Multimodal multi-teacher knowledge distillation with Keyword Attention for Health Mention Classification

Multimodal Multi-Teacher Knowledge Distillation with Keyword Attention for Health Mention Classification

Vishal Krishna Singh[a,*], Niharika Anand[b], Abhijit Mahendra Bhalekar[b], Rajkumar Singh Rathore[c]

[a] *School of Computer Science and Electronics Engineering, University of Essex, Colchester, United Kingdom*
[b] *Department of Computer Science, Indian Institute of Information Technology Lucknow, India*
[c] *School of Technologies, Cardiff Metropolitan University, United Kingdom*

Abstract

Health Mention Classification (HMC) aims to distinguish genuine health-related posts on social media from cases where disease or symptom terms are used figuratively, hyperbolically, or in non-personal contexts. Accurately interpreting such ambiguous expressions requires not only textual understanding but also contextual cues that often arise from conversational emotion and multimodal interactions. To address this challenge, this paper proposes a multimodal multi-teacher knowledge distillation (MTKD) framework for HMC on Reddit. The primary classification model is trained and evaluated on the Reddit Health Mention Dataset (RHMD), where a general-context BERT teacher and a domain-specific BioBERT teacher are independently fine-tuned, and their softened knowledge is distilled into a compact DistilBERT student. A lightweight Keyword Attention module further guides the student to focus on health-related lexical cues while preserving the surrounding contextual information required to distinguish literal and non-literal expressions. In addition, we leverage the MELD dataset as an auxiliary multimodal resource, where textual, acoustic, and visual conversational cues are used to train a multimodal teacher. The semantic and affective conversational knowledge learned by this teacher is transferred to the student through an auxiliary distillation

*Corresponding author
Email addresses: v.k.singh@essex.ac.uk (Vishal Krishna Singh), niharika@iiitl.ac.in (Niharika Anand), mcs24003@iiitl.ac.in (Abhijit Mahendra Bhalekar), rsrathore@cardiffmet.ac.uk (Rajkumar Singh Rathore)

objective. Consequently, the final HMC inference model remains entirely text-based, making it directly deployable on text-only social media platforms such as Reddit while benefiting from multimodal knowledge acquired during training. Experimental results on RHMD demonstrate that the proposed framework achieves an Accuracy of 0.8190, Precision-Macro of 0.8232, Recall-Macro of 0.8191, F1-Macro of 0.8203, and a TN_FM score of 0.8563, outperforming existing approaches. Additional evaluation on the SHMD dataset demonstrates strong robustness under distributional shifts. Despite benefiting from multiple teacher models during training, the final student model remains computationally efficient, requiring only 67.55M parameters with an average inference latency of 9.845 ms per sample.

Keywords: Health Mention Classification, Multi-Teacher Knowledge Distillation, Keyword Attention, DistilBERT, BioBERT, Multimodal Distillation, Reddit Health Mention Dataset

1. Introduction

The proliferation of social media platforms has transformed how individuals communicate and share personal experiences, including health-related issues [1]. The huge quantity of user-generated content, especially on forums such as Reddit characterized by a lack of user identities and longer post formats, has some potential value in public health surveillance [2]. This includes tracking infectious diseases, monitoring mental health, and detecting illicit drug use [3]. In addition to textual content, human communication in real-world settings is inherently multimodal, where cues such as tone, facial expressions, and conversational context play a critical role in interpreting meaning [4]. This is particularly important for distinguishing literal and figurative expressions, which are often ambiguous in text-only environments. At the center of utilizing this data is the Health Mention Classification (HMC) task [5]. HMC attempts to classify whether a text is a personal health mention - a text where the author, or someone they know, describes an experience with a health issue, or symptom [6].

Thus, social media HMC is crucial to ensure health-related data is accurately and reliably gathered from social media [7].

However, a major issue that makes HMC more complex is that the users often use disease or symptom keywords in a manner different from their use in referring to their own health issues [8]. Figurative language (for example, "This project is giving me a migraine"), hyperbole, or talk about news, advice, or general information is common [9]. Failure to account for these uses that are not literal or personal, and instead considering all instances of the keyword as actual health-related mentions, could introduce substantial bias into public health research using data [10]. Reddit, although providing more context due to longer posts, also has its own set of challenges with its varied communities and modes of communication [11]. Such ambiguity is further amplified by the absence of non-textual cues in social media text, which in natural conversational settings would otherwise help disambiguate intent through prosody, facial expressions, and interaction dynamics.

Research in HMC has examined various methods of computation. Initial methods used a combination of handcrafted linguistic features, lexicons, and classical ML models [12]. The incorporation of contextual embeddings from PLMs, and in some cases, attention layers resulted in higher performance. Some methods, like [13] for example, attempt to blur the target health term in order to facilitate a contextual focus. Other methods incorporate sentiment and/or user interaction. However, the problem of balancing general comprehension, the health-related discourse, and the complexity of a platform such as Reddit, in particular the ability to identify the literal and non-literal uses of terms, remains largely unaddressed. Recent advances in multimodal perception and pattern recognition suggest that contextual signals such as emotion, tone, and conversational structure - often learned from multimodal datasets like MELD - can provide complementary insights for interpreting ambiguous expressions [14]. While such signals are not directly available in text-only datasets like RHMD [15], they motivate the need for models that are robust to contextual ambiguity and capable of

approximating such cues through richer textual representations [16].

To address these challenges, we propose a novel framework based on Multi-Teacher Knowledge Distillation (MTKD). The focus of this model is to break down the pre-conceived perceptions of student models by leveraging distilled specializations from multiple teacher models. More importantly, we focus on two teacher models, both of which are fine-tuned on the HMC task.

1. A **general-context teacher (BERT)**. A teacher model (BERT) fine-tuned on general web corpora provides a strong grasp of general language and its constructs, which helps in understanding language patterns and semantics.
2. A **domain-specific teacher (BioBERT)**. A teacher model fine-tuned on extensive biomedical literature assists in providing contextual health-related jargon and its range of usage.

We believe that knowledge distilled into a compact student model (DistilBERT) can further refine the student's ability to accurately identify relevant pieces of information from potentially complex Reddit posts. To achieve this, we embed a lightweight Keyword Attention mechanism. This mechanism guides the student model to focus on health-related words of interest and consider the surrounding context during classification. The proposed framework uses RHMD as the primary supervised HMC dataset. Because RHMD does not contain paired audio or visual signals, the final HMC classifier performs text-only inference. To incorporate the multimodal theme without misrepresenting the RHMD task, we use MELD as an auxiliary multimodal resource during training: text, audio, and visual conversational cues are used to train a multimodal cue teacher, and its emotion-context knowledge is distilled into the student through an auxiliary head. Thus, the model is multimodal during representation learning while remaining practical for Reddit text deployment.

The primary contributions of this work are as below:

- We propose a MTKD framework tailored for HMC on Reddit, where a general-

context BERT teacher and a biomedical BioBERT teacher distill complementary knowledge into a compact DistilBERT student.

- We integrate a Keyword Attention mechanism inside the student model. This layer guides the classifier toward health-related lexical cues while preserving the surrounding context needed to distinguish literal, figurative, and non-personal usage.
- We introduce an auxiliary multimodal distillation component using MELD. A multimodal cue teacher is trained from MELD transcript, audio, and visual cues, and its emotion-context outputs are transferred to the student through an auxiliary distillation objective.
- We provide a detailed empirical analysis on RHMD, including class-wise results, confusion matrix analysis, TN_FM evaluation for figurative mention handling, ablation analysis, and computational efficiency comparison. The final model achieves an F1-Macro of 0.8203, Precision-Macro of 0.8232, Recall-Macro of 0.8191, and TN_FM of 0.8563.

