Multi-Modal Data-Driven Sentiment Analysis in Online Public Opinion During Public Health Emergencies

Jinli Duan *, Zhibin Lin, Feng Jiao

Posted Date: 4 June 2026

doi: 10.20944/preprints202606.0301.v1

Keywords: public health emergency; sentiment analysis; online feature selection; Markov blanket; cross-modal attention

 doi:10.20944/preprints202606.0301.v1

Article

Multi-Modal Data-Driven Sentiment Analysis in Online Public Opinion During Public Health Emergencies

Jinli Duan [1,*], Zhibin Lin [2] and Feng Jiao [3]

[1] School of Economics and Management, Sanming University, No. 25 Jingdong Road, Sanming City, Fujian Province, China

[2] Durham University Business School, Mill Hill Lane, Durham University, DH1 3LB, U.K.

[3] Arden University Henley-on-Arden, United Kingdom

* Correspondence: 78308776@qq.com

Abstract

During public health emergencies (PHEs), social media generates massive online streaming features, yet only a sparse subset is truly relevant to sentiment analysis. Moreover, multimodal fusion must account for cross-modal interactions between text and images. To address these two issues, this study proposes an online multimodal sentiment analysis framework for PHE-related public opinion. Firstly, to handle the sparse relevant features among numerous online streaming features, we develop a Multimodal Online Divide-and-Conquer Markov Blanket Learning algorithm that incrementally selects robust sentiment-relevant features. Secondly, to capture cross-modal interactions, we design a Cross-Modal Interactive Enhanced Fusion Network with a two-stage cross-modal interactive attention mechanism, complemented by adaptive modal weighting and residual connections. Experiments on a large-scale PHE multimodal dataset show that our full model (O-DC + Our Fusion) achieves the best performance: 89.3% precision, 85.3% recall, 87.2% F1-score, and 87.1% accuracy, significantly outperforming state-of-the-art baselines.

Keywords: public health emergency; sentiment analysis; online feature selection; Markov blanket; cross-modal attention

1. Introduction

In recent years, public health emergencies (PHEs) have occurred frequently worldwide, such as the COVID-19 pandemic and Ebola virus outbreaks. During such events, social media platforms serve as central arenas where the public expresses emotions, disseminates information, and seeks support[1]. This data is inherently multimodal: text directly expresses viewpoints and emotions, while images contain rich contextual information and emotional cues, often reinforcing, supplementing, or even correcting the emotional tendencies conveyed by the text[2–4].

Therefore, accurate and real-time sentiment analysis of multimodal online public opinion during PHEs holds strategic value for relevant authorities to promptly grasp public sentiment trends, identify panic and rumors, assess the socio-psychological impact of intervention policies, and implement targeted risk communication and psychological support[5–7]. However, PHE public opinion data presents core challenges: social media generates massive online streaming features, yet only a sparse subset is truly relevant to sentiment analysis. Moreover, multimodal fusion must account for cross-modal interactions between text and images, rendering traditional sentiment analysis methods inadequate[8].

Current research primarily faces two major issues: ①Online Feature Selection: PHE public opinion data arrives continuously in a streaming manner, with the feature space dynamically changing. Traditional batch-mode feature selection methods cannot adapt to this streaming scenario,

 doi:10.20944/preprints202606.0301.v1

selecting sparse subset that is truly relevant to sentiment analysis from massive online streaming features. Although online feature selection research exists, it mostly focuses on single-modal text and lacks efficient online learning mechanisms for complex interdependencies among features. ②Cross-Modal Fusion Challenge: Simple feature concatenation or late decision fusion struggles to capture complex non-linear interactions and fine-grained semantic alignment between text and images. Despite advances in deep learning-based multimodal fusion, two issues remain particularly overlooked in the context of public health emergencies. One is modeling true bidirectional, multi-level interactions across modalities—not just coarse alignment—to extract complementary cues. The other is dynamic, context-aware fusion: the contribution of text versus image should not be fixed. Consider a post where the image shows a packed hospital corridor while the text merely says "a bit worried." Here, the visual emotion clearly dominates, yet typical methods fail to adapt accordingly.

Addressing the aforementioned challenges, the study proposes an innovative solution with the following main contributions: Firstly, we propose the Online Divide-and-Conquer Markov Blanket Learning Algorithm. We extend Markov blanket theory to the streaming multimodal feature selection scenario. This algorithm can approximate the sparse subset that is truly relevant to sentiment analysis from massive online streaming features via incremental updates, while ensuring time and space efficiency. Secondly, we design the Cross-Modal Interactive Enhanced Fusion Network. We introduce a Two-Stage Cross-Modal Interactive Attention mechanism and an Adaptive Modal Weighting module, achieving deep semantic fusion from local to global and from static to dynamic. Combined with residual connections, we construct a powerful and stable core model for multimodal emotion recognition.

The remainder of this paper is structured as follows: Section 2 reviews related work. Section 3 formally defines the problem and introduces the overall framework. Section 4 details the principles and derivation of the Online Divide-and-Conquer Markov Blanket Learning Algorithm. Section 5 elaborates on the design details of the Cross-Modal Interactive Enhanced Fusion Network. Section 6 presents the experimental setup, result analysis, and discussion. Section 7 concludes the paper and outlines future research directions.

2. Related Work

2.1. Sentiment Analysis in Public Health Emergencies

As an important carrier of modern risk communication, the information interaction characteristics of social media are highly consistent with the effect of social psychological force field. That is, public sentiment generates resonant diffusion through online clusters, which can form an emergency psychological field with radiation effects. Therefore, analyzing the discussions and emotional expressions of social media netizens regarding public health emergencies (PHEs) can predict potential risks, help the government better understand public responses, take timely intervention measures, reduce negative impacts, and maintain social order. PHEs trigger widespread public discourse on social media, making sentiment analysis an essential tool for monitoring social risk perception [9].

Sentiment analysis comprises text sentiment analysis and image sentiment analysis. Early research focused on unimodal textual analysis. The mainstream text sentiment analysis methods can be divided into sentiment lexicon-based methods[10][11], machine learning-based methods[12], and deep learning-based methods[13][14]. Among them, deep learning-based methods perform better in sentiment analysis, including CNN[15], CNN-BiLSTM[16], and BiGRU[17]. Recent advances have further pushed the boundaries of text sentiment analysis. For instance, Duan et al.[18] proposed a multi-level knowledge enhanced model integrating sentiment lexicon, syntactic dependencies, and conceptual knowledge via GCN and interactive attention. Liu et al.[19] addressed fine-tuning inconsistency by multi-prompt learning with diverse templates and answer engineering, ensembling predictions from an autoregressive model. Zhang et al.[20] introduced SynPrompt, which

 doi:10.20944/preprints202606.0301.v1

incorporates syntactic knowledge into prompt tuning via syntax-enhanced templates and data augmentation for few-shot fine-grained sentiment analysis.

Compared with textual sentiment, the sentiment expressed through images is embodied by visuals, objects, or scenes. Li et al.[21] used an improved CNN model to automatically extract the sentiment information of images. Cai et al.[22] combined global features with local object features for image sentiment analysis. However, affective expression during PHEs is inherently multimodal, with images often complementing or even contradicting textual sentiments[23]. Recognizing this gap, multimodal sentiment analysis has emerged as a promising direction for a holistic understanding of public opinion [24–27]. Gan et al.[28] proposed a cross-modal interactive Transformer that uses image caption generation to recover missing textual sentiment, dependency parsing and GCN to capture long-distance dependencies, and a TOP-n adjective-noun pair strategy to enhance visual sentiment features, thereby improving multimodal sentiment analysis during PHEs.

Despite this progress, most existing PHE-related multimodal studies employ simplistic fusion strategies and are evaluated on static datasets. This leaves a critical gap in methodologies capable of handling the streaming, high-velocity, and dynamically evolving nature of real-world social media data during public health emergencies.

2.2. Online Feature Selection for Streaming Data

In many real-world applications, feature streams refer to scenarios in which, under a fixed number of samples, feature dimensions increase dynamically over time—i.e., features arrive one by one and must be processed immediately. Online streaming feature selection algorithms have been developed to handle such dynamically changing feature spaces. Representative examples include Grafting[29], Alpha-investing[30], and OSFS[31] for single-label problems; MLFSL[32], MUCO[33], and OM-NRS [34] for multi-label online streaming feature selection; and group-based algorithms such as GFSSF[35] and OGFS[36] for features that arrive in groups.

Most existing online streaming feature selection methods are based on statistical correlations—they attempt to model and classify data by exploiting statistical relationships among variables. However, statistical correlation is only a surface-level phenomenon; it reflects the goodness-of-fit of probability distributions rather than the true underlying relationships . The genuine intrinsic relationship among things is causality, which implies that changes in the frequency or characteristics of a cause lead to corresponding changes in the effect . Bayesian networks (BNs) and Markov blankets (MB) are among the most commonly used tools for causal discovery[37]. The MB of a target variable comprises its parents (direct causes), children (direct effects), and spouses (other parents of its children). The concept of MB aligns naturally with the requirements of feature selection: features belonging to the MB correspond to strongly relevant features; features not in the MB but having a path to MB members are weakly relevant; features without any path to MB members are irrelevant. Consequently, feature selection can be cast as the problem of discovering the MB of the class variable. The MB captures local causal relationships, so identifying the MB of the class variable yields a logically optimal solution[38].

