Decoding Hostility from Conversations through Speech and Text Integration

Jingyu Xin[1], Brooks Gump[1], Stephen Maisto[1], Randall Jorgensen[1], Tej Bhatia[1], Vir Phoha[1], and Asif Salekin[2]

[1]{jxin05, bbgump, samaisto, rsjorgen, tkbhatia, vvphoha}@syr.edu
[2]asif.salekin@asu.edu
[1]Syracuse University
[2]Arizona State University

***Abstract*—Hostility is a complex trait with emotional, cognitive, and behavioral components. Hostility identification in conversational or transactional scenarios can benefit healthcare systems by, for example, predicting cardiovascular disease risks. While conventional hostility assessment relies on interviews, training proficient interviewers and mitigating biases pose significant challenges. In response, this study introduces the GMGF-MIL method to pioneer automatic multi-modal hostility detection in a structured interview. This approach utilizes recurrent neural networks to capture conversational context while integrating a graph neural network-based technique to merge acoustic and textual data. Furthermore, attention-based multiple-instance learning pooling is employed to aggregate utterance-level information. Notably, this is the first paper to introduce a novel multi-modal automated conversational hostility assessment approach, filling a notable gap in existing resources. Our evaluations showcase the efficacy of the GMGF-MIL method, achieving an accuracy of 78% in distinguishing between high- and low-hostile individuals.**

***Index Terms*—Hostility Detection, Multi-Modal Fusion, Acoustic Analysis, Text Analysis, Deep Learning**

I. Introduction

Behavioral hostility refers to an antagonistic interpersonal attitude comprising cognitive, affective, and behavioral elements [1]. The identification of individuals with high hostility holds significant implications in healthcare, notably in patient-provider interactions [2], [3], mental health assessments [4]–[6], and cardiovascular disease (CVD) risk [7]–[10]. Hostility assessment typically involves self-report evaluations [11] and interview-based methods [12]. Due to the transactional nature of hostility, the interview method offers advantages by assessing both recalled instances of hostile behaviors and real-time hostile behaviors in a pseudo-naturalistic environment [13]. Since hostility often emerges in social interactions due to vigilance for others' hostility and perceptions of threat from the interviewer, the Structured Type A Behavioral Interview [14], [15] is a gold standard for assessing transactional hostility, which has been linked to heart diseases [16], [17].

During interviews, hostility is often expressed through negative words and aggressive tones, making both textual transcriptions and audio signals valuable for detection. While interview-based approaches are more reliable, they face challenges such as the need for experienced interviewers, potential interviewer bias, and the time-consuming nature of interview administration and coding. To overcome the limitations of traditional interview-based methods, this study develops an automated approach to identify hostility based on conversational utterances. This makes transactional hostility screening automated, scalable, and cost-effective (e.g., using a voice agent to replace interviewers), enabling efficient assessment of a non-traditional CVD risk factor. Here, the term "utterance" is used to describe spoken statement information, thus comprising both transcribed text and acoustic modalities.

Compared to the existing works that primarily focus on detecting hostile or hateful contents [18]–[20] or predicting stable personality traits [21]–[23], which are not established as proxies of transactional hostility in literature, this paper is *the first to explore how to distinguish individuals high vs. low transactional hostility by integrating text and audio throughout the conversation of gold standard Structured Type A Interview with designed questions.*

To address the research problem, this paper introduces the GMGF-MIL, a novel neural network model comprising three key components: (1) bi-directional gated units (BiGRU) serving as contextual information encoders. Here, the "context" of an utterance encompasses the target individual's (whose hostility is being assessed) acoustic style, choice of words, and affective states of the individuals (i.e., interviewer and interviewee) involved in the utterance; (2) a graph neural network-based fusion mechanism for inter- and intra-modality information integration; and (3) attention-based multiple instance learning pooling layers designed to effectively identify and aggregate important information from multiple utterances. We summarize the key contributions as follows:

- This study pioneers the automatic identification of individuals with high vs. low transactional hostility in conversational interviews, utilizing speech acoustic and transcription data. By leveraging the presented approach and replacing structured interview questions with a voice agent, the interview-based hostility detection process can be fully automated, overcoming the above-discussed limitations of human interviewers.
- This study introduces GMGF-MIL, a method that can

This work was partly supported by NSF IIS SCH #2124285, NSF CNS CPS #2148187, and NIH NIDCD #1R01DC020959-01.
Asif Salekin is the corresponding author.

