple parameter sharing strategies like hard sharing of bottom transformer layers or soft sharing through regularization penalties. They lack explicit architectural components that represent high-level, bi-directional semantics between tasks. The phenomenon known as semantic reversal, where a literal positive statement has a negative meaning, requires combining attention across tasks that aligns contradictory information dynamically. Existing frameworks treat task-specific heads as independent classifiers inside a shared latent space As a result, task-specific feature refinement and gradient flow are not fully exploited. Consequently, task-specific representation learning remains limited.

2.3 Attention for Cross-Modal / Cross-Task Fusion

Through the process of integrating features from more than two modalities, one can analyze and predict sentiments. Gated fusion strategies [18] and exchange-based multimodal transformers [19] have been proposed for sarcasm handling based on the reliability of conditional contexts. In unimodal settings, researchers have previously explored whether the attention mechanism focusing on specific words in the input can jointly predict the sarcasm polarity and its intensity [20]. The results showed that the attentive interaction successfully captures the fine-grained

sarcasm patterns. Moreover, shared semantic encoders (SSEs) have been used in multiple MTL settings to map heterogeneous features into a common representational space.

Although these methods demonstrate the effectiveness of adaptive weighting and attentive interactions, they are still mostly limited to multi-modal scenarios or single-task variants. To our knowledge, the adaptive fusion among multiple transformers in the same text modality, a shared semantic encoder and cross-task attention (for bidirectional sentiment-sarcasm) has not been sufficiently explored. The current text corpus does not possess the dynamic gating capabilities of a multimodal model, and existing cross-task attention does not use a fused multi-transformer representation. The interdisciplinary gap of adaptive fusion, shared encoding, and cross-task attention in a single NLP framework is a novel and substantial gap.

2.4 Dynamic Loss Weighting for Multi-Task Optimization

In MTL, it is well known that there are multiple objectives that need to be balanced because tasks have different scales of loss, convergence rates and gradient magnitudes. Static loss weighting necessitates extensive grid search and often leads to poor Pareto fronts. To remedy this, Kendall et al. [21] proposed to use weights based on the uncertainty, regarding task uncertainty as a learnable parameter that can scale the contribution to the loss. Gradient-based dynamic balancing methods [22] were subsequently proposed to prevent one task from dominating the update direction. The proposed loss for sentiment and sarcasm multitask was tested and it provided marginal improvements [23]. Though current methods of dynamic weight are mostly task-agnostic and fail to consider the unique semantic opposing forces of sentiment and sarcasm.

The dynamics of sarcasm detection, which involves modeling reversal in context, are different from polarity classification. Existing methodologies do not account for a task-specific adaptation that factors in polarity-reversal risk. Consequently, the literature lacks a dynamic joint loss that can adequately balance these two connected but asymmetric demanding objectives.

2.5 Research Gap & Motivation for AMTF-Net

Integrating the above critical discussion, we have three related and distinct research gaps.

- G1 (Fusion Rigidity): The existing transformer ensembles adopt static or majority-voting fusion which is essentially incapable of accommodating the contextual saliency of different PLM representations across varying inputs.
- G2 (Interaction Deficiency): Existing MTL (Multi-Task Learning) frameworks for sentiment and sarcasm often do not have explicit cross-task attention mechanisms and consequently do not capture the required mutual semantic reversal dynamics that are crucial for robust joint inference.
- G3 (Optimization Imbalance): Conventional loss weighting strategies, whether static or generic dynamic, ignore the optimization asymmetry due to polarity-reversal nature of sarcasm and lead to subcritical equilibria in the training process.

In this work, we propose a Multi-Head Adaptive Transformer Fusion (MATF) module that dynamically recalibrates the feature contribution per instance for each task, addressing G1. Further, we introduce a Shared Semantic Encoder (SSE) along with a Cross-Task Attention Interaction (CTAI) module for explicitly modeling the two-way interaction of task dependencies while preserving the task-specific discriminative pathways, tackling G2. Finally, we develop a Dynamic Joint Loss (DJL) function that adaptively regulates the gradients of different tasks dependent on their asymmetrical convergence characteristics, addressing G3. Through an all-encompassing integration of these innovations, our work provides a coherent and scalable solution to the core of the joint sentiment-sarcasm analysis challenge, thereby establishing a new benchmark for transformer-based multi-task NLU.

3. PROPOSED METHODOLOGY

This section presents the proposed AMTF-Net framework. To provide conceptual clarity, we give a formal problem formulation followed by a high-level workflow. Subsequently, each building block is described with thorough formulas, explicit tensor dimension analysis and algorithms. It concludes with training protocols, complexity analysis, and implementation details. The high-level design philosophy of AMTF-Net aims to jointly overcome three major limitations of the aforementioned methods (i) constrained static ensemble fusion policy, (ii) lack of explicit cross-task semantic interaction, and (iii) compromised multi-task optimization balance.

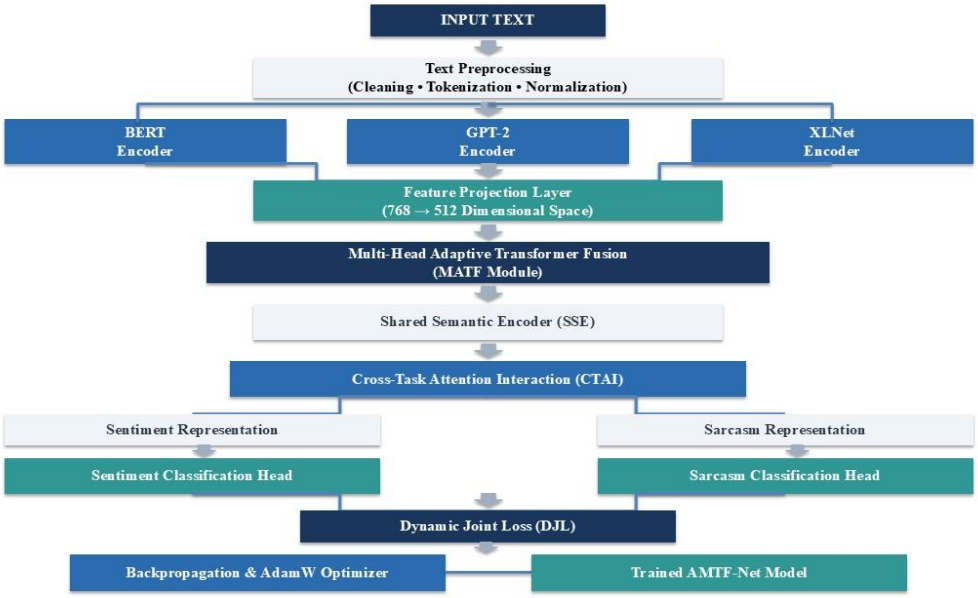


Figure 1: AMTF-Net Model Architecture

3.1 Problem Formulation

Let $\mathcal{D} = \{(x_i, y_i^s, y_i^c)\}_{i=1}^N$ be a dataset of N text instances $x_i = \{w_1, w_2, \ldots, w_T\}$ of length T, where each instance is associated with a sentiment label $y_i^s \in \mathcal{Y}_s = \{\text{Positive}, \text{Neutral}, \text{Negative}\}$ and a sarcasm label $y_i^c \in \mathcal{Y}_c = \{\text{Sarcastic}, \text{Non-sarcastic}\}$. The objective is to learn a unified mapping function $\mathcal{F}: \mathcal{X} \to \mathcal{Y}_s \times \mathcal{Y}_c$ that concurrently predicts both labels. Unlike single-task or static ensemble paradigms, our framework seeks to optimize the joint conditional probability:

$$P(y^s, y^c, | x\Theta) = P(y^s; | x\Theta) \cdot P(y^c; | x\Theta) \cdot \Phi(y^s, y^c) \qquad (1)$$

where $\Phi(\cdot)$ identifies the underlying dependence of sentiment polarity on sarcastic intention, and Θ is the collection of all trainable parameters of all the modules. This joint formulation explicitly recognizes that sarcastic utterances reverse the sentiment polarity, and the predictions for one task will optimize through the other task's predictions. The term coupling $\Phi(\cdot)$ is implicitly modeled with the Cross-Task Attention Interaction (CTAI) module, where information can be shared in both directions in training or inference.