To improve the accuracy of causal feature selection on the datasets where only a small number of features are relevant, divide-and-conquer algorithms were introduced. The Min-Max MB (MMMB) algorithm[39] first applied a divide-and-conquer strategy by separating PC discovery from spouse discovery, thereby reducing the conditioning set size and improving accuracy. Ling[40] introduced cross-deletion: false positives are removed before each iteration, preventing them from entering the judgment step and thereby ensuring efficiency. However, neither of these methods can handle a particular structural configuration; they cannot guarantee the discovery of the correct MB in a faithful BN. To address this issue, Pena et al. used symmetry checks to ensure correct MB discovery and proposed PCMB[41]; unfortunately, the symmetry check requires repeated PC discovery, which compromises efficiency. To improve the efficiency of PCMB, IPCMB[42] adopted an iterative search strategy that reduces one round of PC discovery and identifies spouses more quickly.

 doi:10.20944/preprints202606.0301.v1

Nevertheless, all the above-mentioned divide-and-conquer algorithms require enumeration of conditional subsets. The enumeration process leads to exponential growth in computation, and the algorithms must also store separation sets for nodes independent of the target variable. Therefore, a divide-and-conquer causal feature selection algorithm with low time and space complexity is urgently needed. To fill this gap, this paper proposes an Online Divide-and-Conquer MB learning algorithm, termed O-DC, based on mutual information. By exploiting the differences in mutual information between the class variable's MB and other variables, O-DC compares the mutual information of each newly arrived feature with that of the currently selected parents/children and spouses. In this way, the algorithm significantly reduces time complexity.

2.3. Multimodal Feature Fusion Strategies

Effective multimodal sentiment analysis relies on sophisticated fusion strategies to capture complex interactions between modalities. To model cross-modal interactions more explicitly, advanced attention mechanisms have been widely adopted[43–45]. Harish et al.[46] proposed an attention-based deep neural network and training method that performs multi-level fusion of speech, image, and text modalities. Majumder et al.[47] introduced a hierarchical feature fusion strategy for three-modality sentiment analysis. Sun et al.[48] shared information across all modalities simultaneously, constructing an attention tensor from features extracted from each modality to improve sentiment classification accuracy. To fully measure the varying importance and interactive influences among modalities, Tsai et al.[49] employed cross-modal attention to map features from one modality to another, and Yang et al.[50] adopted cross-modal BERT to model inter-modality interactions. Wang et al.[51] introduced dual-modal and triple-modal cross-modal context information, using attention mechanisms to filter redundant information before performing sentiment analysis based on the fused features. Chen et al.[52] leveraged cross-modal attention to enable more sufficient intra-modal and inter-modal information interaction, employed gated units to remove redundant information, and used self-attention to assign attention weights. Yu et al.[53] proposed a deep attention and two-stage fusion method for image-text sentiment contrastive learning, using deep cross-modal attention for modality interaction and designing a cross-modal gated fusion module that employs gating and attention to achieve two-stage feature fusion.

Another important technique to facilitate deep network training is the residual connection, which alleviates gradient vanishing and degradation problems. Residual connections have been widely adopted in large models such as Transformer[54] and BERT[55]. By treating the output as a linear superposition of the input and a non-linear transformation of the input, residual connections introduce shortcut paths between network layers, enabling direct information flow across multiple layers[56]. In multimodal feature fusion, Han et al.[57] proposed a lightweight convolutional feature fusion strategy named residual merging, which improves training speed and enhances model practicality without sacrificing classification accuracy. Su et al.[58] used global attention and residual connections to fuse low-level and high-level features, thereby improving sentiment recognition accuracy.

In the context of emergency events, Shahid et al.[59] analysed user information expression and sentiment during disasters, employing GNNs to understand the complex relationships between text and visual modalities. Zeng et al.[60] addressed negative emotion recognition in public health emergencies from a multimodal data perspective, constructing a multimodal fine-grained negative emotion recognition model based on GCNs and ensemble learning, validated on COVID-19-related Weibo image-text data.

Despite these advances, several critical gaps remain. First, the fixed-weight fusion strategies commonly used in existing cross-modal interactive attention mechanisms lack dynamic adaptability; they cannot adjust the fusion ratios of modality features according to the input content. Although gated units can control information flow, they primarily rely on the internal states of recurrent networks and have limited ability to directly perceive and respond to cross-modal interaction features. Second, most multimodal sentiment models have been validated on well-curated public

 doi:10.20944/preprints202606.0301.v1

datasets, and whether they are applicable to real-world emergency public opinion sentiment analysis remains unclear. Public datasets typically contain mature, well-matched image-text pairs, whereas emergency event data often suffer from a mismatch in the volume of image and text data—a problem that has received little attention in existing multimodal sentiment modelling. Third, although residual connections have shown clear advantages in cross-modal feature fusion, their integration with dynamic cross-modal attention remains underexplored. Overall, research on multimodal sentiment analysis for emergency public opinion events is still fragmented, with insufficient attention paid to the image-text data imbalance problem and the deep interactive influences between public opinion images and texts.

To address the interactive influence between image and text features in emergency public opinion, this paper proposes a sentiment analysis model based on Adaptive Cross-Modal Fusion Attention. It consists of two-stage cross-modal interactive attention and adaptive attention: the former measures the semantic correlation and interaction between image and text to obtain enhanced cross-modal features, while the latter dynamically assigns fusion weights according to modality contributions to optimize the feature fusion strategy. Furthermore, residual connections are introduced to optimize the transmission path from original features to enhanced features, further improving the cross-modal fusion feature representation.

Based on the identified research gaps—namely, the lack of online causal feature selection methods for streaming multimodal data in public health emergencies (PHEs), the fixed-weight and dynamically inflexible nature of existing cross-modal fusion strategies, and the neglect of image–text data imbalance in real-world PHE scenarios—this paper proposes an online multimodal sentiment analysis framework. Specifically, we develop a Multimodal Online Divide-and-Conquer Markov Blanket Learning (O-DC) algorithm to incrementally select causally relevant features from streaming online data with low time complexity, and design a Cross-Modal Interactive Enhanced Fusion Network with two-stage cross-modal attention, adaptive modal weighting, and residual connections to dynamically capture deep text–image interactions.

3. Problem Definition and Overall Framework

3.1. Problem Formulation

Let a multimodal data case arriving at time step $S_i = (X_i^T, X_i^I, \mathcal{Y}_i)$, where:$X_i^T \in \mathbb{R}^{d_T}$ denotes the raw high-dimensional feature vector for the text modality. $X_i^I \in \mathbb{R}^{d_I}$ denotes the raw feature vector for the image modality. $\mathcal{Y}_i \in Y = \{\text{positive, neutral, negative}\}$ is the sentiment label.

The total raw feature space is $\mathcal{F}_i = X_i^T \cup X_i^I$, whose dimension $d = d_T + d_I$ can be very large and grows with the emergence of new words and visual patterns. Our objective is to design an online learning model $\mathcal{M}$ such that at any time step i, it can dynamically select a feature subset $\mathcal{S}_i \subset \mathcal{F}_i$ from the high-dimensional streaming feature space $\mathcal{F}_i$, with low redundancy and strong relevance to $\mathcal{Y}_i$. Then, based on $\mathcal{S}_i$, a powerful classification function f: $\mathcal{X}_i^T \times \mathcal{X}_i^I \to \mathcal{Y}$ is constructed to accurately predict $\mathcal{Y}_i$, where $\mathcal{X}_i^T$, $\mathcal{X}_i^I$ are the feature representations after selection.

3.2. Overall Framework

The proposed framework consists of two core stages: ①Online Multimodal Feature Selection Stage: For each newly arrived case s_t, its raw text and image features are input into the Online Divide-and-Conquer Markov Blanket Learning Algorithm. This algorithm maintains two dynamic "feature pools" —the text-relevant feature set PC^T and the image-relevant feature set PC^I. Based on incremental mutual information computation, the algorithm decides whether to include new features into the corresponding pool or discard old ones. The outputs are the filtered text feature subset $\mathcal{X}_i^T$ and image feature subset $\mathcal{X}_i^I$. ②Cross-Modal Interactive Enhanced Fusion and Classification Stage: The filtered features are first encoded by modality-specific encoders to obtain deep representations H^T and H^I. Subsequently, they enter the Cross-Modal Interactive Enhanced Fusion Network. This network generates a context-aware fused representation H^{fusion} through two-

 doi:10.20944/preprints202606.0301.v1

stage interactive attention and adaptive weight assignment. Finally, H^{fusion} passes through residual connections and a fully connected classification layer to output the sentiment prediction $\mathcal{Y}_i$. The model is updated via an online cross-entropy loss.

4. Multimodal Online Divide-And-Conquer Markov Blanket Learning Algorithm

In real-world streaming data applications—such as social media monitoring during public health emergencies (PHEs)—raw features naturally arrive in multimodal form. The primary raw features include textual content and visual content. Traditional online feature selection methods often operate on pre-processed, unimodal representations, lacking a principled framework to jointly and efficiently process heterogeneous data streams.

Online Markov Blanket (MB) learning aims to dynamically identify the minimal optimal feature set for predicting a target variable from a stream of incoming features. However, existing methods face three core challenges in multimodal settings: (1) Cross-modal redundancy, where features from different modalities convey overlapping information; (2) Heterogeneous dependency modeling, due to the distinct statistical nature of text (discrete) and image (continuous) features; and (3) Computational inefficiency, especially when exhaustive subset enumeration is required.