2. Proposed Method: MTKD with Keyword Attention

2.1. Problem Definition

Given a Reddit post P_i, consisting of a sequence of tokens $P_i = \{t_1, t_2, \ldots, t_n\}$, where i denotes the post index and n is the number of tokens in the post, the objective is to classify P_i into one of k predefined categories $Y = \{y_1, y_2, \ldots, y_k\}$. These categories represent the nature of the health or symptom term usage within the post. In alignment with the evaluation scheme for the RHMD dataset and prior comparative work [13], we define $k = 3$. The labels correspond to:

Literal Non-Health Mention: The health/symptom term is used, but not in relation to a personal health experience (e.g., news, general discussion).

1. Figurative Non-Health Mention: The health/symptom term is used in a non-literal, figurative, or hyperbolic sense.
2. Health Mention: The post describes a personal health experience of the author or someone they know.

While the main classification objective is defined over textual Reddit posts, we additionally use MELD during an auxiliary multimodal distillation stage. This allows the student model to receive emotion-context supervision from conversational text, audio, and visual cues without requiring audio or video input during final RHMD inference.

2.2. Overview of Architecture

The proposed architecture is illustrated in Figure 1. It consists of two connected training pathways: (i) the RHMD-based HMC distillation pathway and (ii) the MELD-based auxiliary multimodal distillation pathway.

In the RHMD pathway, each Reddit post is cleaned, tokenized, and passed to two independently fine-tuned teacher models. The first teacher is BERT, which captures general linguistic and social-media-style contextual patterns. The second teacher is BioBERT, which contributes biomedical terminology and health-domain semantics. The logits produced by the two teachers are averaged and temperature-scaled to form soft HMC targets. The student is a compact DistilBERT model equipped with a Keyword Attention layer. The student learns from both the ground-truth HMC labels and the softened teacher distributions.

In the auxiliary MELD pathway, conversational utterances are represented using transcript embeddings, acoustic cues, and visual cues extracted from video clips, following prior work showing that multimodal language understanding benefits from modeling interactions across textual, acoustic, and visual streams [17]. These features train a multimodal cue teacher for emotion-context prediction. The softened output distribution of this teacher is then distilled into an auxiliary head of the student model.

This design enables the student to receive multimodal contextual supervision during training without requiring audio or visual input at RHMD inference time. Therefore, the final deployed HMC classifier remains text-only, while its representation learning is multimodal.

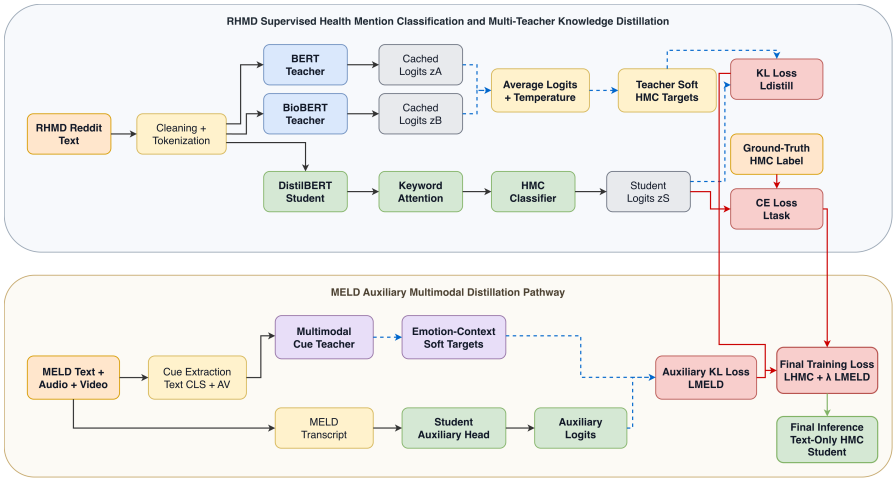

Figure 1: Architecture of the proposed MTKD framework with Keyword Attention and MELD-based auxiliary multimodal distillation.

The sequence of steps in the proposed methodology can be outlined as follows:

1. **Data Preparation:** RHMD posts are cleaned, de-identified, tokenized using model-specific tokenizers, and split into training, validation, and test sets. MELD utterances are separately processed to extract transcript, audio, and visual cues for auxiliary multimodal distillation.
2. **Teacher Training:** BERT and BioBERT are independently fine-tuned on the RHMD training set for the 3-way HMC task. Their logits are cached to accelerate student training.
3. **Student Training:** DistilBERT with Keyword Attention is trained using a weighted combination of ground-truth cross-entropy loss and teacher-student KL-divergence loss.
4. **Auxiliary Multimodal Distillation:** A multimodal cue teacher trained on

MELD provides emotion-context soft targets. These are distilled into a student auxiliary head and contribute an additional training loss.

5. **Evaluation:** The final trained student is evaluated on the unseen RHMD test set. Only text is required at inference time.

2.3. Teacher Models

We use two pre-trained language models as teachers to capture complementary linguistic knowledge. Each teacher is independently fine-tuned on the RHMD training set for the 3-way HMC task using a standard cross-entropy objective.

2.3.1. General Context Teacher (BERT)

We use *bert-base-uncased* as the general-context teacher [18]. BERT is pre-trained on large general-domain corpora and therefore captures broad semantic, syntactic, and pragmatic patterns. This is useful for Reddit HMC because many posts contain informal constructions, exaggerations, humor, and figurative expressions that require general language understanding beyond medical keywords alone.

2.3.2. Domain-Specific Teacher (BioBERT)

To include health-domain knowledge, we use the pre-trained BioBERT model *dmis-lab/biobert-v1.1* [19]. BioBERT is initialized from BERT and further pre-trained on biomedical corpora such as PubMed abstracts and PMC full-text articles. This additional domain-specific pre-training helps BioBERT encode disease names, symptom expressions, biomedical terminology, and clinical phrase patterns more effectively than a general-domain model. In the proposed MTKD framework, BioBERT complements BERT by contributing biomedical knowledge, while BERT contributes broader contextual understanding of informal Reddit language.

2.4. Student Model (DistilBERT with Keyword Attention)

2.4.1. Base Architecture (DistilBERT)

We use DistilBERT *(distilbert-base-uncased)* as the foundational architecture for our student model. DistilBERT is a distilled version of BERT, offering a significant reduction in size and computational cost while retaining a large proportion of BERT's performance, making it suitable for practical applications [20].

2.4.2. Keyword Attention Layer

To guide the student model to focus on salient health-related terms, we integrate a Keyword Attention mechanism over the DistilBERT encoder's hidden states. First, a predefined list of health and symptom keywords (e.g., "headache", "fever", "depression") is used to generate a binary 'keyword_mask' for each input sequence. Next, a dot-product attention mechanism [21] computes attention scores using the [CLS] token's hidden state as the query and the sequence of all token hidden states as keys and values. These scores are then masked using both the standard padding mask and the 'keyword_mask', setting non-keyword token scores to a large negative value before softmax normalization. Finally, a keyword-focused context vector c_{key} is computed as the weighted sum of the value vectors using the resulting attention probabilities α_j. This vector provides a representation specifically focused on the health-related terms within the post. The auxiliary MELD branch does not provide audio or visual input during RHMD inference, but it influences the student during training through auxiliary multimodal distillation.

2.4.3. Final Classification

The final representation for classification is formed by concatenating the standard [CLS] token's hidden state (h_{CLS}), which represents the global context of the post, with the keyword-focused context vector c_{key}. This combined vector $[h_{CLS}; c_{key}]$ is passed through a dropout layer and a linear classification layer. A softmax function

then produces probabilities over the three HMC classes: *Figurative Non-Health*, *Literal Non-Health*, and *Health*. The class with the highest probability is selected as the final prediction.

2.5. Knowledge Distillation Process

The student model is trained using both hard RHMD labels and softened probability distributions generated by the teacher ensemble. For each RHMD instance, the logits from BERT and BioBERT are averaged:

$$z^T_{avg} = \frac{z^A + z^B}{2}, \tag{1}$$

where z^A and z^B denote the output logits of the BERT and BioBERT teachers, respectively.