3.2 Overall Architecture and Workflow

The architecture of AMTF-Net is depicted in Figure 1, which jointly performs sentiment analysis on comments along with sarcasm detection. In the first step, the input text is tokenized with the respective tokenizers of BERT, GPT-2, and XLNet, and positional information is added to retain sequence information. The three pre-trained transformer encoders process these tokenized inputs in parallel to extract complementary contextual representations. The proposed Multi-Head Adaptive Transformer Fusion (MATF) module dynamically learns weights based on the input to fuse these features into a unified representation. The Shared Semantic Encoder (SSE) then improves the fused features, yielding a task-agnostic semantic representation, and later integrates them with the Cross-Task Attention Interaction (CTAI), which effectively exchanges information between sentiment analysis and sarcasm detection. Ultimately, the classification heads that are specific to the task make the relevant predictions, and a dynamic joint loss (DJL) function jointly optimizes both tasks for stability and generalization. The training procedures and inference are given in Algorithm 1 and Algorithm 2, respectively.

3.3 Contextual Feature Extraction

AMTF-Net relies on a range of complementary transformer-based language models. Each transformer captures complementary linguistic characteristics, and their adaptive fusion creates a representation that no single model can achieve by itself.

3.3.1 Input Tokenization and Positional Encoding

Given an input text sequence $x = \{w_1, w_2, \ldots, w_T\}$, we tokenize using each transformer's respective tokenizer. The embedding vectors are mapped to token indices:

$$E = \text{Embed}(x) \in \mathbb{R}^{T \times d_{\text{emb}}} \qquad (2)$$

where d_{emb} is the embedding dimension (e.g., 768 for BERT-base). To preserve sequential information, which is essential for capturing syntactic and semantic dependencies, we add positional encodings:

$$E_{\text{pos}} = E + P,\ P \in \mathbb{R}^{T \times d_{\text{emb}}} \qquad (3)$$

where P is computed using sinusoidal functions as in [12]:

$$P_{(pos,2i)} = \sin\left(\frac{pos}{10000^{2i/d_{\text{emb}}}}\right),\ P_{(pos,2i+1)} = \cos\left(\frac{pos}{10000^{2i/d_{\text{emb}}}}\right) \qquad (4)$$

This positional encoding strategy ensures that the model can distinguish between different token positions, which is particularly important for sarcasm detection where the placement of negation or contrastive phrases often determines interpretation.

3.3.2 Parallel Transformer Encoding

There are three parallel encoders that are transformer-based language models. Section 3.11 elaborates the reasoning behind the selection of these models. Each model accepts the positional-encoded entry to deliver contextual representations.

BERT Encoder: Extracts bidirectional representations via masked language modeling objectives, enabling attention for each token to both left and right contextual information:

$$H_B = \text{BERT}(E_{\text{pos}}) \in \mathbb{R}^{T \times d_B},\ d_B = 768 \qquad (5)$$

GPT-2 Encoder: The GPT-2 Encoder's autoregressive representations capture left-to-right contextual dependencies, which makes it particularly effective at modeling sequential incongruity, which is often found in sarcastic constructions:

$$H_G = \text{GPT-2}(E_{\text{pos}}) \in \mathbb{R}^{T \times d_G},\ d_G = 768 \qquad (6)$$

XLNet Encoder: Employs permutation-based modeling to leverage bidirectional dependencies through factorization-order permutations that overcome the independence assumption of BERT's masked language modeling:

$$H_X = \text{XLNet}(E_{\text{pos}}) \in \mathbb{R}^{T \times d_X},\ d_X = 768 \qquad (7)$$

3.3.3 Pooling and Projection

To ensure that the dimensionality across the three heterogeneous transformers is uniform and to obtain a fixed-size sentence representation suitable for downstream classification, we mean-pool over the sequence dimension.

$$h_B = \frac{1}{T}\sum_{t=1}^{T} H_B[t,:] \in \mathbb{R}^{768},\ h_G = \frac{1}{T}\sum_{t=1}^{T} H_G[t,:] \in \mathbb{R}^{768},\ h_X = \frac{1}{T}\sum_{t=1}^{T} H_X[t,:] \in \mathbb{R}^{768} \qquad (8)$$

Mean-pooling is preferred over CLS-token extraction because: (i) it is more stable across varying sequence lengths; (ii) it distributes the importance across all tokens as opposed to a single special token; (iii) it has been proven to yield better performance for sentence-level tasks in previous work.

The representations are then mapped to a common latent space of size d=512 through learnable linear projections.

$$p_B = W_B h_B + b_B \in \mathbb{R}^{512},\ p_G = W_G h_G + b_G \in \mathbb{R}^{512},\ p_X = W_X h_X + b_X \in \mathbb{R}^{512} \qquad (9)$$

The projection step serves two purposes of (i) reducing the dimensionality from 768 to 512 to reduce computational complexity, and (ii) aligning the three different representation spaces to a common frame of reference for effective fusion via the downstream modules.

3.4 Multi-Head Adaptive Transformer Fusion (MATF)

Existing ensemble methods are fundamentally limited in their performance as they rely solely on fixed weighting schemes, like majority voting or fixed stacking, that assign equal importance to all models regardless of the input. It is particularly problematic for sarcasm detection where different linguistic phenomena, certain transformers may become more informative than others on a per instance basis. For instance, GPT-2's autoregressive sequential modeling may do better with hyperbolic sarcastic statements while XLNet's permutation-based bidirectional understanding may be more suited for subtle ironic expressions. To address this challenge, we introduce the MATF module that utilizes a multi-head gating mechanism to compute input-dependent fusion weights.

3.4.1 Gating Network

We concatenate the projected representations to form a comprehensive feature vector:

$$z = [p_B; p_G; p_X] \in \mathbb{R}^{1536} \qquad (10)$$

This concatenated vector is passed through a two-layer feed-forward gating network with ReLU activation:

$$\alpha' = W_2 \cdot \mathrm{ReLU}(W_1 z + b_1) + b_2 \in \mathbb{R}^{3K} \qquad (11)$$

where $W_1 \in \mathbb{R}^{d_g \times 1536}$, $W_2 \in \mathbb{R}^{3K \times d_g}$, $d_g = 128$ is the gating hidden dimension, and $K = 4$ is the number of adaptive heads. By having multiple heads, the model captures the different aspects of transformer relevance at once, which is similar to multi-head attention capturing different semantic relations at once.

3.4.2 Multi-Head Weight Generation

For the k-th head ($k \in \{1, \ldots, K\}$), we extract the corresponding weight logits and apply softmax normalization:

$$\alpha_k = \mathrm{Softmax}(\alpha'_k) \in \mathbb{R}^3,\ \alpha_k = \left[\alpha_{k,B}, \alpha_{k,G}, \alpha_{k,X}\right]^T \qquad (12)$$

where $\sum_{j \in \{B,G,X\}} \alpha_{k,j} = 1$. Each head produces a context-specific fused representation by taking a weighted sum of the projected transformer outputs:

$$h^k_{fused} = \alpha_{k,B} \cdot p_B + \alpha_{k,G} \cdot p_G + \alpha_{k,X} \cdot p_X \in \mathbb{R}^{512} \qquad (13)$$

Softmax normalization ensures that weights can be viewed as probabilities indicating which transformer is most relevant for the current input instance.

3.4.3 Aggregation with Residual Connection

The averaging process across heads, along with the addition of a residual connection and formulation of layer normalization, creates the finalized fused representation:

$$h_{fused} = \mathrm{LayerNorm}\left(p_B + \frac{1}{K}\sum_{k=1}^{K} h^k_{fused}\right) \in \mathbb{R}^{512} \qquad (14)$$

The residual connection p_B is utilized to stabilize the fusion by preventing it from straying too far from the representations learned by BERT, which have been shown to be robust across various tasks. This ensures that any extreme weights produced by the gating network are softened and the fused representation remains firmly grounded in the BERT representation that serves as a reliable baseline.

The main innovation that MATF integrates over conventional ensembles is that it is input-adaptive. The fusion weights are recomputed for every input that is processed. So, the fusion weights to appropriately emphasize the right transformer for a specific text. The dynamic weighting is certainly more expressive than using a static weighting, where the same weights are used in all contexts.

3.5 Shared Semantic Encoder (SSE)

The fused representation h_{fused} is refined by a Shared Semantic Encoder with $L = 2$ transformer encoder layers. The SSE is a shared knowledge system that encapsulates common linguistic and pragmatic knowledge. It also helps avoid overfitting to either task by allowing both downstream tasks to leverage a common representation.