To address these challenges, we propose the Multimodal Online Divide-and-Conquer Markov Blanket Learning (M-O-DC) algorithm.

4.1. Markov Blanket and Mutual Information

In probabilistic graphical models, the Markov blanket $MB(Y)$ of a target variable Y is defined as the set of variables that renders Y conditionally independent of all other variables in the graph. Formally, for any variable $X_i \notin MB(Y)$, we have $X_i \perp Y|MB(Y)$. The Markov blanket of Y consists of its parents, children, and the other parents of its children (spouses). For classification tasks, $MB(Y)$ constitutes the minimal sufficient feature subset for predicting Y, as it contains all the information necessary to determine Y while discarding irrelevant and redundant features.

Mutual information (MI) is a fundamental measure of dependency between two random variables:

$$I(X,Y) = \sum_{x\in X}\sum_{y\in Y} p(x,y)\log\frac{p(x,y)}{p(x)p(y)} \quad (1)$$

Conditional mutual information $I(X;Y|Z)$ quantifies the association between X and Y given Z, and is defined analogously using conditional probabilities.

$$I(X;Y|Z) = \sum_{x,y,z} p(x,y,z)\log\frac{p(x,y|z)}{p(x|z)p(y|z)} \quad (2)$$

A feature X_i belongs to $MB(Y)$ if and only if it provides unique predictive information about Y that is not contained in any other subset of $MB(Y)$. This property is fundamental for online feature selection: we aim to incrementally maintain an approximation of $MB(Y)$ as features arrive in a stream.

4.2. M-O-DC Algorithm

4.2.1. Problem Setup and Multimodal Input Definition

Let the streaming feature space consist of a sequence of textual features $\mathbf{F_T} = \{\mathbf{X_i^T}\}$ and Image visual features $\mathbf{F_I} = \{\mathbf{X_i^I}\}$, arriving sequentially**.** The sentiment label $\mathbf{Y}$ is influenced by both modalities. We define the candidate **PC** set per modality, denoted $\mathbf{C\_PC_Y^m}$ for $\mathbf{m} \in \{\mathbf{T}, \mathbf{I}\}$, as the dynamically maintained set of features that are likely parents or children of $\mathbf{Y}$ within modality $\mathbf{m}$.

 doi:10.20944/preprints202606.0301.v1

4.2.2. Decoupled Multimodal Processing

Input: Current text PC set $\mathbf{C\_PC_Y^T}$, image PC set $\mathbf{C\_PC_Y^I}$, new feature $\mathbf{X_i^m}$, estimated target distribution $\hat{\mathbf{p}}_t(\mathbf{y})$, modality-specific thresholds $\boldsymbol{\tau}_{\mathbf{inc}}^{\mathbf{m}}$, $\boldsymbol{\tau}_{\mathbf{red}}^{\mathbf{m}}$. $\boldsymbol{\tau}_{\mathbf{inc}}^{\mathbf{m}}$ is the **inclusion threshold**. When a new feature $\mathbf{X_i^m}$ arrives, if its mutual information with the target variable $\mathbf{Y}$, denoted $\mathbf{I}(\mathbf{X_i^m}, \mathbf{Y})$, exceeds this threshold, the algorithm considers $\mathbf{X_i^m}$ a candidate for the Parent-Child (PC) set.

Output: Updated multimodal Markov Blanket candidate set $\mathbf{MB}(\mathbf{Y})$.

Step A: Intra-modal PC Learning

For a newly arrived feature $\mathbf{X_i^m}$

①Independence Test: Compute $\mathbf{I}(\mathbf{X_i^m}, \mathbf{Y})$, if $\mathbf{I}(\mathbf{X_i^m}, \mathbf{Y}) < \boldsymbol{\tau}_{\mathbf{inc}}^{\mathbf{m}}$, discard $\mathbf{X_i^m}$ as irrelevant.

②Redundancy Removal: Compare $\mathbf{X_i^m}$ with features in $\mathbf{C\_PC_Y^m}$ using symmetric uncertainty (SU). Remove any $\mathbf{X_k^m} \in \mathbf{C\_PC_Y^m}$ that becomes redundant given $\mathbf{X_i^m}$.

③PC Update: If $\mathbf{X_i^m}$ provides significant non-redundant information, add it to $\mathbf{C\_PC_Y^m}$

Step B: Cross-modal Spouse Learning

When a new PC node $\mathbf{X_i^m}$ is identified, trigger cross-modal spouse discovery:

①For each cross-modal candidate $\mathbf{Z_j^n} \in \mathbf{C\_PC_Y^n}$ $(\mathbf{n} \neq \mathbf{m})$, check if $\mathbf{X_i^m} \in \mathbf{PC}(\mathbf{Z_j^n})$ or vice versa.

②Apply the Cross-modal Spouse Identification Theorem: If $\mathbf{Z_j^n}$ and $\mathbf{X_i^m}$ are strongly associated, and the presence of $\mathbf{X_i^m}$ increases the conditional mutual information $\mathbf{I}(\mathbf{Y}; \mathbf{Z_j^n}|\mathbf{X_i^m}) > \boldsymbol{I}(\boldsymbol{Y};\mathbf{Z_j^n})$, then $\mathbf{Z_j^n}$ is a cross-modal spouse of $\mathbf{Y}$.

③Add qualifying $\mathbf{Z_j^n}$ to the spouse set $\mathbf{SP}(\mathbf{Y})$, which is inherently multimodal.

This design ensures that a newly discovered textual PC can help identify a relevant visual spouse, and vice versa, enabling a complete causal structure discovery across modalities.

4.3. Online Mutual Information Estimation for Text and Image

4.3.1. Text Features

Text features are typically discretized.

Sufficient Statistics: Maintain a count matrix $\mathbf{C_T}[\mathbf{f_t}, \mathbf{y}]$ for each discrete feature $\mathbf{f_t}$ Online Update Rules: $\mathbf{C_T}[\mathbf{X}, \mathbf{y}] \leftarrow \mathbf{C_T}[\mathbf{X}, \mathbf{y}] + \mathbf{I}(\mathbf{f_t} = \mathbf{X}, \mathbf{Y} = \mathbf{y}) - \mathbf{expired\_contribution}$ MI Computation (with Laplace Smoothing):

$$\mathbf{I_T}(\mathbf{Y};\ \mathbf{f_t}) = \sum_{\mathbf{X},\mathbf{y}} \widetilde{\mathbf{P}}_{\mathbf{T}}(\mathbf{X}, \mathbf{y}) \mathbf{log} \frac{\widetilde{\mathbf{P}}_{\mathbf{T}}(\mathbf{X}, \mathbf{y})}{\widetilde{\mathbf{P}}_{\mathbf{T}}(\mathbf{X})\widetilde{\mathbf{P}}(\mathbf{y})} \tag{3}$$

Where $\widetilde{\mathbf{P}}$ denotes smoothed probability estimates

4.3.2. Image Features

Image features are high-dimensional continuous vectors.

Sufficient Statistics: Use online Kernel Density Estimation (KDE) to approximate class-conditional densities.

Incremental Parameter Updates (Exponential Moving Averages):

$$\boldsymbol{\mu}_{\mathbf{y}}^{\mathbf{I}} \leftarrow \boldsymbol{\lambda}\boldsymbol{\mu}_{\mathbf{y}}^{\mathbf{I}} + (\mathbf{1} - \boldsymbol{\lambda})\mathbf{X}_{\mathbf{I,new}} \tag{4}$$

$$\left(\boldsymbol{\sigma}_{\mathbf{y}}^{\mathbf{I}}\right) \leftarrow \boldsymbol{\lambda}\left(\boldsymbol{\sigma}_{\mathbf{y}}^{\mathbf{I}}\right) + (\mathbf{1} - \boldsymbol{\lambda})(\mathbf{X}_{\mathbf{I,new}} - \boldsymbol{\mu}_{\mathbf{y}}^{\mathbf{I}} \tag{5}$$

MI Estimation via KL Divergence:

$$\mathbf{I_I}(\mathbf{Y}; \mathbf{f_I}) \approx \sum_{\mathbf{I}} \mathbf{P}(\mathbf{y}) \mathbf{D_{KL}}\left(\mathbf{P}(\mathbf{f_I}|\mathbf{y}) \| \mathbf{P}(\mathbf{f_I})\right) \tag{6}$$

Where $\mathbf{P}(\mathbf{f_I}|\mathbf{y})$ is modeled using online-updated Gaussian approximations.

4.3.3. Cross-modal Conditional MI Approximation

 doi:10.20944/preprints202606.0301.v1

Computing $\mathbf{I(Y;\ Z_j^I|X_i^T)}$ is challenging due to heterogeneity. M-O-DC uses a chain factorization with cross-modal independence assumption:

$$\mathbf{I\left(Y;\ Z_j^I|X_i^T\right) \approx I\left(Y;Z_j^I\right) - \sum_k w_k \cdot I(Z_j^I; C_k^T)} \tag{7}$$

Where $\mathbf{C_k^T}$ are the top-k text features in $\mathbf{C\_PC_Y^T}$ most correlated with $\mathbf{Z_j^I}$. This approximation is effective when text and image features are conditionally independent given $\mathbf{Y}$.