2.5.1. Task Loss (L_{task})

The task loss is the standard cross-entropy loss between the student logits and the ground-truth RHMD label:

$$L_{task} = -\sum_{i=1}^{N}\sum_{j=1}^{k} y_{ij}\log(p^S_{ij}), \tag{2}$$

where N is the batch size, $k = 3$ is the number of HMC classes, y_{ij} is the one-hot ground-truth label, and p^S_{ij} is the predicted probability of the student model.

2.5.2. Distillation Loss ($L_{distill}$)

To transfer teacher knowledge, temperature scaling is applied to both teacher and student logits:

$$q^T_j = \frac{\exp(z^T_{j,avg}/T)}{\sum_{l=1}^{k}\exp(z^T_{l,avg}/T)}, \quad q^S_j = \frac{\exp(z^S_j/T)}{\sum_{l=1}^{k}\exp(z^S_l/T)}. \tag{3}$$

The distillation loss is computed as the KL divergence between the teacher and student softened distributions:

$$L_{distill} = D_{KL}\left(\text{softmax}(z^T_{avg}/T) \parallel \text{softmax}(z^S/T)\right) \cdot T^2. \tag{4}$$

The HMC training loss is:

$$L_{HMC} = \alpha L_{task} + (1-\alpha) L_{distill}, \tag{5}$$

where $\alpha = 0.7$ in our experiments. Thus, the task loss receives weight 0.7, while the teacher-student distillation loss receives weight 0.3.

2.5.3. Auxiliary Multimodal Distillation Loss

MELD is not annotated for HMC; therefore, it is not used as a direct HMC training dataset. Instead, it is used to provide auxiliary multimodal context. For each MELD utterance, transcript embeddings, audio cues, and visual cues are used to train a multimodal cue teacher for emotion-context prediction. Let z^M be the output logits of this multimodal teacher and $z^{S,M}$ be the output of the student's auxiliary head for the corresponding MELD transcript. The auxiliary distillation loss is:

$$L_{MELD} = D_{KL}\left(\text{softmax}(z^M/T_M) \parallel \text{softmax}(z^{S,M}/T_M)\right) \cdot T_M^2. \tag{6}$$

The overall student training objective is:

$$L_{total} = L_{HMC} + \lambda L_{MELD}, \tag{7}$$

where $\lambda = 0.15$ controls the contribution of the auxiliary multimodal distillation objective.

Algorithm 1: Multi-Teacher Knowledge Distillation for Health Mention Classification

Input: RHMD training dataset D; BERT teacher M_G; BioBERT teacher M_D; DistilBERT student M_S with Keyword Attention; hyperparameters α, T, λ; tokenizers Tok_G, Tok_D, Tok_S.

Output: Trained student model M_S^*.

Phase 1: Teacher Model Training;

Split D into $D_{train}, D_{val}, D_{test}$ using stratification.;

foreach *teacher model* $M_T \in \{M_G, M_D\}$ *with tokenizer* Tok_T **do**

  Initialize optimizer O_T and scheduler Sch_T.;

  for $epoch = 1$ **to** $Epochs_T$ **do**

    M_T.train();

    foreach *batch* $(x_{batch}, y_{batch}) \in D_{train}$ *tokenized with* Tok_T **do**

      O_T.zero_grad();

      $\text{logits}_T \leftarrow M_T(x_{batch})$;

      $L_{CE} \leftarrow \texttt{CrossEntropyLoss}(\text{logits}_T, y_{batch})$;

      L_{CE}.backward();

      O_T.step(); Sch_T.step();

    Evaluate M_T on D_{val} and save the best checkpoint.;

  Cache teacher logits for D_{train}, D_{val}, and D_{test}.;

Phase 2: Student HMC Training;

Initialize optimizer O_S and scheduler Sch_S.;

for $epoch = 1$ **to** $Epochs_S$ **do**

  M_S.train();

  foreach *batch* $(x_{batch}, y_{batch}) \in D_{train}$ *tokenized with* Tok_S **do**

    Load cached BERT and BioBERT logits for the batch.;

    $\text{logits}_{avg} \leftarrow (\text{logits}_G + \text{logits}_D)/2$;

    $P_{teacher} \leftarrow \texttt{Softmax}(\text{logits}_{avg}/T)$;

    $\text{logits}_S \leftarrow M_S(x_{batch})$;

    $\log P_{student} \leftarrow \texttt{LogSoftmax}(\text{logits}_S/T)$;

    $L_{task} \leftarrow \texttt{CrossEntropyLoss}(\text{logits}_S, y_{batch})$;

    $L_{distill} \leftarrow \texttt{KLDivLoss}(\log P_{student}, P_{teacher}) \cdot T^2$;

    $L_{HMC} \leftarrow \alpha L_{task} + (1 - \alpha) L_{distill}$;

    O_S.zero_grad(); L_{HMC}.backward(); O_S.step(); Sch_S.step();

  Evaluate M_S on D_{val} using F1-Macro and save the best checkpoint.;

return *Final trained student model* M_S^*;

Algorithm 2: Auxiliary Multimodal Distillation using MELD

Input: MELD dataset $D_M = \{(u_j, a_j, v_j, e_j)\}_{j=1}^{M}$, where u_j is transcript text, a_j is audio, v_j is video, and e_j is the emotion label; student model M_S; temperature T_M; auxiliary weight λ.

Output: Student model updated with auxiliary multimodal contextual supervision.

Extract transcript embeddings from MELD utterances.;

Extract lightweight acoustic cues from audio clips, including MFCC, zero-crossing rate, RMS energy, and spectral centroid.;

Extract lightweight visual cues from sampled video frames, including color, brightness, and frame-difference motion statistics.;

Train a multimodal cue teacher M_M on MELD emotion labels using transcript, audio, and visual features.;

Cache softened multimodal teacher logits $P_M = \texttt{Softmax}(z^M / T_M)$.;

for $epoch = 1$ **to** $Epochs_S$ **do**

  foreach *MELD batch* (u_{batch}, P_M) **do**

    Encode transcript u_{batch} using the student text encoder.;

    Compute auxiliary student logits $z^{S,M}$ through the student auxiliary head.;

    $\log P_{S,M} \leftarrow \texttt{LogSoftmax}(z^{S,M} / T_M)$;

    $L_{MELD} \leftarrow \texttt{KLDivLoss}(\log P_{S,M}, P_M) \cdot T_M^2$;

    Update student parameters using λL_{MELD} in addition to the RHMD HMC loss.;

return *Multimodal student model.*

3. System Model: Real-World Application Scenario

To better convey the usefulness and implications of the proposed MTKD model with Keyword Attention, we describe a conceptual system model for the proposed public health surveillance system. This conceptual system model describes the larger pipeline in which the HMC framework is incorporated to process raw social media data into health insights.

3.1. Overview of the Public Health Surveillance System

The conceptual system model, shown in Figure 2, is a continuous pipeline system that aims to monitor public health trends and identify emerging issues from Reddit

data. The overall goal is to filter health-related discussions from the massive amount of user-generated content on social media platforms, properly classifying them according to their nature (literal, figurative, or non-personal) to guarantee the validity of the extracted information.

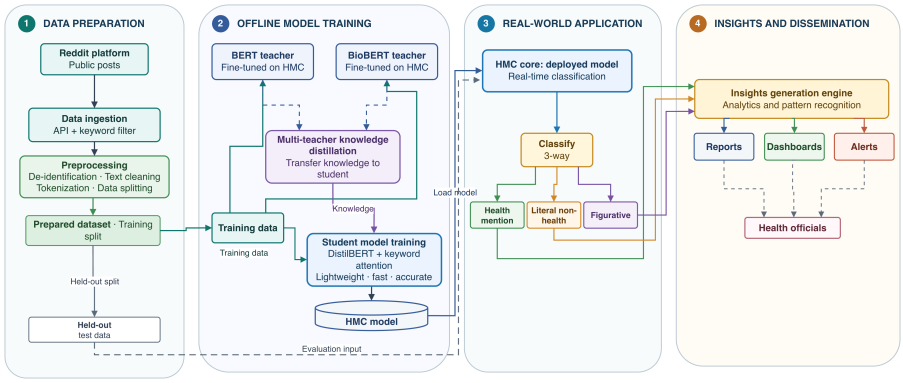

Figure 2: System Model for HMC in Public Health Surveillance.