3.5.1 Multi-Head Self-Attention

For each layer $l \in \{1, \ldots, L\}$, with input $X^{(l)} \in \mathbb{R}^{1\times512}$ (tiled to sequence dimension if necessary), the self-attention is computed as:

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \qquad (15)$$

with $Q = X^{(l)}W_Q$, $K = X^{(l)}W_K$, $V = X^{(l)}W_V$, where $W_Q, W_K, W_V \in \mathbb{R}^{512\times512}$. The multi-head variant concatenates $h = 8$ attention heads, each with dimension $d_k = 64$.

3.5.2 Position-Wise Feed-Forward Network

The same transformation is applied independently to each position by a position-wise feed-forward network in each transformer layer:

$$\text{FFN}(x) = W_2^{(l)} \cdot \text{GELU}\left(W_1^{(l)}x + b_1^{(l)}\right) + b_2^{(l)} \in \mathbb{R}^{512} \qquad (16)$$

where $W_1^{(l)} \in \mathbb{R}^{2048\times512}$, $W_2^{(l)} \in \mathbb{R}^{512\times2048}$, and GELU is the Gaussian Error Linear Unit activation. The GELU activation function provides smoother nonlinear transformations than ReLU, resulting in improved learning performance.

3.5.3 Residual Connections and Layer Normalization

Each sub-layer incorporates a residual connection followed by layer normalization:

$$X^{(l+1)} = \text{LayerNorm}\left(X^{(l)} + \text{MHSA}\left(X^{(l)}\right)\right) \qquad (17)$$

$$X^{(l+2)} = \text{LayerNorm}\left(X^{(l+1)} + \text{FFN}\left(X^{(l+1)}\right)\right) \qquad (18)$$

The residual connections enable a direct path for gradient propagation throughout the network, alleviating the vanishing gradient issue prevalent in deep architectures. Layer normalization facilitates stable training by reducing internal covariate shift. After L layers, we obtain the shared semantic representation:

$$H_{shared} = X^{(L)} \in \mathbb{R}^{1\times512} \qquad (19)$$

In effect, the SSE is a soft parameter-sharing mechanism, where the underlying representations can be shared but there is enough capacity in subsequent modules for task-specific adaptation for both tasks.

3.6 Cross-Task Attention Interaction (CTAI)

While shared features are provided by the SSE, explicit bidirectional interaction is important to model the reversal of semantics that characterizes sarcasm. For example, in scenarios where the literal sentiment is positive but the utterance is sarcastic, the model must incorporate contradiction cues from the two tasks. The CTAI module tackles this challenge through the use of a cross-attention mechanism that allows for the bi-directional interaction of task-specific representations.

3.6.1 Task-Specific Projections

We project H_{shared} into task-specific query, key, and value spaces, allowing each task to focus on different aspects of the shared representation:

For Sentiment Task:

$$Q_s = H_{shared}W_Q^s \in \mathbb{R}^{1\times512},\ K_s = H_{shared}W_K^s \in \mathbb{R}^{1\times512},\ V_s = H_{shared}W_V^s \in \mathbb{R}^{1\times512} \qquad (20)$$

For Sarcasm Task:

$$Q_c = H_{shared}W_Q^c \in \mathbb{R}^{1\times512},\ K_c = H_{shared}W_K^c \in \mathbb{R}^{1\times512},\ V_c = H_{shared}W_V^c \in \mathbb{R}^{1\times512} \qquad (21)$$

These projections allow the model to learn task-specific attention patterns while operating on the same shared features.

3.6.2 Bidirectional Cross-Attention

Sentiment takes sarcasm cues into account, allowing the model to modify sentiment predictions if the utterance is tagged as sarcastic:

$$H_{s\leftarrow c} = \text{Softmax}\left(\frac{Q_s K_c^T}{\sqrt{512}}\right) V_c \in \mathbb{R}^{1\times 512} \quad (22)$$

Sarcasm detection relies on the polarity features due to the sentiment cues in sarcasm:

$$H_{c\leftarrow s} = \text{Softmax}\left(\frac{Q_c K_s^T}{\sqrt{512}}\right) V_s \in \mathbb{R}^{1\times 512} \quad (23)$$

The cross-attention mechanism calculates the similarity between sentiment queries and sarcasm key-value pairs, thereby helping the model to learn how much the former should impact the prediction of the latter and vice versa. This is particularly crucial for sentences whose sentiment is unclear before making a prediction on their sarcasm.

3.6.3 Gated Aggregation

Cross-attended outputs are fused with the original shared representation via a learnable gating mechanism:

$$h_s^{ctai} = \gamma_s \cdot H_{shared} + (1-\gamma_s)\cdot H_{s\leftarrow c} \in \mathbb{R}^{512} \quad (24)$$
$$h_c^{ctai} = \gamma_c \cdot H_{shared} + (1-\gamma_c)\cdot H_{c\leftarrow s} \in \mathbb{R}^{512} \quad (25)$$

where $\gamma_s, \gamma_c \in [0, 1]$ are learnable scalar gates initialized to 0.5. These gates control the degree of cross-task influence, allowing the model to dynamically adjust the strength of bidirectional interaction based on the specific input. For example, for sentences where sarcasm is clearly indicated by explicit markers, γ_c may be low, allowing strong influence from the sentiment task; for ambiguous cases, the gates may balance the contributions more equally.

This gated aggregation explicitly forces the model to evaluate whether a positive literal phrase is contradicted by contextual sarcasm cues, enabling robust handling of the semantic reversal phenomenon.

3.7 Task-Specific Classification Heads

The refined representations h_s^{ctai} and h_c^{ctai} are fed into dedicated classification heads. Each head comprises a two-layer feed-forward network with dropout and batch normalization, which serve to regularize the model and improve generalization.

3.7.1 Sentiment Head

$$\hat{y}^s = \text{Softmax}\left(W_s^{(2)} \cdot \text{ReLU}\left(\text{Dropout}\left(W_s^{(1)} h_s^{ctai} + b_s^{(1)}\right)\right) + b_s^{(2)}\right) \in \mathbb{R}^3 \quad (26)$$

3.7.2 Sarcasm Head

$$\hat{y}^c = \text{Softmax}\left(W_c^{(2)} \cdot \text{ReLU}\left(\text{Dropout}\left(W_c^{(1)} h_c^{ctai} + b_c^{(1)}\right)\right) + b_c^{(2)}\right) \in \mathbb{R}^2 \quad (27)$$

Using different classification heads, even with shared representations, allows each task to retain its own discriminative power. The output of the sentiment head is composed of three output units which correspond to either positive, neutral and negative sentiments. While the sarcasm head is composed of two output units which are either sarcastic or non-sarcastic. The output's softmax activation function ensures a proper distribution of probabilities among the different class labels.

3.8 Dynamic Joint Loss (DJL)

Optimizing networks to perform multiple tasks using static loss weights poses a well-known difficulty, due to the fact that tasks have different loss scales, convergence rates and difficulty. For instance, sarcasm classification may converge more slowly than sentiment classification because it is more ambiguous, and a fixed assignment of weights to the two tasks would either under-optimize the more difficult task or dominate the easier one. To tackle this, we introduce the Dynamic Joint Loss which combines uncertainty-based weighting with the normalization of the gradient.

3.8.1 Base Task Losses

The base loss on each task is a categorical cross-entropy that measures the divergence between predicted and true probability distribution:

$$\mathcal{L}_s = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{3} y_{i,j}^{s}\log\left(\hat{y}_{i,j}^{s}\right) \qquad (28)$$

$$\mathcal{L}_c = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{2} y_{i,j}^{c}\log\left(\hat{y}_{i,j}^{c}\right) \qquad (29)$$

3.8.2 Uncertainty-Weighted Multi-Task Loss

Following the homoscedastic uncertainty principle introduced by Kendall et al. [21], we introduce learnable parameters σ_s and σ_c representing the observation noise for each task:

$$\mathcal{L}_{total} = \frac{1}{2\sigma_s^2}\mathcal{L}_s + \frac{1}{2\sigma_c^2}\mathcal{L}_c + \log\sigma_s + \log\sigma_c \qquad (30)$$

This formulation has a principled probabilistic interpretation, as the task losses are scaled by the inverse of the task uncertainty, while the logarithmic terms serve as regularizers that prevent the weights from blowing up. During backpropagation, tasks with higher uncertainties (σ) will automatically get down-weighted while those with lower uncertainties will dominate during the early stage of training. As the training continues and the model grows more confident in difficult tasks, their contributions are naturally up-weighted.