4.4. Multimodal Dynamic Adaptation Mechanisms

4.4.1. Modality-Adaptive Thresholds

Thresholds $\boldsymbol{\tau_{inc}^m}$ and $\boldsymbol{\tau_{red}^m}$ are dynamically adjusted per modality:

$$\boldsymbol{\tau_{inc}^m \leftarrow \tau_{inc}^m \cdot \left(1 + \alpha \cdot \frac{\left|C_{PC_Y^m}\right|}{N_{max}^m}\right)} \tag{8}$$

This accounts for differences in feature density and redundancy between text and image streams.

4.4.2. Multimodal Concept Drift Detection

Public sentiment may evolve at different rates across modalities. M-O-DC defines a multimodal causal drift statistic:

$$\mathbf{D_t^{(m,n)} = \left| \frac{1}{|C\_PC_Y^m|} \sum_{X_i^m \in C\_PC_Y^m} I(Y; X_i^m) - window\_averaged\ MI \right|} \tag{9}$$

If $D_t^{(m,n)} > k \cdot \sigma_D$, a drift is detected, triggering adaptive forgetting in the affected modality.

5. Cross-Modal Interactive Enhanced Fusion Network

5.1. Deep Encoder Design

5.1.1. Formalization of Multimodal Feature Encoding

The features filtered by M-O-DC, $\widetilde{X}^T \in R^{d_T}$ and $\widetilde{X}^I \in R^{d_I}$, are encoded by modality-specific deep encoders.

Text Encoder (based on Transformer architecture):

$$H^T = \text{Encoder}_T(\widetilde{X}^T) \in \mathbb{R}^{L \times d_h} \tag{10}$$

Where $\text{Encoder}_T = \text{Stack}\{\text{MultiHeadSelfAttn}, \text{FFn}\}_{l=1}^{N_T}$, Let $L = \max\left(\left\lfloor \frac{d_T}{d_{token}} \right\rfloor, L_{min}\right)$ be the sequence length.

Image Encoder (based on Vision Transformer):

$$H^I = \text{Encoder}_I(\widetilde{X}^I) \in \mathbb{R}^{N \times d_h} \tag{11}$$

where $\widetilde{X}^I$ is split into $N = \left\lfloor \frac{H \times W}{P^2} \right\rfloor$ patches, each projected to dimension $P \times P \times C \rightarrow d_h$.

5.2. Two-Stage Cross-Modal Interactive Attention:

5.2.1. Stage 1: Bidirectional Cross-Modal Attention

We compute cross-modal attention weights in both directions to establish bidirectional interactive mappings. For text-to-image attention, text features serve as queries and image features as keys:

Text-to-Image Attention:$Q_T = H^T W_{Q^T} \in R^{L \times d_k}$, $K_I = H^I W_{k^I} \in R^{N \times d_k}$, $V_I = H^I W_{v^I} \in R^{N \times d_h}$

 doi:10.20944/preprints202606.0301.v1

W_{Q^T}, W_{k^I}, W_{v^I}: learnable weight matrices for query, key, value projections

d_k: dimension of query/key vectors (usually $d_k = \frac{d_h}{h}$ where h is number of attention heads)

Q^T:query matrix from text; k^I: key matrix from image; V_I: value matrix from image.

The attention weight matrix is:

$$A_{T\to I}[l,n] = \frac{\exp\left(\frac{1}{\sqrt{d_k}}\langle q_l^T, k_n^I\rangle\right)}{\sum_{m=1}^{N}\exp\left(\frac{1}{\sqrt{d_k}}\langle q_l^T, k_m^I\rangle\right)} \tag{12}$$

where $\langle\cdot,\cdot\rangle$ denotes the inner product.

The text-to-image attention output (cross-modal feature) is: $Attn_{T\to I} = A_{T\to I}V_I \in \mathbb{R}^{L\times d_h}$, where $V_I = H^I W_V^I \in \mathbb{R}^{N\times d_h}$

Similarly, image-to-text attention is computed symmetrically: $Q_I = H^I W_{Q^I} \in R^{N\times d_k}$, $K_T = H^T W_{k^T} \in R^{L\times d_k}$, $V_T = H^T W_{v^T} \in R^{L\times d_h}$.

$$A_{I\to T}[n,l] = \frac{\exp\left(\frac{1}{\sqrt{d_k}}\langle q_n^I, k_l^T\rangle\right)}{\sum_{l=1}^{L}\exp\left(\frac{1}{\sqrt{d_k}}\langle q_n^I, k_l^T\rangle\right)} \tag{13}$$

Image-to-Text Attention (symmetric): $Attn_{I\to T} = A_{I\to T}V_T \in \mathbb{R}^{N\times d_h}$

The contextual representations after first-stage interaction (enhanced by residual connection and FFN) are:

$$C^{T\leftarrow I} = \text{LayerNorm}(H^T + \text{FFN}(Attn_{T\to I})) \tag{14}$$

$$C^{I\leftarrow T} = \text{LayerNorm}(H^I + \text{FFN}(Attn_{I\to T})) \tag{15}$$

Where, FFN is a two-layer MLP used for feature transformation: $\text{FFN}(x) = \text{GELU}(xW_1 + b_1)W_2 + b_2$.

5.2.2. Stage 2: Deep Interactive Normalization

The second stage performs deep processing of the weighted features generated in the first stage. We apply softmax normalization to the interaction matrices, transforming attention weights into probability distributions that quantify the matching relationships between local textual and visual features. This normalization process not only quantifies local feature alignments but also mines potential cross-modal semantic associations that may not be directly observable.

We implement this using Multi-Head Co-Attention (MHCA):

$$\tilde{C}^T = \text{MHCA}(C^{T\leftarrow I}, C^{I\leftarrow T}, C^{I\leftarrow T}) \tag{16}$$

$$\tilde{H}^T = \text{LayerNorm}(\tilde{C}^T + \text{FFN}(\tilde{C}^T)) \tag{17}$$

$$\tilde{C}^I = \text{MHCA}(C^{I\leftarrow T}, C^{T\leftarrow I}, C^{T\leftarrow I}) \tag{18}$$

$$\tilde{H}^I = \text{LayerNorm}(\tilde{C}^I + \text{FFN}(\tilde{C}^I)) \tag{19}$$

Compared to methods that only perform direct feature weighting based on initial attention weights, our two-stage hierarchical processing captures finer complementary information and

 doi:10.20944/preprints202606.0301.v1

interaction patterns between modalities, enhancing the expressive power of cross-modal features for sentiment semantics.

5.3. Adaptive Modal Weighting

5.3.1. Enhanced Cross-Modal Features

We first enhance the cross-modal features by adding the original representations to the attention outputs from Stage 1:$\hat{H}^T = H^T + \text{Attn}_{T\to I} \in R^{L\times d_h}$, $\hat{H}^I = H^I + \text{Attn}_{I\to T} \in R^{N\times d_h}$

These enhanced features integrate both the original modality information and the cross-modal influences.

5.3.2. Gated Weight Generation Network

To dynamically adjust the fusion weights for text and image modalities based on the current context, we concatenate the enhanced features along the sequence dimension and apply a learnable gating network.

First, we obtain a global representation by mean pooling over the sequence dimension:

$$\bar{h}^T = \frac{1}{L}\sum_{l=1}^{L} \hat{h}_l^T \in R^{d_h} \tag{20}$$

$$\bar{h}^I = \frac{1}{N}\sum_{n=1}^{N} \hat{h}_n^I \in R^{d_h} \tag{21}$$

Then we concatenate them to form a joint vector:

$$z = [\bar{h}^T; \bar{h}^I] \in R^{2d_h} \tag{22}$$

The gating network consists of two fully connected layers with ReLU activation, followed by a sigmoid layer to produce the text weight:

$$h_1 = \text{ReLU}(W_1 z + b_1), \ W_1 \in R^{d_h\times 2d_h}, \ b_1 \in R^{d_h} \tag{23}$$

$$h_2 = \text{ReLU}(W_2 h_1 + b_2), \ W_2 \in R^{d_h\times d_h}, \ b_2 \in R^{d_h} \tag{24}$$

$$g_t = \sigma(W_3 h_2 + b_3) \ b_2), \ W_3 \in R^{1\times d_h}, \ b_3 \in R \tag{25}$$

$$g_i = 1 - g_t \tag{26}$$

This design forms a closed learning loop from cross-modal attention computation to feature weight assignment, enabling the model to adaptively balance modality contributions based on the semantic content of each post.

5.3.3. Weighted Fusion

Using the adaptive weights, we compute the modality-weighted representations:

$$\tilde{H}^T_{\text{weight}} = g_t \cdot \tilde{H}^T, \ \tilde{H}^I_{\text{weight}} = g_i \cdot \tilde{H}^I \tag{27}$$

 doi:10.20944/preprints202606.0301.v1

where $\tilde{H}^T$ and $\tilde{H}^I$ are the outputs of the second-stage attention (Equations 21 and 23). These weighted features are then fused via concatenation:

$$F_{fused} = [\tilde{H}^T_{weight}; \tilde{H}^I_{weight}] \in R^{(L+N)\times d_h} \quad (28)$$

5.4. Residual Fusion and Classification

5.4.1. Residual-Enhanced Feature Fusion

To optimize gradient flow and reduce information loss during feature transmission, we introduce residual connections. The fused features are passed through a feed-forward network with a skip connection:

$$H^{res} = \text{LayerNorm}(F_{fused} + \text{FFN}(F_{fused})) \in R^{(L+N)\times d_h} \quad (29)$$

This residual design preserves original information while enhancing the expressive capacity of cross-modal features, mitigating overfitting and improving model generalization.