2156-8111/24/$31.00 ©2024 IEEE

track contexts in conversation, integrate different modalities, and effectively aggregate information from utterances with varying importance in prediction from conversation comprising several multi-person utterances. The novelty comes from formulating the classification as a multiple instance learning (MIL) problem. MIL enhances the significance of utterances closely associated with hostility, thus amplifying their influence in the final embedding used for classification.
- The paper offers insights and interpretations regarding the questions/topics that elicit distinguishable patterns among individuals with low vs. high hostility by visualizing the MIL weights, which can be advantageous for future studies aiming to assess hostility from conversations.

II. Related Work

This section first discusses the transactional model of hostility, differentiating our work from studies focused on identifying hostile content and predicting personality traits. Then, it introduces state-of-the-art (SOTA) approaches for the related task of emotion recognition in conversations.

A. Transaction Model of Hostility

The transactional model of hostility explains how hostile individuals interpret and react to social interactions [7], [13]. It suggests that hostility involves more than just responding to daily stressors with increased and prolonged cardiovascular and neuroendocrine reactivity; Instead, hostile individuals, through their thoughts and actions, create more frequent, severe, and enduring contacts with stressors, leading to cardiovascular reactivity within the context of interpersonal interactions [24], [25].

a) Transactional Hostility vs. Content-Based Hostility: Hostile, toxic, and hateful speech or posts are reflections of hostility [26]. Studies have aimed to detect hostile or hateful content across various mediums, including images [27], [28], textual posts [18], [19], [29], and videos [20], [30] on social media platforms. However, these efforts predominantly focus on content analysis and provide limited insights into individuals' hostility in a transactional environment and its health implications.

b) Transactional Hostility vs. Big 5 Personality Traits: The Big Five Personality Traits include Open-mindedness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism, describing individuals' stable patterns of behavior, cognition, and emotion across various contexts [31]. Studies have explored automatic personality recognition based on these traits using multiple modalities from dyadic interactions [21]–[23], [32], [33]. While hostility is likely related to the Big Five model, the transactional model of hostility specifically examines the dynamics of hostile responses in social interactions, such as in structured Type A interviews. In contrast, the Big Five model offers a broader overview of stable personality characteristics across various contexts.

Compared to content-based hostility detection and personality prediction, this study focuses on identifying transactional hostility in structured interview conversations. By developing an automatic detection method in this setting, we aim to pave the way for future research in this field.

B. Multi-Modal Emotion Recognition in Conversation

Studies on emotion recognition in conversation (ERC) commonly integrate information from multiple modalities, such as text, audio, video, etc., to extract complementary semantic information [34]. The common approaches for modality integration include early fusion that concatenates initial feature vectors from various modalities [35], [36], late fusion that concatenates embeddings of different modalities generated from independent embedding generating neural networks [37], [38], and utterance-level interaction fusion that explicitly models the relationships between modalities [39]–[41].

Acknowledging the sequential nature of conversational utterances, studies underscore the mutual influence of utterance contexts [42], [43]. Models capable of encoding sequential information, such as recurrent neural networks and transformers, have been widely employed [44]–[46]. Moreover, graph neural networks have emerged as promising tools for encoding speaker interactions and sentiment implications within and across modalities [47]–[54].

While emotion is a component of hostility, this study differs from ERC with a distinct scenario. ERC involves detecting the emotion conveyed in each utterance, whereas our objective is to detect individuals with high hostility from others based on patterns exhibited across all utterances within an interview conversation.

III. Problem Formulation and Data

There are two individuals in an interview: the interviewer, who poses questions, and the interviewee, who responds. The i-th utterance is denoted as u_i and incorporates two modalities - transcribed text and acoustic signal. The feature representation of an utterance is expressed as $u_i = \{u_i^a, u_i^t\}$, where u_i^a and u_i^t represent the feature vectors of the acoustic and textual modalities, respectively. We arrange the utterances in the temporal order they are spoken, modeling the interview conversation as $U = [u_1, u_2, \ldots, u_n]$, where n is the number of utterances. An utterance is spoken either by the interviewer or the interviewee, therefore, we denote that

$$U_Q = \{u_i | u_i \in U, u_i \text{ is spoken by the interviewer}\} \quad (1)$$

$$U_A = \{u_j | u_j \in U, u_j \text{ is spoken by the interviewee}\} \quad (2)$$

where U_Q and U_A are the sets of utterances from the interviewer and interviewee respectively.

The target of this study is to infer whether the interviewee exhibits high hostility (positive class, i.e., class 1) or low hostility (negative class, i.e., class 0) based on the utterances U exchanged during the conversation.