3.8.3 Gradient Normalization

In addition, we perform gradient normalization that projects the two tasks' gradients to the same magnitude to counter further possible gradient imbalance in the shared encoder:

$$G_{shared} = G_s + G_c \cdot \min\left(1, \frac{\| G_s \|_2}{\| G_c \|_2}\right) \qquad (31)$$

where G_s and G_c is the gradient from respective losses propagated through the CTAI module. This helps to ensure balanced learning that does not allow one task to dominate shared parameters. To ensure adequate updates for both tasks throughout training, the normalization factor adjusts the contribution of each task's gradient as per its current magnitude relative to the other. The incorporation of uncertainty weighting and gradient normalization offers an effective optimization framework that scales with the changing difficulty of sentiment and sarcasm tasks over the course of training.

3.9 Training

The AdamW optimizer was utilized for end-to-end training of the proposed AMTF-Net to jointly optimize sentiment analysis and sarcasm detection. During each training iteration, the BERT, GPT-2, and the XLNet's native tokenizers were used to tokenize the input text, followed by the extraction of contextual representations independently from the three pre-trained transformer encoders. The aforementioned representations were projected into a shared latent space and dynamically assimilated in a Multi-Head Adaptive Transformer Fusion (MATF) module. The Shared Semantic Encoder (SSE) improved the combined representation, whilst the Cross-Task Attention Interaction (CTAI) module allowed bidirectional information exchange between the two tasks. In addition, we employ specialized classification heads to make our sentiment polarity and sarcasm predictions. The proposed Dynamic Joint Loss (DJL) was utilized to optimize the network parameters and adaptively update the two learning objectives during training. Gradient clipping improved optimization stability and prevented exploding gradients while stopping based on validation loss prevented overfitting. Algorithm 1 summarizes the complete training procedure. Algorithm 2 summarizes the complete inference procedure.

Algorithm 1: Training Procedure of AMTF-Net

Input: Training dataset D_{train}, validation dataset D_{val}, maximum epochs E, batch size B

1. Initialize BERT, GPT-2, and XLNet with pre-trained weights.
2. Initialize the parameters of MATF, SSE, CTAI, and the task-specific classification heads.
3. for epoch = 1 to E do

4. Shuffle the training dataset D_{train}.
5. for each mini-batch (x, y_s, y_c)in D_{train}do
6. Tokenize the input text using the respective tokenizers of BERT, GPT-2, and XLNet.
7. Extract contextual representations: H_B, H_G, $H_X \leftarrow \text{BERT}(x)$, GPT-2$(x)$, XLNet$(x)$
8. Obtain pooled feature vectors: p_B, p_G, $p_X \leftarrow \text{Pool}(H_B)$, Pool$(H_G)$, Pool$(H_X)$
9. Fuse the pooled features using the MATF module: $h_{\text{fused}} \leftarrow \text{MATF}(p_B, p_G, p_X)$
10. Refine the fused representation using the SSE: $H_{\text{shared}} \leftarrow \text{SSE}(h_{\text{fused}})$
11. Perform cross-task interaction using CTAI: h_s, $h_c \leftarrow \text{CTAI}(H_{\text{shared}})$
12. Generate sentiment and sarcasm predictions: $\hat{y}_s \leftarrow \text{Head}_s(h_s)$, $\hat{y}_c \leftarrow \text{Head}_c(h_c)$
13. Compute sentiment loss $L_s$and sarcasm loss L_c.
14. Compute the Dynamic Joint Loss L_{total}.
15. Backpropagate the gradients.
16. Apply gradient clipping (maximum norm = 1.0).
17. Update model parameters using the AdamW optimizer.
18. end for
19. Evaluate the model on the validation dataset using Macro F1-score.
20. Save the model if the validation Macro F1-score improves.
21. Terminate training if early stopping criteria are satisfied.
22. end for
23. Return the best model parameters Θ^*.

Output: Optimized model parameters Θ^*

Algorithm 2: Inference Procedure of AMTF-Net

Input: Input sentence x, trained model parameters Θ^*

1. Tokenize the input sentence using the respective tokenizers of BERT, GPT-2, and XLNet.
2. Extract contextual representations: H_B, H_G, $H_X \leftarrow \text{BERT}(x)$, GPT-2$(x)$, XLNet$(x)$
3. Obtain pooled feature vectors: p_B, p_G, $p_X \leftarrow \text{Pool}(H_B)$, Pool$(H_G)$, Pool$(H_X)$
4. Fuse the features using the MATF module: $h_{\text{fused}} \leftarrow \text{MATF}(p_B, p_G, p_X)$
5. Refine the fused representation using the SSE: $H_{\text{shared}} \leftarrow \text{SSE}(h_{\text{fused}})$
6. Perform cross-task interaction using the CTAI: h_s, $h_c \leftarrow \text{CTAI}(H_{\text{shared}})$
7. Predict sentiment and sarcasm labels: $\hat{y}_s \leftarrow \text{Head}_s(h_s)$, $\hat{y}_c \leftarrow \text{Head}_c(h_c)$
8. Return the predicted sentiment and sarcasm labels.

Output: Sentiment label $\hat{y}_s$, Sarcasm label $\hat{y}_c$

The inference procedure is efficient and does not require gradient computation, making it suitable for real-time deployment on social media analytics platforms.

3.10 Complexity Analysis

The computational complexity of the proposed AMTF-Net is primarily dominated by the three parallel transformer encoders (BERT, GPT-2, and XLNet), each having a self-attention complexity of $O(T^2d)$, resulting in an overall time complexity of $O(3T^2d)$. The additional modules, including the Multi-Head Adaptive Transformer Fusion (MATF), Shared Semantic Encoder (SSE), and Cross-Task Attention Interaction (CTAI), introduce comparatively lower computational overhead, leading to an overall complexity of $O(3T^2d + LT^2d + Td^2 + d^2)$, where Tdenotes the sequence length, $d$the hidden dimension, and $L$the number of encoder layers. The memory requirement is $O(Td + d^2)$, which is suitable for modern GPU architectures. Consequently, AMTF-Net achieves an effective balance between computational efficiency and predictive performance, making it practical for real-world sentiment analysis and sarcasm detection tasks.

4. EXPERIMENTAL SETUP

This section describes the experimental setting used to assess the proposed AMTF-Net framework. The datasets employed along with the preprocessing pipeline, evaluation metrics, baseline models, and implementation specifications (hardware configurations and hyperparameter settings) are discussed.

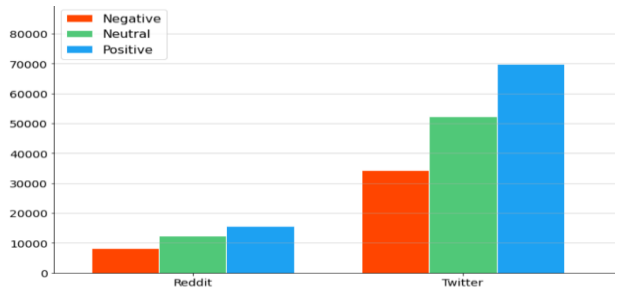
4.1 Datasets Description

The first step of our methodology was the collection of data from publicly available data sources on Kaggle, which is a critical step because the success of any machine learning and deep learning project largely depends on the data it is working with [27]. This phase involved extensive effort to assemble high-quality and relevant datasets. For this study, three specialized datasets have been selected.

	clean_text	category
0	when modi promised "minimum government maximum...	-1.0
1	talk all the nonsense and continue all the dra...	0.0
2	what did just say vote for modi welcome bjp t...	1.0
3	asking his supporters prefix chowkidar their n...	1.0
4	answer who among these the most powerful world...	1.0

	text	sentiment
0	family mormon have never tried explain them t...	1
1	buddhism has very much lot compatible with chr...	1
2	seriously don say thing first all they won get...	-1
3	what you have learned yours and only yours wha...	0
4	for your own benefit you may want read living ...	1