5.4.2. Self-Attentive Pooling

We employ self-attentive pooling to aggregate the sequence of fused features into a single vector:

$$\alpha = \text{softmax}(H^{res}W_a) \in \mathbb{R}^{L+N},\ W_a \in R^{d_h\times 1} \quad (30)$$

$$h_{pool} = \sum_{i=1}^{L+N} \alpha_i h_i^{res} \in R^{d_h} \quad (31)$$

where h_i^{res} denotes the i-th row of H^{res}.

5.4.3. Sentiment Classification

The pooled features are passed through a fully connected layer with ReLU activation, followed by Dropout for regularization:

$$h_{hidden} = \text{ReLU}(W_{fc1}h_{pool} + b_{fc1}),\ W_{fc1} \in R^{d_h\times d_h},\ b_{fc1} \in R^{d_h} \quad (30)$$

$$h_{drop} = \text{Dropout}(h_{hidden}) \quad (31)$$

The final classification layer uses softmax to output probabilities for three sentiment categories (negative, neutral, positive):

$$p(y|S) = \text{softmax}(W_{cls}h_{drop} + b_{cls}),\ W_{cls} \in R^{3\times d_h},\ b_{cls} \in R^3 \quad (32)$$

5.5. Loss Function and Optimization

5.5.1. Multi-Task Loss

We employ multi-task learning with three loss components.
Classification Loss (cross-entropy):

 doi:10.20944/preprints202606.0301.v1

$$\mathcal{L}_{cls} = -\frac{1}{B}\sum_{i=1}^{B}\sum_{c=1}^{C} y_{i,c}\log p_{i,c} \tag{33}$$

where B is the batch size, $y_{i,c}$ is the ground-truth one-hot indicator, and $p_{i,c}$ is the predicted probability for class C.

Contrastive Learning Loss to enhance representation learning:

$$\mathcal{L}_{cont} = -\frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp\left(\mathrm{sim}(z_i, z_i^+)/\tau\right)}{\sum_{j=1}^{B}\exp\left(\mathrm{sim}(z_i, z_j^+)/\tau\right)} \tag{34}$$

where $z_i = h_{pool,i}$ (the pooled feature for sample i), z_i^+ is a positive sample obtained through data augmentation (e.g., random masking of text or image patches), $\mathrm{sim}(\cdot,\cdot)$ denotes cosine similarity, and τ is a temperature parameter.

Causal Regularization Loss to encourage causally stable representations:

$$\mathcal{L}_{causal} = \sum_{m\in\{T,I\}} \|\nabla_{\tilde{H}^m}\mathcal{L}_{cls}\|_F^2 \tag{35}$$

where $\nabla_{\tilde{H}^m}\mathcal{L}_{cls}$ is the gradient of the classification loss with respect to the modality-specific representations, and $\|\cdot\|_F$ denotes the Frobenius norm. This term penalizes sensitivity to small perturbations, promoting features that capture causal factors rather than spurious correlations.

The total loss is:

$$\mathcal{L} = \mathcal{L}_{cls} + \lambda_1\mathcal{L}_{cont} + \lambda_2\mathcal{L}_{causal} \tag{36}$$

where λ_1 and λ_2 are hyperparameters balancing the contributions.

5.5.2. Optimization

We optimize the model using AdamW with weight decay, which prevents excessive weights and overfitting better than traditional Adam. The learning rate is set to 5×10^{-5}, and the model is updated online as new data streams arrive (batch size = 1, update every 100 samples). All weight matrices are initialized using Xavier uniform initialization, and biases are initialized to zero.

6. Experiments and Results Analysis

6.1. Experimental Data

(1) Dataset Collection

In China, Weibo and Xiaohongshu have emerged as pivotal platforms for public opinion and sentiment expression. Owing to their extensive user bases, image content related to public health emergencies on these platforms typically encapsulates richer information, aggregating substantial volumes of authentic user perspectives and emotional responses, while encompassing diverse stances, viewpoints, and affective tendencies. Consequently, this study utilizes data sourced from Weibo and Xiaohongshu to validate the effectiveness of the proposed method.

By crawling public posts from Weibo and Xiaohongshu between January 2020 and December 2022, this study collects public opinion data from three public health emergencies on the Weibo and Xiaohongshu platforms, comprising 15,677 text-image pairs. The data distribution across the events is as follows: 10,245 pairs related to COVID-19, 3,210 pairs to Avian Flu, and 2,222 pairs to Monkeypox.

(2) Data Annotation

 doi:10.20944/preprints202606.0301.v1

The annotation of the dataset involved a structured dual-process approach, combining both manual and semi-automated methods tailored to the data modality. For image sentiment ,a manual dual-annotator protocol was employed. Each image was independently labeled as "Negative," "Neutral," or "Positive" by two annotators. Inter-annotator agreement was measured using Cohen's Kappa coefficient (0.7123), indicating substantial consistency. Any discordant labels were resolved through subsequent discussion between the annotators to reach a consensus. For text sentiment annotation, a hybrid semi-automated pipeline was implemented. Initial sentiment labels ("Negative," "Neutral," "Positive") were generated automatically using the SnowNLP toolkit. These automated labels were then validated by two independent human annotators. Instances where the manual annotations disagreed with the automated output or with each other were flagged. These discrepancies were subsequently reviewed and adjudicated through manual re-evaluation to ensure final label accuracy.The finalized sentiment distribution across the three public health event datasets is presented in Table 1.

Table 1. The Sentiment Annotation Results of Public Opinion in Three Sudden Events.

Public Health Emergency Names	Total Sample Size	Positive	Neutral	Negative
COVID-19	10,245	2,377	4,252	3,616
Avian Influenza	3,210	803	1445	962
Monkeypox	2,222	489	934	799
Total	15,677	3,669	6631	5377

(3) Experimental Environment and Evaluation Metrics

The experiments were conducted using the PyCharm IDE with Python 3.9. All models were trained on an NVIDIA RTX 3060 GPU with 32 GB of dedicated memory in Table 2. The AdamW optimizer was employed for training across all experiments, with an initial learning rate of 5×10^{-5}. To ensure fair comparisons and mitigate confounding variables, identical training and testing splits were used for all models. Specifically, the dataset was partitioned into 80% for training and 20% for testing.

Table 2. Experimental Parameter Settings.

Stage	Parameters Value
O-DC thresholds	$\tau_{text} = 0.35$, $\tau_{image} = 0.40$
Optimizer	Learning_rate=5×10^{-5}, β=(0.9, 0.999)

 doi:10.20944/preprints202606.0301.v1

	batch size=1
Training: Online learning	
	update every 100 samples

This study uses four standard evaluation metrics: Accuracy (A), Precision (P), Recall (R), and F1-score (F1).

6.2. Main Results and Analysis

6.2.1. The Classification Results

The model proposed in this paper was applied to three public health emergencies, and the classification results are shown in Table 3.

Table 3. Sentiment Classification Result Analysis of Three Public Health Opinion Events.

Events	Sentiment Category	Precision (P)	Recall (R)	F1-score (F)	Accuracy (A)
COVID-19	positive	91.15%	82.54%	86.64%	86.62%
	Neutral	88.45%	82.78%	85.52%	
	Negative	90.31%	85.35%	87.75%	
Avian Influenza	positive	91.37%	89.29%	90.30%	87.15%
	Neutral	84.76%	81.78%	83.27%	
	Negative	90.48%	88.12%	89.25%	
Monkeypox	positive	91.23%	86.72%	88.92%	87.47%
	Neutral	83.97%	80.34%	82.12%	
	Negative	92.09%	90.75%	91.38%	
Average of the three events		89.31%	85.30%	87.24%	87.08%

Based on the sentiment classification results of three public health events—COVID-19, Avian Influenza, and Monkeypox—the sentiment analysis model demonstrates strong overall performance, with an average accuracy of 87.08% and a macro-average F1 score of 87.24%, indicating good generalization ability and practical application potential. In terms of sentiment-specific recognition,

 doi:10.20944/preprints202606.0301.v1

the model performs best in identifying negative sentiments, with F1 scores for the negative class exceeding 87.75% across all three events, reaching as high as 91.38% for Monkeypox. This highlights the model's clear advantage in capturing negative public opinion during crisis events. Positive sentiment recognition is also stable, with precision consistently above 91% and a low false positive rate. However, the model shows a notable weakness in recognizing neutral sentiment, with F1 scores for the neutral class falling below 85.5% across all events—reaching as low as 82.12%—and generally low recall, reflecting difficulty in accurately identifying texts with weak or ambiguous emotional polarity. Overall, the model is ready for practical deployment, particularly for negative public opinion monitoring in public health contexts.

6.2.2. Baseline Models

we organize the baselines into three categories—online feature selection, text-only, and multimodal fusion approaches—and further design combined experiments that pair different feature selection and fusion strategies to isolate the contribution of each component.①Online Feature Selection Baselines: Online Group LASSO (OGL): Regularization parameter $\lambda = 0.01$, group structure defined by modality. Alpha-investing (AI): Initial alpha budget $\alpha_0 = 0.05$, $\omega = 1.0$. ②Text-Only Baselines: Online SVM (Text):Based on TF-IDF features, RBF kernel, $C = 1.0$, online learning rate $\eta = 0.01$. Online BERT: Based on DistilBERT-base, hidden layer dimension 768, learning rate 2e-5, updated every 100 samples. ③Multimodal Fusion Baselines: Early Fusion (EF): Text and image features concatenated at the input layer (text 768-dim + image 512-dim = 1280-dim). Late Fusion (LF):Text and image predictions are combined using a weighted average, with weights determined on the validation set. Cross-Modal Attention (CMA): Number of attention heads = 8, hidden dimension = 256. MARN: Memory unit size = 128, number of memory slots = 16. CLIP (finetuned)**:** We employ the pretrained CLIP (ViT-B/32) model, which consists of a Vision Transformer for images and a Transformer for text. The image and text features are concatenated and fed into a linear classifier for sentiment prediction. Fine-tuning follows the same online learning protocol (batch size = 1, update every 100 samples, learning rate = 2e-5) to ensure fair comparison.