3.2. System Components and Workflow

3.2.1. Data Ingestion and Preprocessing

The system starts with continuous data ingestion from Reddit, usually done through APIs that enable keyword-based retrieval of health-related posts and comments. After ingestion, raw text data are subjected to a comprehensive preprocessing pipeline. This includes de-identification (substitution of usernames, URLs, and other potentially private information with anonymized tokens), text cleaning (converting to lowercase, translating emojis to text, and normalizing whitespaces), and tokenization using specialized language model tokenizers.

3.2.2. Health Mention Classification Core

Categorization of each of the posts into any of the three health categories, Health Mention, Literal Non-Health Mention, and Figurative Non-Health Mention, is the HMC core responsibility. This is important to ensure that our model puts each post

under the right health category. To illustrate this, it would be more accurate and desirable to categorize the post such as, “This exam gave me a heart attack”, as a Figurative Non-Health Mention rather than categorizing it as a Health Mention. This prevents a false positive and is able to use the post in the analysis. The student model Keyword Attention layer also helps in making this analysis easier as it prioritizes the appropriate health-related terms that appear in the post and thereby minimizes ambiguity. The auxiliary MELD distillation stage provides additional contextual supervision during training, which may help the student better distinguish figurative expressions from genuine health mentions.

3.2.3. Insights Generation and Aggregation

After categorization, the data is processed to generate actionable insights. Filtering removes non-health mentions, particularly figurative ones, ensuring only actual personal health reports enter the surveillance pipeline. Categorization further groups the data by specific diseases or symptoms. Finally, Aggregation and Trend Analysis, groups the mentions by time, demographics, or geographic regions. This allows public health organizations to monitor irregular spikes in symptoms, track the proliferation of health concerns, and analyze health event-related sentiments [22].

3.2.4. Reporting and Alerting

The findings are then reported for policy making. For instance, a large number of correctly classified “Health Mention" messages about a particular symptom in a particular region may prompt an alert for a possible localized outbreak. On the other hand, a large number of “Figurative Non-Health" messages about a health crisis may suggest public concern or a need for better communication.

4. Experiments

4.1. Dataset

We conduct the primary HMC experiments on the publicly available Reddit Health Mention Dataset (RHMD) introduced by Naseem et al. [15]. The RHMD dataset consists of $10,015$ manually annotated Reddit posts, each containing at least one of 15 common disease or symptom terms. The original annotations include four labels: Personal Health Mention (PHM), Non-Personal Health Mention (NPHM), Figurative Health Mention (FHM), and Hyperbolic Health Mention (HBHM).

For our experiments, and to allow direct comparison with the 2 + 1 evaluation scheme used by prior RHMD work [13], we transform these four labels into a 3-way classification task. The original NPHM class is mapped to *Literal Non-Health Mention*. The FHM and HBHM classes are combined into *Figurative Non-Health Mention*, because both correspond to non-literal uses of health-related terms. The PHM class is mapped to *Health Mention*. The final labels used in this paper are therefore:

- **Literal Non-Health Mention:** a health or symptom term occurs in the post, but the post does not describe a personal health experience.
- **Figurative Non-Health Mention:** a health or symptom term is used metaphorically, hyperbolically, or idiomatically.
- **Health Mention:** the post describes a personal or closely related health experience.

Table 1: Label Distribution for the RHMD Dataset in the 3-Way Classification Scheme

RHMD Dataset: 3-Way Classification Scheme	
Class	**Number of Posts**
Figurative Non-Health Mention	3430
Literal Non-Health Mention	3225
Health Mention	3360
Total Posts	**10015**

We use a stratified split of the RHMD dataset, resulting in 7010 training samples, 1502 validation samples, and 1503 test samples. The class distribution is preserved across the splits to avoid evaluation bias caused by class imbalance. To incorporate multimodal contextual information, we additionally use the publicly available MELD dataset [23] as an auxiliary resource. MELD contains conversational utterances with text, audio, and visual information, along with emotion labels. Since MELD is not annotated for HMC, it is not used for direct HMC supervision or RHMD test evaluation. Instead, it is used to train a multimodal cue teacher whose softened emotion-context outputs are distilled into the student model through an auxiliary training objective.

Table 2: Auxiliary MELD Configuration Used for Multimodal Distillation.

Auxiliary MELD Multimodal Configuration	
Item	**Value**
MELD utterances used	999
Emotion categories	7
Audio window	3 seconds
Video frames sampled per clip	3
Audio feature dimension	32
Visual feature dimension	9
Multimodal cue teacher best validation F1-Macro	0.2973
Auxiliary distillation weight λ	0.15

4.2. Experimental Settings

4.2.1. Preprocessing

The Reddit posts are processed before being passed to the teacher and student models:

- **Text Cleaning:** We convert posts to lowercase, replace URLs and usernames with generic placeholders, convert emojis into textual descriptions when present, normalize extra whitespace, and preserve negation terms such as “not”, “no”, and “never”, because they are essential for contextual interpretation.

- **Tokenization:** The BERT, BioBERT, and DistilBERT models use their respective pre-trained tokenizers from the Hugging Face Transformers library [24].

Model-specific tokenization is used to avoid vocabulary mismatch between BERT and BioBERT.

- **Padding and Truncation:** Input sequences are padded or truncated to a maximum length of 256 tokens.
- **Auxiliary Multimodal Cue Extraction:** For MELD, we extract transcript embeddings, acoustic cues, and visual cues. The acoustic representation includes MFCC statistics, zero-crossing rate, RMS energy, and spectral centroid. The visual representation is computed from sampled video frames using color, brightness, and frame-difference motion statistics. These cues are used only for the auxiliary multimodal distillation objective and are not required at final RHMD inference time.

4.2.2. Model Parameters and Training Details

The hyperparameters and training details for our experiments were held constant for fair comparison. The important parameters for the teacher and student models are listed in Table 3. Both models were trained with the AdamW optimizer with an epsilon of 1×10^{-8} and a linear learning rate scheduler with zero warmup steps. All experiments were performed using PyTorch [25] on NVIDIA *T*4 GPUs. The random seed was set to 42. For faster training, the BERT and BioBERT teacher logits were precomputed and cached before student training.

4.3. Evaluation Metrics

To evaluate our proposed model, we use standard classification metrics: Accuracy, Macro-averaged Precision, Macro-averaged Recall, and Macro-averaged F1-score. We also compute the Weighted F1-score to account for class imbalance. These metrics are calculated using scikit-learn [26] with 'zero_division = 0'.

To specifically assess the model's ability to correctly handle non-literal (figurative and hyperbolic) mentions, we define the TN_FM (True Negatives for Figurative

Table 3: Hyperparameter Settings for Teacher, Student, and Auxiliary Multimodal Training.

Model and Training Hyperparameter Configuration		
Parameter	**Teacher Models**	**Student Model**
Base Model	BERT [18] / BioBERT [19]	DistilBERT [20]
Checkpoint	`bert-base-uncased` `dmis-lab/biobert-v1.1`	`distilbert-base-uncased`
Training Details		
Learning Rate	2×10^{-5}	5×10^{-5}
Batch Size	16	16
Epochs	4	5
Max Seq. Length	256	256
Optimizer	AdamW	AdamW
Scheduler	Linear	Linear
Knowledge Distillation		
Temperature (T)	N/A	2.0
Task Loss Weight (α)	N/A	0.7
Distillation Loss Weight ($1 - \alpha$)	N/A	0.3
MELD Auxiliary Weight (λ)	N/A	0.15

α denotes the weight assigned to the RHMD cross-entropy task loss, while $1 - \alpha$ denotes the weight assigned to the RHMD teacher–student distillation loss. The MELD auxiliary weight λ is used only during auxiliary multimodal distillation.