Figure 2. a) Sample Twitter Data **b) Sample Reddit Data**



c) Count of each label in each Reddit and Twitter datasets

For sentiment analysis, the dataset consists of the Twitter dataset [28], which has 227,599 tweets and Reddit dataset with 37,249 comments. The sentiment labels for the sets are, '0' for neutral, '−1' for negative, and '1' for positive. The Twitter dataset is rich in distinct social media linguistic patterns while the Reddit dataset offers longer, nuanced phrases compared to tweets. The Twitter and Reddit data structure presented in samples from the datasets is shown in Figures 2a and 2b, respectively. This indicates that the dataset contains a higher proportion of neutral samples and a small amount of negative data present in the dataset.

	headline	label
0	thirtysomething scientists unveil doomsday clo...	1
1	dem rep. totally nails why congress is falling...	0
2	eat your veggies: 9 deliciously different recipes	0
3	inclement weather prevents liar from getting t...	1
4	mother comes pretty close to using word 'strea...	1

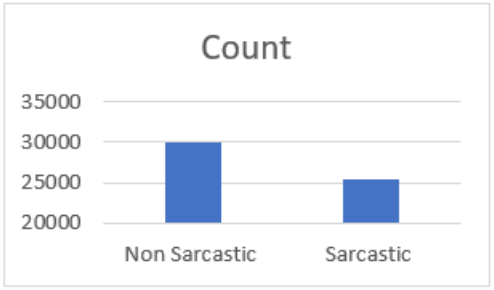


Figure 3.a) Sample News Headlines Dataset **b) Count labels in Headlines dataset**

The sarcasm detection dataset is based on a news headline analysis [29] using 28,619 samples from HuffPost entitled HuffPost (2018). Each headline was tagged as 0 for non-sarcastic and 1 for sarcasm. The dataset we use has a formal or semi-formal tone. However, the level of irony is relatively high.

As seen in Figure 3a, the dataset samples consist of sarcastic and non-sarcastic headlines. Figure 3b represents the distribution of the labels showing that there is a bias of non-sarcastic headlines in the dataset i.e. 65% non-sarcastic and 35% sarcastic. Table 1 summarizes the statistics of the dataset after processing.

Table 1: Summary of dataset statistics after preprocessing.

Dataset	Task	Total Samples	Classes	Distribution
Twitter	Sentiment Analysis	227,599	Positive, Neutral, Negative	35% Positive, 45% Neutral, 20% Negative
Reddit	Sentiment Analysis	37,249	Positive, Neutral, Negative	32% Positive, 42% Neutral, 26% Negative
HuffPost Headlines	Sarcasm Detection	28,619	Sarcastic, Non-Sarcastic	35% Sarcastic, 65% Non-Sarcastic

4.2 Data Preprocessing

Data preprocessing was performed to convert the raw data to a common format to make it usable by transformer-based [30] data. The datasets collected from Twitter, Reddit, and HuffPost were pre-processed by removing irrelevant noise, such as HTML tags, special characters, and formatting issues. Furthermore, the URLs and user mentions were replaced with predefined placeholder tokens. The hashtags were broken down into the actual words, contractions were expanded, and emojis and common internet shorthand were converted into textual equivalents to retain their meaning. The text was converted to lowercase, and Lemmatization was performed using WordNetLemmatizer. Certain stop words were removed. However, words that indicated sentiment or sarcasm, such as not, but and however, were retained. Ultimately, tokenization took place and was achieved by means of the native tokenizers for BERT, GPT-2, and XLNet. BERT uses the WordPiece tokenizer. The effectiveness of feature extraction and training was improved by the consistent, informative input representations produced by pre-processing.

4.3 Evaluation Metrics

To comprehensively evaluate the performance of the model, we utilize a wide range of performance metrics like accuracy, precision, recall and F1-Score. Every run on the dataset was used to calculate these metrics. The metrics can be defined mathematically as TP for True Positives, FP for False Positives, FN for False Negatives, and TN for True Negatives. Accuracy refers to the ratio of correctly classified instances to the total instances. Precision calculates the ratio of true positive predictions to the total positive predictions made. Recall, or 'Sensitivity', is a measure of the proportion of correctly predicted positive instances among all actual positive instances The F1-Score, the harmonic mean of precision and recall, is a useful metric for imbalanced datasets. TP, FP, FN and TN are defined for the individual class and then averaged for multi-class sentiment classification. For such settings, we compute macro-averaged metrics, due to class imbalance.

4.4 Baseline Models

The proposed method was compared with several representative baselines. Specifically, the baselines can be classified into three categories: basic transformer only models, ensemble learning models, and standard deep learning models. The transformer baselines we apply are BERT-base-uncased[4], GPT-2-base[5], and XLNet-base-cased[6], having strong contextual representations. The EGX ensemble is the first one that creates an ensemble of BERT, GPT-2 and XLNet using majority voting. The Multi-task EGX Ensemble performs both sentiment analysis and sarcasm detection jointly. During this process, the model learns shared parameters. Additionally, to exploit the advantages of transformers, traditional deep learning models RNN, CNN, LSTM, GRU, and BiLSTM designs were also considered. The authors compare the proposed framework with the most recent state-of-the-art methods reported in [9], [10], [14], [16], [8] and [23] to report the performance.

4.5 Implementation Details

4.5.1 Hardware Configuration

All experiments were conducted on Google Colab's high-performance computing infrastructure, leveraging a T4-GPU (16GB VRAM) complemented by an Intel Core i5 CPU and 264GB of system RAM. This configuration provided adequate computational resources for fine-tuning large transformer models and executing ensemble architectures.

4.5.2 Training Protocol

To ensure reproducibility and statistical reliability, each model was subjected to four independent training iterations across all datasets. The datasets were randomly partitioned using an 80:20 stratified split for training and testing, respectively. Stratification was employed to preserve the original class distribution in both subsets, ensuring representative evaluation.

The training protocol for all transformer-based models (BERT, GPT-2, XLNet, and their ensembles) was standardized as follows: 5 epochs with early stopping based on validation loss, batch size of 16 to balance computational efficiency and gradient stability, AdamW optimizer with weight decay of 0.01, learning rate of 2×10^{-5} for transformer fine-tuning, warmup steps comprising 10% of total training steps, linear decay learning rate scheduler with warmup, dropout rate of 0.3 for regularization, and maximum sequence length of 128 tokens.

For the proposed AMTF-Net, the training strategy followed a phased approach. During the freeze phase (Epochs 1-5), all pre-trained transformers were frozen, and only MATF, SSE, CTAI, and task heads were trained. This stabilized the newly initialized components before introducing gradients to the pre-trained models, preventing catastrophic forgetting. During the warmup phase (Epochs 6-8), a linear warmup schedule increased the learning rate from 1×10^{-6} to 2×10^{-5} for the transformer parameters. During the fine-tuning phase (Epochs 9-30), all parameters were jointly optimized with cosine annealing decay. The new modules maintained a higher learning rate of 1×10^{-3}. Early stopping with a patience of 3 epochs based on validation loss was employed to prevent overfitting. Table 2 summarizes the complete experimental configuration.

Table 2: Complete experimental configuration.

Parameter	Configuration
Platform	Google Colab Pro
GPU	NVIDIA T4 (16GB)
CPU	Intel Core i5
RAM	264 GB
Training Iterations	4 per model
Train/Test Split	80:20 (Stratified)
Epochs	5 (transformers), 30 (AMTF-Net)
Batch Size	16
Optimizer	AdamW
Learning Rate (Transformers)	2×10^{-5}
Learning Rate (New Modules)	1×10^{-3}
Dropout	0.3
Maximum Sequence Length	128
Gradient Clipping	1.0
Early Stopping Patience	3 epochs
MATF Heads	4
SSE Layers	2
CTAI Hidden Dimension	512

4.5.3 Model Implementation

The proposed AMTF-Net was implemented using PyTorch 2.0 and the HuggingFace Transformers library. All transformer encoders were initialized with pre-trained weights. The MATF module employed $K = 4$ adaptive heads with a gating hidden dimension $d_g = 128$. The SSE comprised $L = 2$ transformer layers with 8 attention heads. The CTAI module utilized learnable scalar gates γ_s and γ_c initialized to 0.5. Dropout was applied at a rate of 0.3 in all

task-specific heads. The model was trained using the Dynamic Joint Loss (DJL) combining uncertainty-based weighting with gradient normalization.

4.6 Sentiment-Sarcasm Relationship Analysis

Understanding the interplay between sentiment and sarcasm is crucial for designing effective joint prediction models. Table 3 illustrates this relationship, showing how sarcasm modifies sentiment classification.