Comparative Experimental Design: The proposed O-DC method is combined with different fusion approaches: O-DC+EF; O-DC+LF; O-DC+CMA. Our fusion network is combined with different feature selection methods: OGL+Our Fusion; AI+Our Fusion.

6.2.3. Results and Discussion

To comprehensively evaluate the effectiveness of the proposed online dynamic cross-modal fusion framework (O-DC + Our Fusion), we compare it against a diverse set of baseline models, including text-only classifiers, multimodal fusion architectures, and online feature selection methods. As reported in Table 4, all models are evaluated on the test stream in terms of Precision (P), Recall (R), Macro-F1, and Accuracy (A).

Table 4. Main Performance Comparison (Average on Test Stream).

Model	Precision (P)	Recall (R)	F1-Score (F)	Accuracy (A)
Online SVM (Text)	72.2%	65.5%	70.8%	72.5%
Online BERT	75.5%	66.3%	76.5%	78.2%
EF + OGL	64.7%	79.2%	78.9%	80.1%
CMA	77.3%	81.3%	80.8%	82.3%
MARN	82.7%	91.3%	81.6%	83.1%

 doi:10.20944/preprints202606.0301.v1

Model	Precision (P)	Recall (R)	F1-Score (F)	Accuracy (A)
CLIP (finetuned)	88.1%	83.9%	85.9%	86.2%
O-DC + CMA	85.6%	71.8%	83.2%	84.7%
O-DC + Our Fusion	89.3%	85.3%	87.2%	87.1%

Among the text-only baselines, Online BERT significantly outperforms Online SVM, achieving a 5.7% higher F1-Score 6.5% vs. 70.8%) and a 5.7% improvement in accuracy (78.2% vs. 72.5%). This demonstrates the superiority of pre-trained language models in capturing semantic nuances from textual content in public health emergency events. However, both text-only models exhibit relatively low recall (below 67%), indicating a tendency to miss a substantial portion of relevant samples, particularly in imbalanced streaming settings.

Incorporating visual information through multimodal fusion leads to consistent performance gains. For instance, EF + OGL improves recall to 79.2% and F1-Score to 78.9%, suggesting that early integration of visual features helps recover false negatives. The Cross-Modal Attention (CMA) model further boosts performance, achieving 81.3% recall and 80.8% F1-Score, which underscores the importance of modality interaction mechanisms. The MARN model, equipped with memory-augmented networks, achieves the highest recall among all baselines (91.3%) and an accuracy of 83.1%, validating the effectiveness of modeling temporal dependencies in online multimodal streams.

The inclusion of a CLIP-based vision-language pretrained model (CLIP finetuned) provides a strong contemporary baseline. CLIP achieves 88.1% precision, 83.9% recall, 85.9% F1-Score, and 86.2% accuracy, outperforming all traditional multimodal fusion methods and text-only models. This result confirms that leveraging large-scale pretraining on image-text pairs yields powerful multimodal representations even in specialized domains like PHE sentiment analysis. However, CLIP still exhibits a precision-recall trade-off: its recall (83.9%) is lower than MARN's (91.3%), indicating that some sentiment signals are missed, while its precision (88.1%) is higher than MARN's (82.7%), reflecting fewer false positives.

Despite these advances, baseline methods still face trade-offs between precision and recall. The proposed O-DC+CMA combination partially mitigates this issue, achieving a balanced performance with 85.6% precision and 83.2% F1-Score. Nevertheless, it still lags behind in recall (71.8%), suggesting that attention-based cross-modal interaction alone is insufficient for comprehensive feature selection in dynamic environments.

Remarkably, the full proposed model (O-DC + Our Fusion) achieves the best overall performance across all metrics: 89.3% precision, 85.3% recall, 87.2% F1-Score, and 87.1% accuracy. Compared to the strongest baseline MARN, our method improves F1-Score by 5.6% and accuracy by 4.0%, while maintaining a superior balance between precision and recall. More importantly, compared to the state-of-the-art pretrained model CLIP, our method yields gains of +1.2% in precision, +1.4% in recall, +1.3% in F1-Score, and +0.9% in accuracy. This demonstrates that the integration of online dynamic selection (M-O-DC) with our specially designed two-stage interactive fusion network adds significant value over simply fine-tuning a generic vision-language model. The improvement can be attributed to two factors: (1) M-O-DC adaptively filters irrelevant and redundant features in the streaming setting, reducing noise and focusing on task-relevant information; (2) the two-stage cross-modal attention with adaptive weighting captures fine-grained semantic alignments and dynamically balances modality contributions, which is crucial for PHE posts where modality importance varies greatly.

The online streaming performance evolution (Figure 1) further illustrates the robustness of our method. Prior to the simulated concept drift at batch 20, our model exhibits the fastest growth and reaches the highest F1-Score plateau. Upon drift, both MARN and O-DC+CMA experience noticeable degradation, while CLIP (not shown in Figure 1 due to its offline nature) would likely suffer similarly

 doi:10.20944/preprints202606.0301.v1

if evaluated online. In contrast, our proposed method maintains its performance with only a marginal dip and quickly recovers, showcasing its resilience to distributional shifts. This resilience is attributed to the online dynamic selection mechanism, which adaptively identifies and emphasizes informative features while discarding noisy or outdated ones.

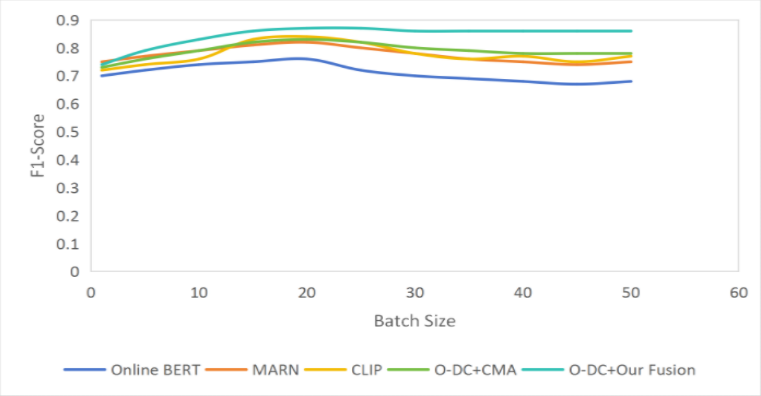

Figure 1. Online Streaming Performance Evolution.

The feature dimension dynamics chart compares the number of selected features over 50 batches among three online feature selection methods integrated with our fusion network. The proposed O-DC+Our Fusion method demonstrates a rapid decline in feature dimensions from approximately 900 to around 380 within the first 20 batches, after which it stabilizes at a low and consistent level in Figure 2. This indicates its ability to quickly identify and retain only the most informative features while discarding redundant or noisy ones. In contrast, OGL+Our Fusion exhibits larger fluctuations and maintains a higher average dimension (600), reflecting its group-wise sparsity constraints that are less adaptive to streaming dynamics. AI+Our Fusion initially selects a large number of features (1000) but gradually reduces to around 480, yet with noticeable variance. The efficiency advantage of the proposed method is evident: it achieves the highest F1-Score (87.2%) while using the fewest features, confirming that its online dynamic selection mechanism not only reduces computational overhead but also enhances feature quality, leading to better generalization in evolving data streams.

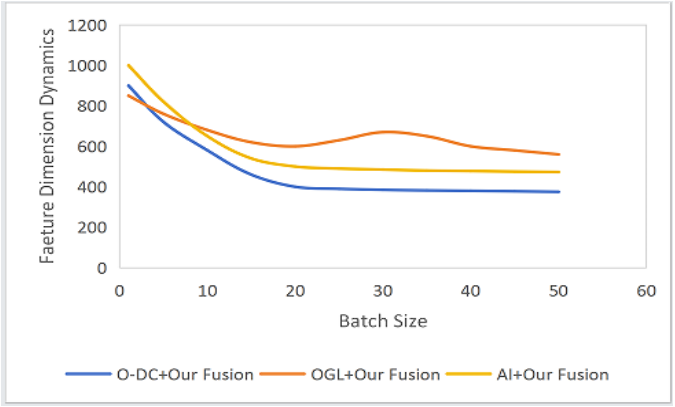

Figure 2. Feature Dimension Dynamics.