A. Dataset Overview

To the best of our knowledge, there is no large public multi-modal dataset that contains the label of hostility measure for the individual in interviews or conversations. Thus, we

processed a dataset comprising more than 3000 utterances from 91 interview conversations that occurred as part of the Multiple Risk Factor Intervention Trial (MRFIT) [55]. The interview was designed by Chesney *et al.* [15], and the interviewee's hostile behaviors were assessed using the Interpersonal Hostility Assessment Technique (IHAT) [12]. IHAT evaluates hostility by assessing the interviewee's irritated tones, impatience, and implicit or explicit responses indicating disdain for the questions asked, among other indicators [12], [15].

Some questions (along with their follow-up questions) in the interview intentionally pose challenges or interruptions, which are more likely to elicit hostile reactions from the interviewee. We selected a total of 12 questions-answers (suggested as highly effective in triggering hostility by [12], [15]), each of which may be followed by several additional questions-answers. These questions are arranged in the temporal order they are asked during the interview, as presented in Table I.

TABLE I

A LIST OF SELECTED INTERVIEW QUESTIONS. *The interviewers are trained to ask the questions in a standardized manner to present the subject with a stimulus that will educe hostile behavior if it is part of the subject's behavioral repertoire. The capitalized words are emphasized with a crisp, abrupt, staccato style* [15].

1. Are you SATISFIED with your job level? Does your job carry HEAVY responsibility?
2. Are there times that you feel RUSHED or under PRESSURE? When you ARE under PRESSURE does it bother you?
3. How would your WIFE (CLOSE FRIEND) describe you in those terms - as HARD-DRIVING and AMBITIOUS or as relaxed and easy-going?
4. When you are ANGRY or UPSET, do people around you know about it? How do you show it?
5. When you play competitive games with children, do you ALWAYS let them WIN on PURPOSE? Why or why not?
6. If a car in your lane is going FAR TOO SLOWLY for you, would you MUTTER and COMPLAIN to yourself? What do you do about it?
7. Most people who work have to get up fairly early in the morning, in your particular case, uh-what-time-uh-do-you-uh, ordinarily uh-uh-uh-get up?
8. If you make a DATE with someone for, oh, two o'clock in the afternoon, would you BE THERE on TIME? If you are kept waiting, do you RESENT it?
9. If you see someone doing a job rather SLOWLY and you KNOW you could do it faster and better yourself, does it make you RESTLESS to watch them? Would you be tempted to STEP IN AND DO IT yourself?
10. What IRRITATES you most about your work, or the people with whom you work?
11. When you go out in the evening to a restaurant and you find eight or ten people WAITING AHEAD OF YOU for a table, will you wait?
12. How do you feel about WAITING in lines - bank lines, supermarket lines?

Typically, a higher IHAT score indicates greater hostility. To set a threshold for positive and negative classes and avoid bias in our small subset, we use the median IHAT score (0.06) reported by a previous study [8] with a larger population of 518 individuals. This categorizes the interviewees into high-(positive) and low-hostile (negative) classes, comprising 46 and 45 individuals, respectively, as depicted in Fig. 1.

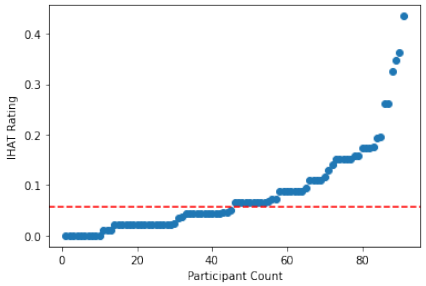

Fig. 1. Each point in this figure represents one interviewee and his/her IHAT rating in our data. The red dashed line is the median score (0.06) reported in [8]; we use it to assign hostility labels to interviewees. The ones above the red line are high-hostile (positive), while the ones below the red line are low-hostile (negative).

B. Data Processing and Feature Extraction

The raw acoustic data for each interview is provided as a single .wav file, and the time stamps for individual utterances are unavailable. To identify the specific questions outlined in Table I, we employed the Google Speech-to-Text service [56] to perform speaker diarization on each interview recording, thereby segmenting the entire conversation into multiple smaller clips. Subsequently, the transcript for each clip was generated using the same API. By comparing the automatically generated transcripts with the accurate transcripts, we extracted the audio clips containing the utterances corresponding to the questions and answers listed in Table I from all interview recordings. Further, the acoustic data is paired with the textual data.