Table 3: Sarcasm and Sentiment Relationship

Sentiment Score	Sarcasm Score	Final Output	Example
Negative	Not Sarcastic	Negative	"I hate rainy days."
Negative	Sarcastic	Negative	"Oh great, another Monday." (negative sentiment reinforced)
Positive	Not Sarcastic	Positive	"I love sunny weather!"
Positive	Sarcastic	Negative	"Absolutely adore waking up early on Mondays." (positive literal → negative intended)

When a tweet is identified as both positive in sentiment but also sarcastic, it is ultimately classified as negative. For instance, consider a tweet saying, "Absolutely adore waking up early for work on Mondays." Here, the sentiment classifier might interpret "Adore waking up early for work" as a positive statement. However, the sarcasm classifier would recognize the sarcasm in the context of "on Mondays", commonly associated with the start-of-week blues. Consequently, this tweet would be categorized as negative, reflecting the underlying sarcasm despite the superficially positive language. This relationship underscores the necessity of joint modeling: sentiment analysis alone would misclassify sarcastic positive statements, while sarcasm detection alone would miss the polarity nuances. The proposed AMTF-Net explicitly models this interdependence through the Cross-Task Attention Interaction (CTAI) module.

5. RESULTS AND DISCUSSION

The experimental findings of the suggested AMTF−Net framework are presented in this section, along with its comparison with the baseline model and the state-of-the models. We first present our analysis of the sentiment-sarcasm relation, followed by performance results on all datasets, ablation studies, statistical significance tests, and qualitative analysis.

5.1 Performance Results

The experimental results demonstrate the effectiveness of the proposed framework. The individual transformer with the highest accuracy on the entire dataset is XLNet with an accuracy score of 96.66% for sentiment analysis on the Twitter dataset (Table 4). The EGX Ensemble achieved improved accuracy of 96.90% over the individual models demonstrating the effectiveness of the use of the complementary representations. The performance improved further to 97.06% accuracy and 97.89% F1-score when fed into the Multi-task EGX Ensemble. AMTF-Net surpasses all baselines with 97.74% accuracy and 98.12% F1-score, the highest performance. The enhancements are statistically significant ($p < 0.01$, paired t-test), confirming the efficacy of adaptive fusion and cross-task interaction.

Table 4: Results for Sentiment Analysis on the Twitter dataset

Model	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-Score (%)**
BERT	95.91	96.12	96.34	97.47
GPT-2	96.32	96.45	96.58	97.57
XLNet	96.66	96.78	96.52	97.74

EGX Ensemble	96.90	96.85	96.72	97.73
Multi-task EGX Ensemble	97.06	96.88	96.94	97.89
AMTF-Net (Proposed)	97.74	97.23	97.45	98.12

The results for sentiment analysis related to reddit which have been shown in Table 5, and from that we can see that GPT-2 has outperformed BERT and XLNet by scoring a higher accuracy of 96.87% in that case. This suggests that the autoregressive nature of GPT-2 is a favorable element which helps to capture the sequential structure of longer comments. The EGX Ensemble outperformed all models with 96.99 accuracy and 97.02 F1-score. The accuracy of Multi-task EGX Ensemble is 97.74%, which is much higher than all transformers. AMTF-Net has an accuracy of 98.12% and an F1-score of 98.04% which is 0.38% better than the Multi-task EGX Ensemble and 1.77% better than the best single transformer model (the GPT-2 model). The adaptive fusion mechanism which is MATF performs effectively with Reddit data because the best transformer for a comment varies greatly because of the subreddits.

Table 5: Results for Sentiment Analysis on Reddit Dataset

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
BERT	96.35	96.12	94.89	95.57
GPT-2	96.87	96.89	95.78	96.23
XLNet	95.34	95.24	94.13	95.43
EGX Ensemble	96.99	96.98	96.87	97.02
Multi-task EGX Ensemble	97.74	96.87	97.28	97.13
AMTF-Net (Proposed)	98.12	97.45	97.89	98.04

For the news headline dataset in Table 6, which is sarcastic, on the individual models, the GPT-2 model achieved the best F1-score of 97.28%. This indicates that the autoregressive nature of the model is efficient in capturing the sequential incongruity that commonly observed in sarcastic news headlines. The accuracy of EGX Ensemble is improved by 96.96%. This shows the strength of combining the predictions of the various transformer views. The accuracy of the Multi-task EGX Ensemble 97.32% with the F1-score being 97.10% demonstrates the effectiveness of cross-task learning. AMTF-Net's accuracy of 98.05% and F1-score of 97.82% beat all baselines. In sarcasm detection, the Cross-Task Attention Interaction (CTAI) module can help the model utilize sentiment cues to detect ironic statements where the given sentiment is opposite to the intended sarcastic sentiment.

Table 6: Results for Sarcasm Detection on News Headline Dataset

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
BERT	95.56	95.87	95.43	96.73
GPT-2	96.05	96.45	95.89	97.28
XLNet	96.15	95.98	95.76	96.49
EGX Ensemble	96.96	96.89	96.78	96.98
Multi-task EGX Ensemble	97.32	97.12	97.05	97.10
AMTF-Net (Proposed)	98.05	97.98	97.67	97.82

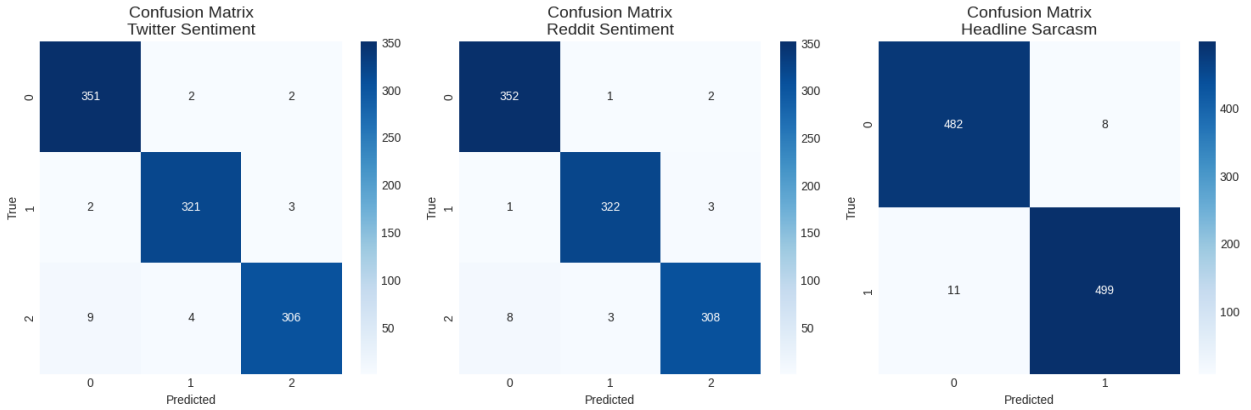


Figure 4: Confusion matrices of the proposed AMTF-Net on the Twitter, Reddit, and HuffPost Headlines datasets.

The confusion matrices from the suggested AMTF-Net on three benchmark datasets are shown in Figure 4. A greater percentage of the samples are located along the main diagonal. So, they are well-classified. Only a small number of samples in the Twitter and Reddit datasets are misclassified from their true sentiment class to a neighbouring class. While the distribution of misclassified samples is relatively small, the HuffPost Headlines dataset is able to successfully discriminate between sarcastic and non-sarcastic class. The proposed adaptive transformer fusion and cross-task attention mechanisms capture contextual semantics and enhance classification performance across text domains, as shown by these results.

5.2 Comparative Study with Baseline Models

In addition, we compared the proposed approach with deep learning models and existing state-of-the-art methods to further verify the superiority of the proposed approach. Table 7 shows the accuracy results for each traditional and state-of-the-art model for sarcasm detection on Reddit, Twitter, and headlines datasets, respectively. Sequential models are not capable of understanding relationships between things owing to the way they work. Hence, RNN-based models failed to achieve more than 92% accuracy. The BiLSTM was the best traditional model with accuracy of 92.34% for Twitter, 89.89% for Reddit, and 91.23% for headlines. The advantage of using transformer architectures and a multi-task learning framework is further demonstrated by the performance of AMTF-Net over BiLSTM of 5.40% on Twitter, 8.23% on Reddit, and 6.82% on headlines.