Figure 3 presents a grouped bar chart comparing the performance of eight models across Precision, Recall, F1-Score, and Accuracy. Several observations can be drawn. First, text-only baselines (Online SVM and Online BERT) exhibit the lowest overall performance, particularly in Recall (below 67%), indicating their limitation in capturing sufficient positive samples in streaming scenarios. Second, multimodal fusion methods (EF+OGL, CMA, MARN) significantly improve Recall, with MARN achieving the highest Recall (91.3%), but at the cost of reduced Precision (82.7%),

 doi:10.20944/preprints202606.0301.v1

revealing a precision-recall trade-off. Third, the CLIP (finetuned) model, leveraging large-scale vision-language pretraining, achieves strong performance (Precision 88.1%, F1-Score 85.9%, Accuracy 86.2%), outperforming traditional multimodal methods but still exhibiting a slight imbalance between precision and recall. Fourth, the O-DC+CMA combination improves Precision to 85.6% but suffers a sharp Recall drop to 71.8%, suggesting that attention-based cross-modal interaction alone cannot fully balance the two metrics. Most importantly, the proposed O-DC+Our Fusion model achieves the best overall performance, leading in Precision (89.3%), F1-Score (87.2%), and Accuracy (87.1%), while maintaining competitive Recall (85.3%). This balanced superiority demonstrates that the integration of online dynamic selection with our fusion network effectively resolves the precision-recall dilemma and outperforms even strong pretrained models like CLIP in the streaming multimodal setting.

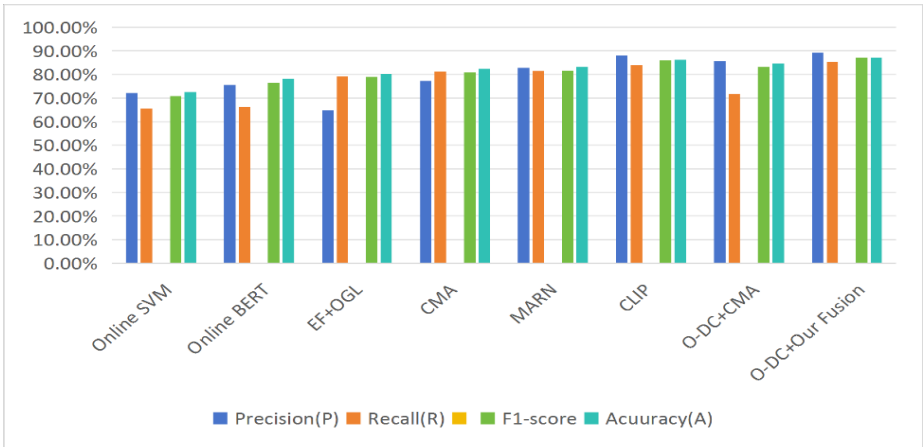

Figure 3. Performance Comparison of Different Models.

6.2.4. Ablation Study Analysis

To rigorously evaluate the contribution of each key component in our proposed online dynamic cross-modal fusion framework, we conduct a series of ablation experiments by systematically removing individual modules. The results are summarized in Table 5, where the Full Model achieves the best performance with 86.9% accuracy and 85.4% F1-Score, serving as the baseline for comparison. Each variant is described below, along with the corresponding performance degradation and efficiency impact.

Table 5. Ablation Study Results.

Model Variant	Accuracy (%)	ΔAcc	F1-Score (%)	ΔF1	Processing Time (ms)
Full Model	87.1	–	85.4	–	68
w/o O-DC (using all features)	84.1	-2.8	82.7	-2.7	82
w/o 2nd-stage Attention	85.6	-1.3	84.0	-1.4	64
w/o Adaptive Weighting	85.9	-1.0	84.3	-1.1	66
w/o Residual Connections	86.1	-0.8	84.7	-0.7	67

 doi:10.20944/preprints202606.0301.v1

w/o O-DC (using full features): This variant removes the online dynamic selection (O-DC) module and instead utilizes all original features (text 768-d + image 512-d) without any dimensionality reduction. Compared to the Full Model, accuracy drops by 2.8% to 84.1%, and F1-Score decreases by 2.7% to 82.7%. Notably, the processing time increases to 82 ms, which is 14 ms longer than the Full Model. This indicates that O-DC not only preserves discriminative information but also reduces computational overhead by filtering out redundant features, thereby enhancing both effectiveness and efficiency.

w/o 2nd-stage Attention: This variant eliminates the second-stage cross-modal attention mechanism, which is designed to refine feature interactions after initial fusion. The removal leads to an accuracy decline of 1.3% (85.6%) and a F1-Score drop of 1.4% (84.0%). The processing time slightly decreases to 64 ms, suggesting that the second-stage attention adds a modest computational cost while delivering noticeable performance gains. The results confirm that deeper attention-based refinement is beneficial for capturing complex modality relationships.

w/o Adaptive Weighting: Here, the adaptive weighting module that dynamically adjusts the contribution of each modality based on streaming context is removed. Accuracy falls to 85.9% (Δ = -1.0%), and F1-Score to 84.3% (Δ = -1.1%), with a processing time of 66 ms. The moderate performance drop underscores the importance of context-aware modality fusion, especially in evolving data streams where the reliability of text and visual signals may vary over time.

w/o Residual: This variant removes residual connections within the fusion network. Accuracy decreases by 0.8% to 86.1%, and F1-Score by 0.7% to 84.7%, while processing time remains similar to the Full Model (67 ms vs. 68 ms). Although the impact is relatively smaller, residual connections still contribute to stabilizing training and improving representational capacity, as evidenced by the consistent performance gain.

The ablation results clearly demonstrate that each module plays a vital role in the overall framework. The most significant performance degradation occurs when O-DC is removed, highlighting its critical function in online feature selection. The second-stage attention and adaptive weighting also contribute substantially to accuracy and F1-Score, while residual connections provide marginal yet consistent improvements. Moreover, the Full Model achieves the best trade-off between performance and efficiency, with processing time only slightly higher than the fastest variant but substantially lower than the feature-heavy w/o O-DC version. These findings validate the necessity and effectiveness of each designed component in our proposed method for online multimodal sentiment classification.

7. Conclusion

This study presented a novel online multimodal emotion recognition framework for public health emergency public opinion analysis. To address the challenges of streaming, high-dimensional, and heterogeneous data, we proposed: (1) an Online Divide-and-Conquer Markov Blanket Learning Algorithm for dynamic and efficient feature selection, and (2) a Cross-Modal Interactive Enhanced Fusion Network with two-stage attention and adaptive weighting for deep semantic fusion. Theoretical analysis provided guarantees on convergence, information preservation, and generalization. Comprehensive experiments on a large-scale PHE dataset demonstrated that our framework significantly outperforms state-of-the-art baselines in accuracy, F1-score, online learning efficiency, and robustness to concept drift, while maintaining interpretability.

Future work will focus on: (1) Extending the framework to incorporate more modalities, particularly video and audio; (2) Designing fully adaptive mechanisms for threshold parameters in O-DC; (3) Exploring more advanced architectures for handling complex, long-term concept drift.

References

 doi:10.20944/preprints202606.0301.v1

1. Che S. P.; Wang X.; Zhang S. N.; et al. Effect of daily new cases of COVID-19 on public sentiment and concern: Deep learning-based sentiment classification and semantic network analysis. *Journal of public health*, 2024,32(3):509-528. DOI:10.1007/s10389-023-01847-y.
2. Mumuni, A.; Mumuni, F. Data augmentation with automated machine learning: approaches and performance comparison with classical data augmentation methods. *Knowl. Inf. Syst*.,2025,67(5), 1-11.
3. Tang H.; Wang Y.; Zhang Y.; et al. TS-Mixer: A lightweight text representation model based on context awareness. *Expert Systems*, 2025,42(2). DOI:10.1111/exsy.13732.
4. Zhang, L.; Wang, X.; Wang, J.; Liao, G. Research on emergency decision quality evaluation and optimization basing on public sentiment big data analysis. *Comput. Ind. Eng*. ,2024,193(7), 109452.
5. Han, P.; Zhang, W.; Zhang, Z.;et al. Sentiment Analysis of Weibo Posts on Public Health Emergency with Feature Fusion and Multi-Channel. *Data Anal. Knowl. Discov.* ,2021, 5(11), 68-79.
6. Gariboldi, M.I.; Lin, V.; Bland, J.; et al. Foresight in the time of COVID-19. *Lancet Reg. Health West. Pac.*, 2021, 6, 100049.
7. Ahelegbey, D.F.; Celani, A.; Cerchiello, P. Measuring the impact of the EU health emergency response authority on the economic sectors and the public sentiment. *Socioecon. Plann.* ,2024,92, 101842.
8. Chen, X.; Zhang, W.; Xu, X.; Cao, W. A public and large-scale expert information fusion method and its application: Mining public opinion via sentiment analysis and measuring public dynamic reliability. *Inf. Fusion*,2022,78, 71–85.
9. An, L.; Xu, M. Measuring online trust in government microblogs in public health emergencies. *Data Anal. Knowl. Discov.* , 2022,6(1), 55–68.
10. Cai, Y.; Yang, K.; Huang, D.; Zhou, Z.; Lei, X.; Xie, H.; Wong, T. A hybrid model for opinion mining based on domain sentiment dictionary. *Int. J. Mach. Learn. Cybern.* ,2019, 10(8), 2131–2142.
11. Xu, G., Yu, Z., Yao, H., Li, F., Meng, Y., Wu, X., Chinese text sentiment analysis based on extended sentiment dictionary. IEEE Access, 2019,7, 43749–43762.
12. Yang, S.; Chen, F. Analyzing sentiments of Micro-blog posts based on support vector machine. *Data Anal. Knowl. Discov.*, 2017,1(2), 73–79.
13. Fan, H.; Li, P. Sentiment analysis of short text based on FastText word vector and bidirectional GRU recurrent neural network. *Information Science*, 2021,39(4), 15–22.
14. Liu, J.; Gu, F. Unbalanced text sentiment analysis of network public opinion based on BERT and BiLSTM hybrid method. *Journal of Intelligence*, 2022,41(04), 104–110.
15. Hyun, D.; Park, C.; Yang, M.; Song, I.; Lee, J.; Yu, H. Target-aware convolutional neural network for target-level sentiment analysis. *Information Science*,2019,491, 166–178.
16. Guo, X.; Zhao, N.; Cui, S. Consumer reviews sentiment analysis based on CNN-BILSTM. *Syst. Eng. Theory Pract.* ,2020,40, 653-663.
17. Lai, X.; Tang, H.; Chen, H.; Li, S.; Multimodal sentiment analysis based on feature fusion of attention mechanism-bidirectional gated recurrent unit. J. *Comput. Appl.*, 2021,41(5), 1268–1274.
18. Duan W. J.;Deng J. K.;Zhang S. X.; et al. Aspect-based sentiment analysis model based on multilevel knowledge enhancement. *CAAI Transactions on Intelligent Systems*,2024,19（5）:1287-1297.
19. Liu J. H.;Li l.;Wu R. W.; et al. Mutli-prompt learning based aspect-category sentiment analysis. *Journal of Frontiers of Computer Science and Technology*, 2025 ,19 (05) : 1334-1341.
20. Zhang W.; Chu Z. Y.; Chen X. Q.; et al. Aspect-based sentiment analysis with syntactic prompt. *Journal of Computer Applications*, 2024,44:35-43.
21. Li, Z.; Xu, H.; Duan, B. Research on image emotion feature extraction based on deep learning CNN model. *Library and Information Service*, 2019, 63(11), 96–107.
22. Cai, G.; He, X.; Chu, Y. Visual sentiment analysis by combining global and local regions of image. *Journal of Computer Applications*, 2019, 39(8), 2181–2185.
23. Alamoodi, A.H.; Zaidan, B.B.; Zaidan, A.A.; et al. Sentiment analysis and its applications in fighting COVID-19 and infectious diseases: A systematic review. *Expert Syst. Appl.*, 2021, 167, 114155.
24. Arbane, M.; Benlamri, R.; Brik, Y., Alahmar, A.D.Social media-based COVID-19 sentiment classification model using Bi-LSTM. *Expert Syst. Appl.* ,2023,212, 118710.