Mentions) metric. Assuming C_{FNH} is the set of true "Figurative Non-Health" posts, and C_{NH} is the set of posts predicted as either "Literal Non-Health" or "Figurative Non-Health":

$$\text{TN\_FM} = \frac{|\{p \in C_{FNH} \mid p \in C_{NH}\}|}{|C_{FNH}|} \tag{8}$$

Because the improvement over the strongest reported RHMD baseline is modest, we report the comparison conservatively and do not mark statistical significance in the final baseline table. The auxiliary MELD component is evaluated as a multimodal distillation resource rather than as a direct HMC benchmark, since MELD does not contain RHMD-style HMC labels.

4.4. Baselines

We compare our proposed MTKD model with Keyword Attention against the following:

- **HMCNET (2+1 Scheme):** The results reported by Naseem et al. [13] for their proposed HMCNET model when evaluated on the RHMD dataset using

Table 4: Comparative Analysis with Baseline Models on the RHMD Dataset (3-Way / 2+1 Scheme). Our proposed model results are from the test set.

Comparison with Existing HMC Models			
Model	**F1-Score**	**Precision**	**Recall**
BiLSTM-Senti [29]	0.70	0.70	0.70
HMCNET [15]	0.78	0.77	0.78
WESPAD [27]	0.59	0.60	0.60
JiangLSTM [28]	0.63	0.63	0.63
FeatAug+ [12]	0.51	0.52	0.51
MASK-Net [13]	0.81	0.81	0.81
Proposed MTKD + MELD Aux. (Ours)	**0.8203**	**0.8232**	**0.8191**
Underline indicates the best-performing literature baseline model.			

the 3-way (2+1) classification scheme (Table 6 in the mentioned paper). This represents the most relevant state-of-the-art for direct comparison on this specific task and dataset.

- **Single Teacher Models (Ablation):** As part of our ablation study, we also report the performance of:
  - Fine-tuned BERT (our General Context Teacher) used as a standalone classifier.
  - Fine-tuned BioBERT (our Domain-Specific Teacher) used as a standalone classifier.
- **Student Model without KD (Ablation):** DistilBERT with Keyword Attention trained solely on the ground-truth labels without knowledge distillation.

The MASK-Net paper [13] also compares against other baselines like WESPAD [27], FeatAug+ [12], JiangLSTM [28], and BiLSTM-Senti [29] on the RHMD 2+1 scheme. Our primary comparison is with their proposed HMCNET model, which outperformed these other baselines in their study.

5. Results

This section presents the performance of the proposed MTKD model with Keyword Attention and MELD-based auxiliary multimodal distillation on the RHMD dataset.

We compare our approach with established baselines and analyze its effectiveness in overall HMC, class-wise classification, figurative mention handling, ablation behavior, and computational efficiency. The RHMD test set remains the final benchmark for HMC evaluation.

5.1. Overall Comparison with Baselines

Table 4 summarizes the overall performance of the proposed MTKD model against several baseline methods on the RHMD dataset using the 3-way (2 + 1) classification scheme. The final student model achieves an F1-Macro score of 0.8203, Precision-Macro of 0.8232, and Recall-Macro of 0.8191. The improvement over MASK-Net is modest in absolute F1-Macro terms; therefore, we do not interpret the result as a large margin gain. Instead, the result shows that the proposed framework remains competitive with the strongest RHMD baseline while using a compact student at inference time and adding explicit keyword-guided and multimodal training signals.

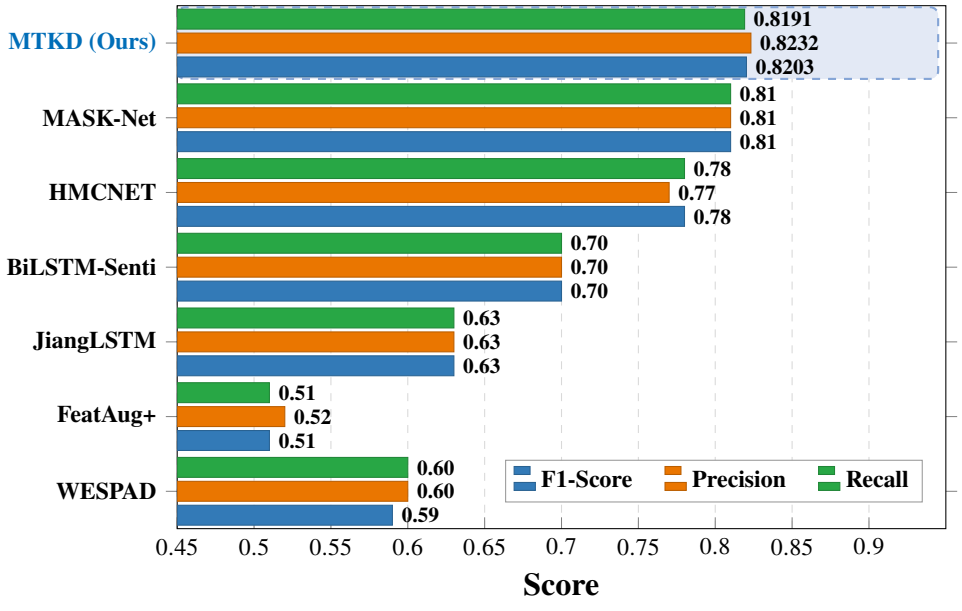

Figure 3: Comparative Analysis with Baseline Models on RHMD Dataset.

5.2. Figurative Language Identification (TN_FM)

One of the main difficulties in HMC is the proper detection of the figurative or hyperbolic employment of health-related and symptom-related terms, to avoid their misclassification as actual personal health-related posts. We assess this using the TN_FM metric, which calculates the proportion of actual "Figurative Non-Health" mentions that are properly classified as any type of non-health.

The proposed MTKD model with Keyword Attention and MELD auxiliary distillation obtains a TN_FM score of 0.8563 (441 out of 515 actual figurative mentions correctly classified as non-health) on the RHMD test set. This is slightly higher than the TN_FHM+HBHM score of 0.85 reported for MASK-Net [13]. Although the improvement is modest, it is important because false interpretation of figurative health expressions as genuine health mentions can negatively affect downstream monitoring systems. The result indicates that the combination of teacher distillation and keyword-guided contextual representation helps preserve sensitivity to figurative expressions.

5.3. Auxiliary Multimodal Distillation using MELD

To address multimodal context more directly, MELD is used as an auxiliary multimodal distillation resource rather than only as a post-hoc qualitative dataset. We use a stratified subset of 999 MELD utterances containing seven emotion categories: anger, disgust, fear, joy, neutral, sadness, and surprise. For each utterance, we extract transcript-based text embeddings, acoustic cues from the audio stream, and visual cues from sampled video frames. These features are used to train a multimodal cue teacher for emotion-context prediction. The teacher's softened output distributions are then distilled into the student through an auxiliary head.

This auxiliary component is not treated as direct HMC supervision because MELD does not contain HMC labels. Instead, it provides an external multimodal context signal that is useful for learning conversational patterns such as emotional tone, surprise,

sarcasm, or affective emphasis. At inference time on Reddit posts, the model requires only text input.

5.4. Class-Specific Analysis

The performance of the proposed method is analyzed on metrics such as precision, recall, and F1-score for each of the three classes on the RHMD test set. The results as presented in Table 5, show that the model achieves F1-scores of 0.8050, 0.8556, and 0.8004 for the Figurative Non-Health, Literal Non-Health, and Health classes, respectively.