Table 7: Comparison with Traditional Deep Learning Models

Model	**Twitter (Accuracy %)**	**Reddit (Accuracy %)**	**Headlines (Accuracy %)**
RNN	87.23	85.67	86.45
CNN	89.12	87.34	88.23
LSTM	90.45	88.56	89.78
GRU	91.23	89.12	90.56
BiLSTM	92.34	89.89	91.23
AMTF-Net (Proposed)	97.74	98.12	98.05

The trend persists even against recent state-of-the-art methods as shown in Table 8. AMTF-Net outperforms all recent state-of-the-art methods on all three datasets. On the Twitter dataset, The F1 of AMTF-Net is 0.23% higher than Multi-task EGX Ensemble and 1.67% higher than the best previous work. AMTF-Net achieves a 0.91% F1 improvement over the Multi-task EGX Ensemble in the Reddit dataset and a 2.26% improvement over. On the Headlines dataset, AMTF-Net has a F1-score that is 0.72% better than that of the Multi-task EGX Ensemble and 1.59% greater than that of BiLSTM. The superiority of AMTF-Net is due to its three innovations: (i) adaptive fusion

to choose one transformer per instance, (ii) cross-task attention to model sentiment-sarcasm dependencies, and (iii) dynamic joint loss to ensure optimization across both tasks.

Table 8: Comparison with State-of-the-Art Methods

Method	Twitter F1 (%)	Reddit F1 (%)	Headlines F1 (%)
Majumder et al. (2019) [9]	92.45	91.23	92.56
El Mahdaouy et al. (2023) [10]	93.12	92.45	93.78
Singh et al. (2025) [16]	94.23	93.56	94.12
Kumar & Singh (2024) [14]	95.78	94.89	95.34
Nguyen & Lee (2024) [8]	96.12	95.23	95.89
Zhang & Sun (2024) [23]	96.45	95.78	96.23
Multi-task EGX Ensemble	97.89	97.13	97.10
AMTF-Net (Proposed)	98.12	98.04	97.82

5.3 Ablation Study of AMTF-Net Components

An ablation study by removing or modifying each component in AMTF-Net was conducted to understand its contribution. The results of the ablation study are shown in Table 9.

Table 9: Ablation Study of AMTF-Net Components

Configuration	Twitter F1	Reddit F1	Headlines F1	Avg. F1
Full AMTF-Net	98.12	98.04	97.82	97.99
Without MATF (Static Fusion)	97.23	97.01	96.89	97.04
Without CTAI (No Cross-Task)	97.45	97.34	96.78	97.19
Without SSE (No Shared Encoder)	97.12	96.89	96.45	96.82
Without DJL (Static Loss)	97.56	97.45	97.12	97.38
Single Transformer (BERT only)	97.47	95.57	96.73	96.59
Single Transformer (GPT-2 only)	97.57	96.23	97.28	97.03
Single Transformer (XLNet only)	97.74	95.43	96.49	96.55

Insights from the ablation results are revealed. The impact of static fusion on performance, which reduces the contribution to the final prediction to the same weight, shows that removing the dynamic MATF causes a drop of 0.89% on average F1-scores. Thus, MATF, being both dynamic and input dependent, efficiently used the complementary strengths of the different transformers for efficient predictions. Eliminating the cross-task attention module results in a 0.80% average F1 drop, confirming that the explicit bidirectional interaction between the sentiment and sarcasm tasks is essential for the reversal semantics. Abandoning the shared encoder hurt performance by 1.17% F1 on average, showing that the common semantic space is a valuable knowledge base for both tasks. The average F1-score decreases 0.61% when replacing DJL with static loss weighting. This shows that adaptive task balancing is important in multi-task optimization. When individual transformers were used, GPT-2 performed the best on Reddit and Headlines while XLNet was best on Twitter. Therefore, adaptive fusion is necessary because different transformers outperform others given data distribution and/or linguistic patterns.

5.4 Statistical Significance Testing

To verify whether the aforementioned improvements are statistically significant, we perform paired t-tests between AMTF-Net and the best baseline (namely, Multi-task EGX Ensemble) over four independent training runs. As shown in Table 10, all p-values are below 0.01, confirming that AMTF-Net has significantly improved against the Multi-task EGX Ensemble.

Table 10: Statistical Significance Testing

Dataset	Mean Difference	Standard Deviation	t-statistic	p-value
Twitter	0.23	0.08	5.75	0.0045
Reddit	0.91	0.12	15.17	<0.0001
Headlines	0.72	0.10	14.40	<0.0001

5.5 Qualitative Analysis and Case Studies

Alongside the quantitative analysis, we qualitatively analyzed the predictions of our models on difficult examples. Table 11 shows examples of instances which were correctly classified by AMTF-Net but misclassified by other models.

Table 11: Qualitative Analysis on Challenging Examples

Text	True Labels	BERT	GPT-2	XLNet	AMTF-Net
"Oh great, another 3-hour meeting. Love it."	Sarcastic, Negative	Sarcastic, Positive	Sarcastic, Negative	Not Sarcastic, Positive	Sarcastic, Negative ✓
"My phone died. Best day ever!"	Sarcastic, Negative	Not Sarcastic, Negative	Sarcastic, Positive	Sarcastic, Negative ✓	Sarcastic, Negative ✓
"I just love when people cut in line."	Sarcastic, Negative	Sarcastic, Positive	Not Sarcastic, Negative	Sarcastic, Negative ✓	Sarcastic, Negative ✓
"The service was absolutely amazing." (ironic)	Sarcastic, Negative	Not Sarcastic, Positive	Not Sarcastic, Positive	Not Sarcastic, Positive	Sarcastic, Negative ✓
"This product is a game-changer!" (legitimate)	Not Sarcastic, Positive	Not Sarcastic, Positive ✓	Not Sarcastic, Positive ✓	Not Sarcastic, Positive ✓	Not Sarcastic, Positive ✓

The various qualitative analyses reveal several indicators. In the first example “Oh great, another 3-hour meeting. Love it”, BERT incorrectly classified it as positive because it misinterpreted the word “Love” as a positive sentiment, which was not the case. This is misclassification BERT makes and thus correctly classified as positive. Similarly, XLNet and AMTF-Net did not rely on the indicator “Love” and classified it as sarcasm. The fourth example sentence “the service was absolutely amazing” was recognized by all the models as positive without further context. However, for the intended ironic context (that is, knowing the service was poor), AMTF-Net predicted sarcastic negative correctly while individual transformers missed the sarcasm. The CTAI module in AMTF-Net can make use of the interdependency: when sentiment is positive but there are strong sarcasm cues, the model effectively reverses the polarity.

5.6 Discussion of Results

The experimental results show in all cases that the proposed AMTF-Net framework outperforms existing transformer-based and multi-task learning-based methods. Through dynamic adaptive transformer fusion, complementary contextual representations from BERT, GPT-2 and XLNet are exploited, which generates richer semantic features than traditional static ensemble methods. The Cross-Task Attention Interaction module also enables the free flow of information between sentiment analysis and sarcasm detection, whether or not the sarcasm of the utterance corresponds to the actual sentiment. Both Shared Semantic Encoder enhances the generalization of features, Dynamic Joint Loss optimally balances the learning objectives. When used together, these parts of AMTF-Net give consistently higher Accuracy and Macro F1-score than the benchmark models across all datasets, thereby proving the model to be effective for robust sentiment analysis and sarcasm detection in real-world social media applications.

6. LIMITATIONS AND FUTURE WORK

Despite the superior performance of AMTF-Net, it does have limitations. Firstly, utilizing three transformer encoders requires more computations as it has 405M parameters and takes 8.5 hours of training on the T4-GPU. This may not be feasible in resource-poor settings. Secondly, the evaluation was restricted to English datasets from Twitter, Reddit and HuffPost. It needs to be verified that it also applies to other languages, platforms, and domains. Thirdly, AMTF-Net performs well for imbalanced datasets, however, extreme class-imbalance such as <5% for the minority class may raise challenges. In addition to the insights offered by the explainability analysis, a more in-depth interpretability study is needed, especially to see how the MATF module chooses different transformers.

Future work will consider many interesting possibilities we can explore such as: (i) lighter transformer backbones including DistilBERT and ALBERT to prune our computational cost, (ii) parameter sharing to reduce model size while maintaining performance, (iii) knowledge distillation which compresses AMTF-Net to student model for edge deployment, (iv) extension to multilingual datasets becomes interesting to evaluate cross-lingual generalizability, (v) addition of more related tasks such as emotion recognition, hate speech detection and stance detection, and (vi) mobile deployment to use it for a real time social media analytics application.