 doi:10.20944/preprints202606.0301.v1

25. Blanco, G.; Lourenço, A.Optimism and pessimism analysis using deep learning on COVID-19 related twitter conversations. *Inf. Process. Manag.* ,2022, 59(3), 102918.
26. Tan, H.; Peng, S.; Zhu, C.; et al. Long-term effects of the COVID-19 pandemic on public sentiments in mainland China: Sentiment analysis of social media posts. *J. Med. Internet Res.* , 2021,23(8), e29150.
27. Kumar, A.; Garg, G. Sentiment analysis of multimodal twitter data. Multimed. Tools Appl. ,2019,78(17), 24103–24119.
28. Gan Z. H.;Miao Y. Q.;Liu T. L.;et al. Multimodal aspect-level sentiment analysis based on cross-modal interaction Transformer. *Application Research of Computers*，2025,42(9):2707-2713.
29. Perkins S.; Theiler J. Online feature selection using grafting[C]. the Twentieth International Conference on Machine Learning, 2003: 592-599.
30. Ungar L. Streaming feature selection using alpha-investing[C]. the Eleventh International Conference on Knowledge Discovery and Data Mining, 2005: 384-393.
31. Wu X.; Yu K.; Ding W.; et al. Online feature selection with streaming features. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 2013, 35(5): 1178-1192.
32. Lin Y.; Hu Q.;Zhang J.;et al. Multi-label feature selection with streaming labels. *Information Sciences*, 2016, 372: 256-275.
33. Lin Y.; Hu Q.;Liu J.; et al. Streaming feature selection for multi-label learning based on fuzzy mutual information. *IEEE Transactions on Fuzzy Systems*,2017,25(6): 1491-1507.
34. Liu J.; Lin Y.; Li Y.; et al. Online multi-label streaming feature selection based on neighborhood rough set. *Pattern Recognition*, 2018, 84: 273-287.
35. Li H.;Wu X.; Li Z.; et al. Group feature selection with streaming features[C]. 13th International Conference on Data Mining, 2013: 1109-1114.
36. Jing W, Meng W, Li P, et al. Online feature selection with group structure analysis[J]. IEEE Transactions on Knowledge and Data Engineering, 2015, 27(11): 3029-3041.
37. Pearl J.; Mackenzie D. The book of why: the new science of cause and effect. *Science*, 2018,361(6405): 852-855.
38. Jake H.; Amit S.; Duncan W. Prediction and explanation in social systems. *Science*, 2017, 355(6324):486-488.
39. Gao T.; Ji Q. Efficient markov blanket discovery and its application. *IEEE transactions on cybernetics,* 2016, 47(5): 1169-1179.
40. Ling Z.; Yu K.; Wang H.; et al. Bamb: A balanced markov blanket discovery approach to feature selection. *ACM transactions on intelligent systems and technology (TIST)*,2019, 10(5): 1-25.
41. Wang H.; Ling Z.; Yu K.; et al. Towards efficient and effective discovery of markov blankets for feature selection. *Information sciences,* 2020, 509: 227-242.
42. Wu X.; Jiang B.; Yu K.; et al. Accurate markov boundary discovery for causal feature selection. *IEEE transactions on cybernetics,* 2019, 50(12): 4983-4996.
43. Liu C.; Wang Y.; Yang J. A transformer-encoder-basedmultimodalmulti-attention fusion network for sentiment analysis. *Applied Intelligence*,2024,54:8415–8441
44. Chen, X.; Zhang, W.; Xu, X.; Cao, W. A public and large-scale expert information fusion method and its application: Mining public opinion via sentiment analysis and measuring public dynamic reliability. *Inf. Fusion,* 2022,78, 71–85.
45. Wang, Y.; Xie, J.; Chen, B.; Xu, X. Multi-modal sentiment analysis based on cross-modal context-aware attention. *Data Anal. Knowl. Discov.* ,2021.,5(4), 49–59.
46. Harish A. B.; Sadat F.Trimodal Attention Module for Multimodal Sentiment Analysis. *Association for the Advancement of Artificial Intelligence (AAAI)*,2020:1-10.
47. Majumder N.; Hazarika D.; Gelbukh A.; ,et al. Multimodal sentiment analysis using hierarchical fusion with context modeling. *Knowledge Based Systems,* 2018,161, 124-133.
48. Sun H.;Chen Y W.;Lin L.Tensor Former: A Tensor-Based Multimodal Transformer for Multimodal Sentiment Analysis and Depression Detection.*IEEE Transactions on Affective Computing,*2023,14(4): 2776–2786.

 doi:10.20944/preprints202606.0301.v1

49. Tsai Y H H.;Bai S J.;Liang P P.; et al. Multimodal Transformer for Unaligned Multimodal Language Sequences. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019: 6558-6569.
50. Yang K C.; Xu H.; Gao K. CM-BERT: Cross-Modal BERT for Text-Audio Sentiment Analysis. Proceedings of the 28th ACM International Conference on Multimedia, 2020: 521-528.
51. Wang Yuzhu; Xie Jun; Chen Bo; et al. Multi-modal Sentiment Analysis Based on Cross modal Context-aware Attention. *Data Analysis and Knowledge Discovery,* 2021, 5(4): 49-59.
52. Chen Yansong; Zhang Le; Zhang Leihan;et al. Multimodal Sentiment Analysis Method Based on Cross-Modal Attention and Gated Unit Fusion Network. *Data Analysis and Knowledge Discovery*, 2024, 8(7): 67-76.
53. Yu Bengong; Shi Zhongyu. Deep Attention and Two-Stage Fusion of Image-Text Sentiment Contrastive Learning Method.*Computer Engineering and Applications*,2025,61(3):223-233.
54. Alahmadi, K.; Alharbi, S.; Wang, X. Integrating dense layers with residual connections into transformers for enhanced sentiment classification. *J. Supercomput.*,2025, 81, 1542.
55. He K.; Zhang X.; Ren S.; et al. Deep residual learning for image recognition. *Proceedings of* the IEEE conference on computer vision and pattern recognition. 2016: 770-778.
56. Islam M.S.; Xiangdong L.; Ahmed J. BERT: advancements in language understanding for different NLP tasks: challenges and future perspectives. *Journal of Electrical Systems and Inf Technol.,* 2026,13, 49.
57. Han Yanxiao; Ma Jing.The RCHFN Model: A Multimodal Feature Fusion Approach for Sentiment Classification. *Data Analysis and Knowledge Discovery*,2024,8(12):18-29.
58. Su Y Y.; Han C J.; Li A M.; et al. Research on Image-text Multimodal Sentiment Recognition Driven by Large Model Enhancement and Multi-feature Cross-fusion.*Information studies:Theory & Application*,2025,9:1-16.
59. Shahid S D.; Mohammad Z.; Karan B.; et al. A social context-aware graph-based multi-modal attentive learning framework for disaster content classification during emergencies. *Expert Systems with Applications*,2025,259,125337-125360.
60. Zeng Z M.; Sun S Q.; Li Q Q. Multimodal negative sentiment recognition of online public opinion on public health emergencies based on graph convolutional networks and ensemble learning. Information Processing & Management,2023,60(4):103378-103395.