Table 5: Classification Report for the Proposed MTKD Model on the RHMD Test Set.

RHMD Classification Performance				
Class	**Prec.**	**Recall**	**F1**	**Support**
Figurative NH	0.8004	0.8097	0.8050	515
Literal NH	0.8941	0.8202	0.8556	484
Health	0.7751	0.8274	0.8004	504
Accuracy			**0.8190**	1503
Macro Avg	**0.8232**	**0.8191**	**0.8203**	1503
Weighted Avg	0.8224	0.8190	0.8198	1503

Table 6: Confusion Matrix of the Proposed MTKD Model on the RHMD Test Set.

RHMD Confusion Matrix			
True Label	**Predicted Label**		
	Fig. NH	**Lit. NH**	**Health**
Figurative NH	**417**	24	74
Literal NH	40	**397**	47
Health	64	23	**417**

The confusion matrix, as shown in Table 6, further reveals that the largest remaining errors occur between "Figurative Non-Health" and "Health" classes. This is expected because both categories often contain explicit health terms, and the model must rely on surrounding context to determine whether the mention is literal or figurative.

6. Analysis and Discussion

6.1. Robustness Analysis on RHMD

6.1.1. Granular Per-Class Metrics

In addition to standard precision and recall, we computed specificity, false positive rate (FPR), and false negative rate (FNR) for each class on the RHMD test set (Table 7). The model maintains high specificity for all classes, especially for Literal Non-Health mentions. The Figurative Non-Health class has an FNR of 0.1903, showing that figurative expressions remain challenging but are handled with reasonable reliability.

6.1.2. *Performance on Challenging Data Subsets*

We further assessed model performance on subsets of the RHMD test data grouped by negation presence and sentiment polarity (using VADER [30]). As shown in Table 8, performance remains stable regardless of negation. The model achieves its strongest F1-Macro (0.8396) on positive sentiment posts, which frequently contain figurative expressions, while maintaining robust performance (0.7938) on negative sentiment posts. This suggests the model captures deeper contextual relationships rather than relying on superficial linguistic cues.

Table 7: Per-Class Metrics on the RHMD Test Set.

RHMD Per-Class Analysis			
Class	**Specificity**	**FPR**	**FNR**
Figurative NH	0.8947	0.1053	0.1903
Literal NH	0.9539	0.0461	0.1798
Health	0.8789	0.1211	0.1726

Table 8: Performance across RHMD Test-Set Slices.

RHMD Data-Slice Analysis			
Data Slice	**Accuracy**	**F1-Macro**	**Support**
With Negation	0.8232	0.8051	311
Without Negation	0.8161	0.8116	1191
Positive Sentiment	0.8422	0.8396	564
Neutral Sentiment	0.8314	0.8108	261
Negative Sentiment	0.7917	0.7938	677

6.1.3. *Model Confidence and Calibration*

To assess prediction confidence, we analyzed the model's output probabilities. The average confidence for correct predictions is 0.9509, compared to 0.8467 for incorrect predictions, indicating that the model tends to be more confident when it makes correct decisions. The calibration curves shown in Fig. 4 further demonstrate that the predicted probabilities closely track the ideal diagonal, suggesting that the model is reasonably well calibrated across all three classes.

The calibration curves in Fig. 4 show whether the model's predicted probabilities are aligned with empirical correctness. This is important for health-related monitoring because a model should not only classify accurately but also express reliable confidence. Well-calibrated predictions allow downstream systems to set safer alert thresholds and defer uncertain cases for human review.

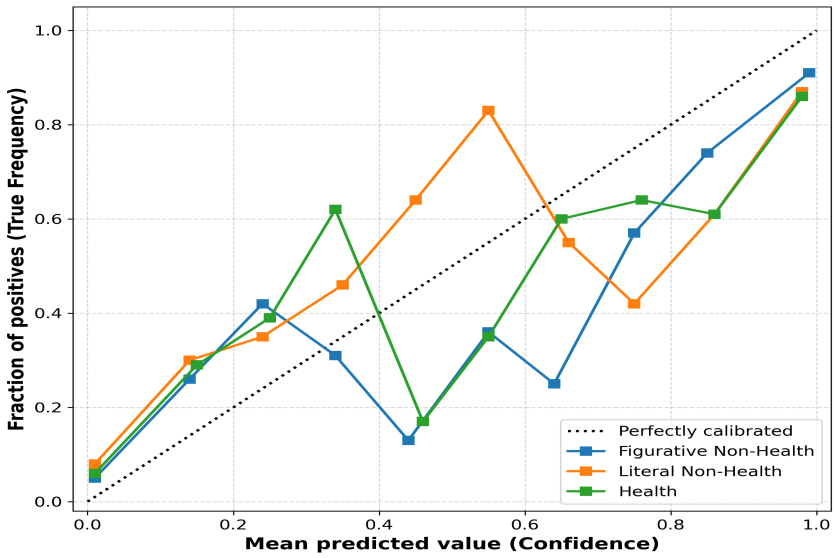

Figure 4: Calibration curves for each class on the RHMD test set. Curves closer to the diagonal indicate better calibration.

6.2. Ablation Analysis

To better understand the contribution of individual components, we conduct an ablation study by systematically removing key elements of the proposed MTKD framework.

Table 9: Ablation Study on the RHMD Test Set.

Ablation Analysis of the Proposed Framework	
Model Variant	**F1-Macro**
Full MTKD + Keyword Attention + MELD Auxiliary Distillation	**0.8203**
MTKD + Keyword Attention without MELD Auxiliary Distillation	0.8203
w/o Keyword Attention	0.8091
Single Teacher (BERT only)	0.8047
Single Teacher (BioBERT only)	0.8073

The results indicate that both multi-teacher distillation and keyword attention contribute to overall performance. Removing Keyword Attention reduces F1-Macro from 0.8203 to 0.8091, a drop of 0.0112, which shows that health-keyword-guided attention helps contextual disambiguation. The MELD auxiliary distillation stage does not materially change the aggregate RHMD F1-Macro in this run, but it provides an explicit

multimodal training pathway and addresses conversational context without requiring audio or video input at final Reddit inference time.

6.3. Computational Efficiency Analysis

To justify the additional training complexity introduced by the teacher models, we measure parameter size and inference latency. The teacher models are used only during training and logit caching. During deployment, only the compact DistilBERT student with Keyword Attention is used. Although MTKD requires training two large teachers, the deployed model is substantially smaller than either teacher. The final student has 67.55M parameters, compared with 109.48M for BERT and 108.31M for BioBERT, and its measured inference latency is 9.845 ms/sample.

Table 10: Computational Efficiency Comparison.

Computational Efficiency Analysis			
Model	**Params. (M)**	**Latency (ms)**	**Inference**
DistilBERT + Keyword Attention	**67.55**	**9.845**	**Yes**
BERT Teacher	109.48	18.429	No
BioBERT Teacher	108.31	18.442	No

Note: Teacher models are used only during training.

Table 11: Overall Performance on the SHMD Dataset.

SHMD Performance	
Metric	**Score**
Accuracy	0.8278
F1-Macro	**0.8266**
F1-Weighted	0.8266
TN_FM	0.7870

6.4. Generalization to a Synthetic Dataset (SHMD)

As a supplementary robustness analysis, we evaluate the trained HMC student on the unseen, synthetically generated SHMD dataset. The overall results, presented in Table 11, show strong generalization performance, with an F1-Macro score of 0.8266.

These results suggest that the model retains reasonable performance under the synthetic distribution shift introduced by SHMD. Since SHMD is a synthetic robustness set rather than the primary benchmark, these results are used only as supplementary evidence of generalization.(Table 12) and the corresponding granular error analysis (Table 13), offering further evidence of the model's robustness and generalizability. At the same time, the lower TN_FM score on SHMD suggests that synthetic figurative

expressions can be more challenging than those observed in RHMD, reinforcing the importance of contextual disambiguation.