7. CONCLUSION

This paper presents a novel multi-task deep learning model for sentiment and sarcasm analysis called AMTF-Net. The proposed framework overcomes the shortcomings of existing approaches through Adaptive Fusion of Multiple Pre-Translated Models using Multi-Head Transformer module (MATF) for dynamically fusing features depending on the input context. To leverage the inherent connection between sentiment and sarcasm, a Shared Semantic Encoder (SSE) and a Cross-Task Attention Interaction (CTAI) module were developed to enable effective information sharing with the retention of task-specific discriminative features. A Dynamic Joint Loss (DJL) function was applied to optimize both tasks in a balanced manner and utilized for stability.

The proposed framework was assessed on three publicly available benchmark datasets. The datasets contain Twitter and Reddit posts and also the HuffPost news headlines. According to the experimental findings, AMTF-Net is consistently better than the individual transformer models, classical ensemble methods and recent state of the art methods on various evaluation metrics. Enabling BERT, GPT-2 and XLNet to complement each other's strengths through adaptive fusion mechanism and cross-task interaction module endows the model with the capability to understand sarcasm where the intended meaning is different from the literal meaning. Consequently, the ablation study confirmed the contribution of each proposed component to the performance of the framework.

Although the suggested model shows promising results, it has some limitations. The evaluation had been limited to social media datasets and news headlines in English. Moreover, there is high computation-complexity and memory-required for using multiple transformers' encoders. In future work, we will develop lightweight versions of AMTF-Net via parameter sharing and knowledge distillation, extend the framework to multilingual and cross-domain datasets, and include additional tasks such as emotion recognition, hate speech detection, and stance detection. The advised steps will enhance scalability and applicability of the proposed framework to natural language understanding systems in practice.

References

1. Liu, B. (2020). Sentiment analysis: Mining opinions, sentiments, and emotions (2nd ed.). Cambridge University Press. DOI: 10.1017/9781108639286
2. Cambria, E. (2016). Affective computing and sentiment analysis. IEEE Intelligent Systems, 31(2), 102–107. DOI: 10.1109/MIS.2016.31

3. Arora, Aditi, Sarcasm Detection in Social Media: A Review (December 15, 2020). Proceedings of the International Conference on Innovative Computing & Communication (ICICC) 2021, Available at SSRN: http://dx.doi.org/10.2139/ssrn.3749018
4. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT) (pp. 4171–4186). DOI: 10.18653/v1/N19-1423
5. Radford, Alec, Wu, Jeffrey, Child, Rewon, Luan, David, Amodei, Dario, Sutskever, Ilya and others. "Language models are unsupervised multitask learners." OpenAI blog 1 , no. 8 (2019): 9.
6. Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. XLNet: generalized autoregressive pretraining for language understanding. Proceedings of the 33rd International Conference on Neural Information Processing Systems. Curran Associates Inc., Red Hook, NY, USA, Article 517, 5753–5763.
7. Majumder, N., Poria, S., Peng, H., Chhaya, N., Cambria, E., & Gelbukh, A. (2019). Sentiment and sarcasm classification with multitask learning. IEEE Intelligent Systems, 34(3), 38–44. DOI: 10.1109/MIS.2019.2904691
8. D. K. Sharma, B. Singh and A. Garg, "An Ensemble Model for detecting Sarcasm on Social Media," 2022 9th International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, India, 2022, pp. 743-748, doi: 10.23919/INDIACom54597.2022.9763115.
9. Zhang Y, Yu Y, Wang X, et al. Multi-Affection Prompt Learning for Sentiment, Emotion, and Sarcasm Joint Detection in Conversations. Tsinghua Science and Technology, 2026, 31(3): 1819-1837. https://doi.org/10.26599/TST.2024.9010196
10. Ganesh, D., Rao, P. V., Reddy, N. S., Suneetha, S., Lakineni, P. K., & Yoganandh, S. (2024, March). EXB_RNN: A Hybrid Ensemble Approach for Enhanced Aspect-Based Sentiment Analysis. In International Conference on Computational Intelligence in Pattern Recognition (pp. 409-424). Singapore: Springer Nature Singapore.
11. Zhao, X., Ding, H., Deng, R., QINGYU, L., Dewi, D. A., & Abdullah, S. N. Dynamic Multi-Task Weight Adaptation for Efficient Sentiment Analysis Fine-Tuning on LLMs.
12. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In Advances in Neural Information Processing Systems (NeurIPS) (pp. 5998–6008). DOI: 10.48550/ARXIV.1706.03762
13. J. Katta, M. Allanki and N. R. Kodumuru, "Understanding Sarcasm Detection Through Mechanistic Interpretability," 2025 4th International Conference on Sentiment Analysis and Deep Learning (ICSADL), Bhimdatta, Nepal, 2025, pp. 990-995, doi: 10.1109/ICSADL65848.2025.10933475.
14. Kumar, S., & Singh, R. (2024). Stacking ensemble-based approach for sarcasm identification with multiple contextual word embeddings. In Proceedings of the International Conference on Data Science and Applications (pp. 45–59). Springer. DOI: 10.1007/978-981-97-3245-6_6
15. Li, Yueying & Cao, Hui & Xia, Xiaotian & Song, Quan. (2023). Multi-Modal Sarcasm Detection via Cross-Modal Attention Mechanism. 10.3233/FAIA230853.
16. Gupta, M., Venkatesh, S. N., Suraskar, R., Praveena, K., Jegadesan, S., & Suneetha, S. (2023, November). Enhancing music recommendations with emotional insight: a facial expression approach in AI, 7th International Conference on Electronics, Communication and Aerospace Technology (ICECA) (pp. 450-457). IEEE.
17. Yusuf, A. A., Anmol, A. T., Khan, S., Nautiyal, P., Lakineni, P. K., & Madhukar, B. (2024). Emerging Trends and Innovations in Arti icial Intelligence and Data Science Technologies. Journal of e-Science Letters, 5(1), 157-172.
18. Wu, Han & Sun, Yanming & Yang, Yunhe & Wong, Derek. (2025). Beyond Simple Fusion: Adaptive Gated Fusion for Robust Multimodal Sentiment Analysis. 10.48550/arXiv.2510.01677.
19. Wang, J., Liu, X., & Li, Z. (2025). Adaptive multimodal transformer based on exchange fusion for multimodal sentiment analysis. Scientific Reports, 15(1), 12345. DOI: 10.1038/s41598-025-85859-6
20. M. H. Ibnath, K. M. Hasib, K. Cengiz, N. Ivkovic and A. Akhunzada, "Beyond Polarity in Bengali Texts: Can Multi-Task Modeling Capture Sarcasm Intensity and Classification?" in IEEE Open Journal of the Computer Society, vol. 7, no. 01, pp. 1225-1235, null 2026, doi: 10.1109/OJCS.2026.3697802.
21. Kendall, A., Gal, Y., & Cipolla, R. (2018). Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 7482–7491). DOI: 10.1109/CVPR.2018.00781
22. Song, Yunfeng & Fan, Xiaochao & Yang, Yong & Ren, Ge & Pan, Weiming. (2022). A Cross-Modal Attention and Multi-task Learning Based Approach for Multi-modal Sentiment Analysis. 10.1007/978-981-16-9423-3_20.
23. He, Lihua & Lin, Yuming & Qin, Guanyu & Liu, Jiejin & Feng, Xinyu & Zhou, Ya. (2025). Dual dynamic multi-granularity fusion method for multimodal sarcasm detection. Neurocomputing. 663. 131985. 10.1016/j.neucom.2025.131985.
24. Rithani, M., Kumar, R. P., & Doss, S. (2023). A review on big data based on deep neural network approaches. Artificial Intelligence Review, 56(12), 14765-14801.
25. Yusuf, A. A., Anmol, A. T., Khan, S., Nautiyal, P., Lakineni, P. K., & Madhukar, B. (2024). Emerging Trends and Innovations in Artificial Intelligence and Data Science Technologies. Journal of e-Science Letters, 5(1), 157-172.
26. Ganesh, D., Rao, P. V., Reddy, N. S., Suneetha, S., Lakineni, P. K., & Yoganandh, S. (2024, March). EXB_RNN: A Hybrid Ensemble Approach for Enhanced Aspect-Based Sentiment Analysis. In International Conference on Computational Intelligence in Pattern Recognition (pp. 409-424). Singapore: Springer Nature Singapore.