Table 12: Classification Report for the SHMD Dataset.

SHMD Classification Performance				
Class	**Prec.**	**Rec.**	**F1**	**Supp.**
Figurative NH	0.8517	0.7682	0.8078	4000
Literal NH	0.8451	0.9263	0.8838	4000
Health	0.7874	0.7890	0.7882	4000
Macro Avg	**0.8281**	**0.8278**	**0.8266**	**12000**
Weighted Avg	0.8281	0.8278	0.8266	12000

Table 13: Granular Metrics on SHMD.

SHMD Per-Class Analysis			
Class	**Spec.**	**FPR**	**FNR**
Figurative NH	0.9331	0.0669	0.2318
Literal NH	0.9151	0.0849	0.0737
Health	0.8935	0.1065	0.2110

The confusion matrix in Fig. 5 visualizes these results. The model's excellent performance on "Literal Non-Health" (F1 = 0.8838) is confirmed by its very low False Negative Rate (FNR) of 0.0737. However, the TN_FM score drops to 0.7870. The granular metrics in Table 13 reveal why: the FNR for "Figurative Non-Health" is 0.2318, meaning the model failed to identify over 23% of true figurative posts. The confusion matrix confirms these were predominantly misclassified as "Health" (852 instances). This suggests the synthetic figurative patterns were more challenging than those in the original training data.

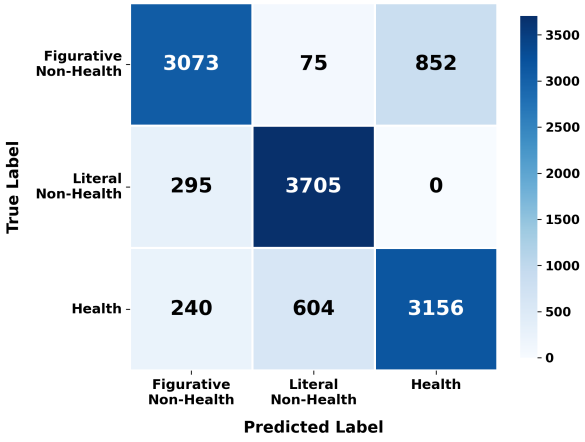

Figure 5: Confusion Matrix of the Model on the Synthetic SHMD Dataset.

Finally, the calibration curves on the SHMD dataset (Fig. 6) show how the model's

confidence behaves on this new data distribution. Similar to the RHMD analysis, the model remains well-calibrated on the synthetic dataset, indicating that its confidence scores are reliable even on unseen data distributions.

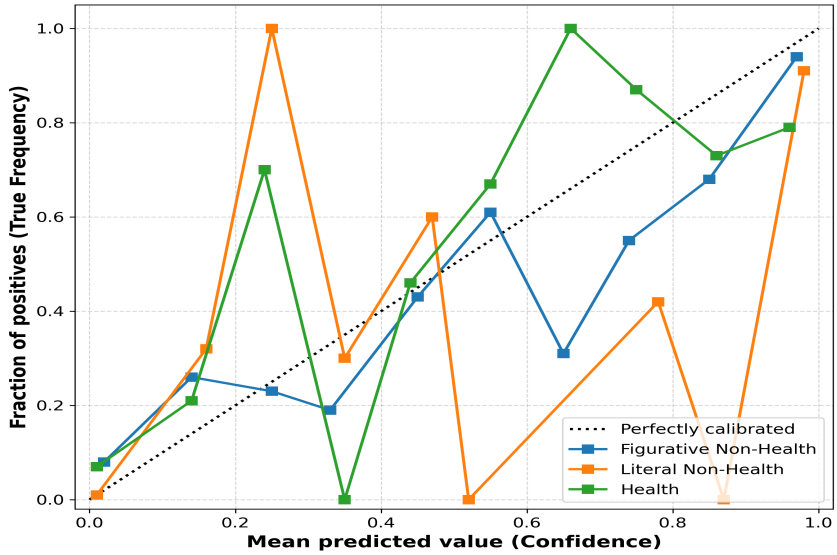

Figure 6: Calibration Curves for Each Class on the Synthetic SHMD Dataset.

6.5. Auxiliary MELD Distillation Results

MELD is used as an auxiliary multimodal resource and not as a direct HMC evaluation dataset. In the optimized run, we used 999 stratified MELD utterances with seven emotion labels. The multimodal cue teacher was trained using transcript embeddings together with audio and visual cue features. Its best validation F1-Macro was 0.2973. This value should not be compared with RHMD HMC scores because the label spaces and tasks are different. Its purpose is to provide softened emotion-context supervision that exposes the student model to multimodal conversational patterns during training. The auxiliary MELD branch therefore supports the multimodal aspect of the proposed framework while preserving the practical deployment setting of Reddit HMC, where only text is available at inference time.

6.6. Error Analysis

Table 6 shows that the most important remaining error source is the confusion between *Figurative Non-Health* and *Health*. Specifically, 74 true *Figurative Non-Health* posts are incorrectly classified as *Health*, while 64 true *Health* posts are misclassified as *Figurative Non-Health*. These errors arise because both categories can contain the same disease or symptom terms, and the distinction depends on whether the surrounding context signals a genuine personal health experience or a non-literal expression.

The model also makes fewer errors between *Literal Non-Health* and *Health*, with 47 true *Literal Non-Health* posts predicted as *Health* and 23 true *Health* posts predicted as *Literal Non-Health*. This indicates that general health discussions and personal health narratives remain challenging, but the Keyword Attention and multi-teacher distillation help reduce some of these errors. The TN_FM score of 0.8563 further shows that most figurative mentions are still correctly treated as non-health, although figurative-to-health errors remain an important direction for future improvement.

7. Conclusion

This paper addresses HMC on Reddit, with particular emphasis on distinguishing genuine personal health mentions from literal non-personal mentions and figurative or hyperbolic uses of disease and symptom terms. We proposed a multimodal MTKD framework in which BERT and BioBERT act as complementary teachers for a compact DistilBERT student with Keyword Attention. The BERT teacher contributes general contextual understanding, while BioBERT contributes biomedical-domain knowledge. The Keyword Attention layer helps the student focus on health-related terms while still using their surrounding context for disambiguation. Additionally, MELD is incorporated as an auxiliary multimodal resource. Transcript, audio, and visual cues from MELD were used to train a multimodal cue teacher, whose softened emotion-context outputs were distilled into the student through an auxiliary head. The final HMC in-

ference remains text-based, which matches the RHMD and Reddit deployment setting, but the training process benefits from auxiliary multimodal supervision.

Experiments on RHMD show that the proposed method achieves an Accuracy of 0.8190, F1-Macro of 0.8203, Precision-Macro of 0.8232, Recall-Macro of 0.8191, and TN_FM of 0.8563. Although the aggregate F1 improvement over prior RHMD baselines is modest, the model remains competitive while using a smaller student at inference time. The efficiency analysis shows that the deployed student uses 67.55M parameters and 9.845 ms/sample inference latency, whereas the larger BERT and BioBERT teachers are required only during training. Future work will focus on stronger multimodal alignment between social-media text and audio-visual conversational cues, larger multimodal auxiliary datasets, and improved handling of difficult figurative-to-health confusion cases.

Declarations

Conflict of Interest	The authors declare no competing interests.
Funding	No specific funding was received for this work.
Data Availability	Data and code are available from the corresponding author upon reasonable request.
Author Contributions	All authors contributed equally, approved the final manuscript, and agree with its submission.
Ethics Approval	Not applicable.
Consent to Participate	Not applicable.
Consent for Publication	All authors consent to publication.
Submission Declaration	This manuscript is original, has not been published previously, is not under consideration elsewhere, and is not an extension of previously published work.
AI Declaration	Grammarly was used solely for grammar and spelling correction. No generative AI was used for content generation, data analysis, or scientific interpretation.
Acknowledgments	The authors acknowledge the publicly available datasets used in this study.

References

[1] H. Lin, L. Li, Y. Zhao, Y. Zhao, Personalityllm: A zero-shot detection method for big five personality traits from social avatars, IEEE Transactions on Computational Social Systems 13 (4) (2026) 4295–4312. doi:10.1109/TCSS.2026.3694667.

[2] P. I. Khan, I. Razzak, A. Dengel, S. Ahmed, Performance comparison of transformer-based models on twitter health mention classification, IEEE Transactions on Computational Social Systems 10 (3) (2023) 1140–1149. doi:10.1109/TCSS.2022.3143768.

[3] K. Lan, C. L. P. Chen, Z. Zhang, T. Zhang, Contrastive adversarial tuning: Enhancing discriminability and robustness of llms for emotion recognition in conversation, Pattern Recognition 174 (2026) 113025. doi:https://doi.org/10.1016/j.patcog.2025.113025.

[4] V. K. Singh, J. Singh, R. Singh Rathore, D. Nema, Gradient-based compression and approximate computing for deep network optimization in multimedia data processing, IEEE Transactions on Consumer Electronics 72 (1) (2026) 1699–1710. doi:10.1109/TCE.2026.3656716.

[5] P. I. Khan, S. A. Siddiqui, I. Razzak, A. Dengel, S. Ahmed, Improving health mention classification of social media content using contrastive adversarial training, IEEE Access 10 (2022) 87900–87910. doi:10.1109/ACCESS.2022.3200159.

[6] R. John, V. S. Anoop, S. Asharaf, Health mention classification from user-generated reviews using machine learning techniques, in: S. Anwar, A. Ullah, Á. Rocha, M. J. Sousa (Eds.), Proceedings of International Conference on Information Technology and Applications, Springer Nature Singapore, Singapore, 2023, pp. 175–188.

[7] R. Ullah, S. Akbar, R. Nawaz, Z. Ali, V. K. Singh, S. A. C. Bukhari, Digital biomarkers for early detection of alzheimer's disease: A comprehensive review and bibliometric analysis, Journal of Dementia and Alzheimer039;s Disease 3 (2) (2026).

[8] M. B. Abisado, R. L. Rodriguez, M. L. G. Bautista, M. F. Pacis, B. S. Fabito, Healthph: A mobile and web-based multilingual ai platform for respiratory disease surveillance using geospatial social media analytics (2025) 78–81doi:10.1109/ICCRI67201.2025.11262460.

[9] C. Ren, J. Chen, R. Li, Y. Chen, T. Wang, W. Zheng, X. Zhang, B. Hu, Hierarchical feature distillation model via dual-stage projections and graph embedding label propagation for emotion recognition, Pattern Recognition 171 (2026) 112143. doi:https://doi.org/10.1016/j.patcog.2025.112143.

[10] Y. Zou, X. Hu, P. Li, Y. Wu, J. Hu, Online multi-label classification under noisy and changing label distribution, Pattern Recognition 173 (2026) 112892. doi:https://doi.org/10.1016/j.patcog.2025.112892.

[11] X. Li, J. Mao, H. Shi, L. Chen, Fine-grained evaluation for offensive speech detection on social media, Pattern Recognition 174 (2026) 113000. doi:https://doi.org/10.1016/j.patcog.2025.113000.

[12] A. Iyer, A. Joshi, S. Karimi, R. Sparks, C. Paris, Figurative usage detection of symptom words to improve personal health mention detection, in: A. Korhonen, D. Traum, L. Màrquez (Eds.), Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 1142–1147. doi:10.18653/v1/P19-1108.

[13] U. Naseem, S. Thapa, Q. Zhang, J. Rashid, L. Hu, Mask-net: Robust health mention classification by masking a disease or symptom terms, IEEE Trans-

actions on Computational Social Systems 12 (4) (2025) 1607–1616. doi:10.1109/TCSS.2024.3492143.

[14] S. Dong, J. Wei, J. Zhao, Y. Zhu, J. Zhou, Z. Shao, M. Niu, X. Tan, Y. Jiang, R. Qin, Cmtnet: A collaborative mamba-transformer network with spatial-temporal cross-fusion for speech emotion recognition, Pattern Recognition 175 (2026) 113159. doi:https://doi.org/10.1016/j.patcog.2026.113159.

[15] U. Naseem, J. Kim, M. Khushi, A. G. Dunn, Identification of disease or symptom terms in reddit to improve health mention classification, in: Proceedings of the ACM Web Conference 2022, WWW '22, Association for Computing Machinery, New York, NY, USA, 2022, pp. 2573–2581. doi:10.1145/3485447.3512129.

[16] T. Baltrušaitis, C. Ahuja, L.-P. Morency, Multimodal machine learning: A survey and taxonomy, IEEE Transactions on Pattern Analysis and Machine Intelligence 41 (2) (2019) 423–443. doi:10.1109/TPAMI.2018.2798607.

[17] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, R. Salakhutdinov, Multimodal transformer for unaligned multimodal language sequences, in: A. Korhonen, D. Traum, L. Màrquez (Eds.), Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 6558–6569. doi:10.18653/v1/P19-1656.

[18] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: J. Burstein, C. Doran, T. Solorio (Eds.), Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. doi:10.18653/v1/N19-1423.

[19] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, J. Kang, Biobert: a pre-trained biomedical language representation model for biomedical text mining, Bioinformatics 36 (4) (2020) 1234–1240. doi:10.1093/bioinformatics/btz682.

[20] V.Sanh, Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter., Proceedings of Thirty-third Conference on Neural Information Processing Systems (NIPS2019) (2019).

[21] C. Subakan, M. Ravanelli, S. Cornell, M. Bronzi, J. Zhong, Attention is all you need in speech separation, in: ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2021, pp. 21–25. doi:10.1109/ICASSP39728.2021.9413901.

[22] Y. Mejova, M. Tizzani, Vaccine hesitancy on youtube: A competition between health and politics (2025) 1–8.doi:10.1109/DPH66411.2025.11198076.

[23] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, R. Mihalcea, MELD: A multimodal multi-party dataset for emotion recognition in conversations, in: A. Korhonen, D. Traum, L. Màrquez (Eds.), Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 527–536. doi:10.18653/v1/P19-1050.

[24] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Brew, Transformers: State-of-the-art natural language processing, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Association for Computational Linguistics, Online, 2020, pp. 38–45. doi:10.18653/v1/2020.emnlp-demos.6.

[25] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al., Pytorch: An imperative style, high-performance deep learning library, Vol. 32, 2019.

[26] P. Gupta, A. Bagchi, Essentials of python for artificial intelligence and machine learning (2024). doi:10.1007/978-3-031-43725-0.

[27] P. Karisani, E. Agichtein, Did you really just have a heart attack? towards robust detection of personal health mentions in social media, in: Proceedings of the 2018 World Wide Web Conference, WWW '18, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 2018, p. 137–146. doi:10.1145/3178876.3186055.

[28] K. Jiang, S. Feng, Q. Song, R. A. Calix, M. Gupta, G. R. Bernard, Identifying tweets of personal health experience through word embedding and LSTM neural network, BMC Bioinformatics 19 (Suppl 8) (2018) 67–74. doi:10.1186/s12859-018-2198-y.

[29] R. Biddle, A. Joshi, S. Liu, C. Paris, G. Xu, Leveraging sentiment distributions to distinguish figurative from literal health reports on twitter, in: Proceedings of The Web Conference 2020, WWW '20, Association for Computing Machinery, New York, NY, USA, 2020, p. 1217–1227. doi:10.1145/3366423.3380198.

[30] C. J. Hutto, E. Gilbert, VADER: A parsimonious rule-based model for sentiment analysis of social media text, in: Proceedings of the International AAAI Conference on Web and Social Media, Vol. 8, 2014, pp. 216–225. doi:10.1609/icwsm.v8i1.14550.