As described in Section III, each utterance is represented as $u_i = \{u_i^a, u_i^t\}$, comprising acoustic and textual information. Hostility is expressed through both verbal and paraverbal components (involving words and vocal stylistics) [15], sharing connections with language and vocal expressions of sentiment and emotion [57]. Following the related work on ERC [52], [58], [59], we employed pre-trained FastText word embeddings [59] and OpenSmile toolkit [60] with IS10 configuration [61] to represent the textual (u_i^t) and acoustic (u_i^a) modalities.

IV. METHODOLOGY

A. Challenges and Design Choices

This section rationalizes our design choices for detecting hostility in conversations, which includes contextual information encoding, modality fusion, and pooling of utterance embeddings.

1) Encoding of Contextual Information.: The first challenge is *"How to track the contexts in the conversation?"*

In conversations, the interpretation of an utterance can be influenced by its surrounding context [62]. Hostile behavior often emerges in social interactions, such as through the *vocal style* and *choice of words* employed by the interviewee in response to the interviewer's questions. As an individual's hostility can be influenced by others' behavior, the *affective information of both individuals* involved in the respective context

(i.e., utterance) is crucial for assessing hostility. Therefore, it is essential to implement a mechanism capable of capturing the contextual information and corresponding speakers' affective states within the interview conversation. Thus, we utilize the one bi-directional gated unit (BiGRU), which is known for its high capability in interpreting sequential data [63] to encode the sequential contextual information from textual modality. Compared to other sequential models like Long Short-Term Memory (LSTM) networks and transformers, GRU has fewer parameters, making it a more suitable choice for our small dataset. Furthermore, another shared BiGRU network is employed to track the speakers' states for better affective information detection affiliated with the respective context. The details are discussed in Section IV-B1.

2) Fusion of Acoustic and Textual Modalities.: The second challenge is *"How to effectively fuse the two modalities?"*

Text and audio modalities exhibit distinct patterns yet are inherently interconnected. Integrating complementary semantic information across modalities can enhance the model's understanding capability [64]–[66]. Given the extensive studies and promising results of graph-based multi-modal fusion techniques in emotion and sentiment analysis during conversations [49], [51]–[54], [67], we opt for a graph-based fusion approach to merge textual and acoustic information. The details about graph construction and fusion are given in Section IV-B2.

3) Aggregation of Utterances via Pooling.: The third challenge is *"How to consolidate information from a fluctuating number of utterances, each with differing degrees of importance, to infer hostility?"*

This study targets identifying the interviewee's high vs. low hostility based on multiple utterances, and the number of utterances varies among interviews. Additionally, as outlined in prior work, not all questions are equally effective in eliciting hostility patterns [12]. Therefore, employing a simple mean pooling over the embeddings of all utterances could diminish the representation capability of those utterances that are closely associated with hostility. To enhance the aggregation of information from multiple utterances regardless of the utterance number, we formulate hostility identification into a multiple instance learning (MIL) problem. By deploying the attention-based MIL pooling method [68], the varying number of utterances can be handled, and the information from the most indicative utterances can be amplified. Further discussion is provided in Section IV-B3.

B. GMGF-MIL Approach

To address the challenges, we design **G**RU-based **M**ulti-modal dynamic **G**raph **F**usion MIL (GMGF-MIL) approach. Different modules of GMGF-MIL are discussed in the following sections. Fig 2 shows the overall architecture of the method.

1) Contextual Information Encoding with BiGRU: Considering sentences from speech transcriptions are more structured and contain richer contextual information than audio data, we only encode contexts within the textual modality. This approach allows for a more nuanced understanding of utterances

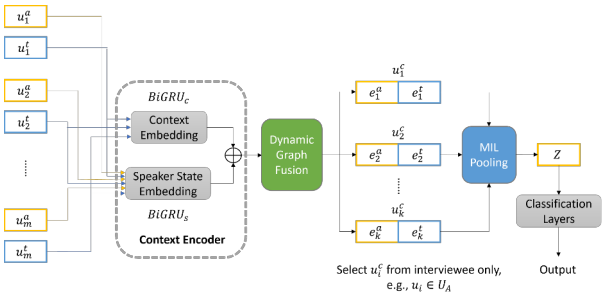

Fig. 2. GMGF-MIL Model Architecture

in linguistics. Given a bi-directional gated unit ($BiGRU_c$) for context tracking, the context embedding c_i^t for u_i^t is computed as:

$$c_i^t, h_i^c = BiGRU_c(u_i^t, h_{i-1}^c) \tag{3}$$

where h_i^c is the hidden state. Additionally, a prior study [44] indicates that the individual speakers' states during a conversation are informative for affect information detection. Thus, we use a shared $BiGRU_s$ to encode the speakers' states from both modalities:

$$s_{i,Q}^m, h_i^s = BiGRU_s(u_i^m, h_{i-1}^s), m \in \{a, t\}, u_i \in U_Q \tag{4}$$

$$s_{j,A}^m, h_j^s = BiGRU_s(u_j^m, h_{j-1}^s), m \in \{a, t\}, u_j \in U_A \tag{5}$$

where $s_{i,Q}$ and $s_{j,A}$ are the speaker state embedding for the interviewer and interviewee, respectively. Based on the context embedding and speaker state embedding, we define:

$$x_i^t = c_i^t + s_i^t \qquad x_i^a = u_i^a + s_i^a \tag{6}$$

Both the new textual (x_i^t) and acoustic (x_i^a) embeddings comprise each speaker's affective state information from respective modalities, where x_i^t comprises contextual information and x_i^a comprises acoustic signal information. The next section will discuss how we perform modality fusion on x_i^t and x_i^a.

2) Graph-based Dynamic Fusion for Multi-modal Integration: Integration of multi-modality has two components discussed below.

Graph Construction: Following the studies that use the graph-based method to model utterances in conversation [51], [52], [54], [58], we construct an undirected graph to fuse the acoustic (x_i^a) and textual (x_i^t) embeddings.

To build a graph of a conversation, each utterance is represented by two nodes: one textual node x_i^t and one acoustic node x_i^a. Therefore, for a conversation with n utterances, there are a total of $2n$ nodes. We follow two rules to connect nodes: (1) two nodes from the same modality are connected, i.e., x_i^t and x_j^t, enhancing the intra-modality context information; (2) two nodes of different modalities from the same utterance are connected, i.e., x_i^a and x_i^t, enhancing the inter-modality complementarity. The weights of the edges are determined by the cosine similarity between the two nodes [69]:

$$A_{ij} = 1 - \frac{arccos(sim(n_i, n_j))}{\pi} \tag{7}$$

Dynamic Fusion with Graph Convolution Based on the graph built in the previous section, we apply the dynamic fusion module designed by [52] to perform the inter- and intra-modality information fusion using a K-layer gating graph convolution. The graph convolution for the k-th layer is defined as [70]:

$$H^k = \sigma(((1-\alpha)\tilde{P}H^{k-1}+\alpha H^0)((1-\beta_{k-1})I_n+\beta_{k-1}W^{k-1})) \tag{8}$$

where $\tilde{P} = \tilde{D}^{-1/2}\tilde{A}\tilde{D}^{-1/2}$ is the graph convolution matrix with the renormalization trick [71]. σ denotes the ReLU activation function and W^k is the weight matrix of the k-th layer. H^0 is the initial representation of nodes (i.e., x_i^t and x_a^i from Equation (6)). I_n is an identity mapping matrix. α is a hyper-parameter that ensures the final representation of each node consists of a fraction of its initial features; $\beta_k = \log(\frac{\gamma}{k}+1)$ with the hyper-parameter γ ensures the decay of the weight matrix adaptively increases as the number of layers increases [70].

On top of the graph convolutional layer, a gating mechanism is applied to help the model learn intrinsic sequential patterns of contextual information in different semantic spaces by controlling how many contexts are to be stored and removing redundant information [52], [72]. An LSTM is used to implement the gating mechanism, and the initial hidden state h^0 is set to zero. The output of the k-th layer H_{out}^k is defined as:

$$g^k, h^k = LSTM(H^k, h^{k-1}) \tag{9}$$

$$H_{out}^k = H^k + g^k \tag{10}$$

Fig 3 demonstrates how the graph is constructed and updated using the graph convolutional layers and gating mechanism. After K layers, the embedding of each node is extracted from H_{out}^K and denoted as e_i^t or e_i^a, corresponding to the textual and acoustic modality of the i-th utterance, respectively. By concatenating e_i^t and e_i^a, we obtain the embedding for the utterance $u_i^c = [e_i^a, e_i^t]$, which encapsulates a fusion of inter- and intra-modality contextual information along with intrinsic sequential patterns.

3) Attention-based Multiple Instance Learning Pooling: This section discusses the strategy of aggregating information from utterance embeddings to assess hostility.

To address the challenge posed by the varying number of utterances in interview conversations and to focus on the utterances most relevant to hostility detection, we formulate the task as a multiple instance learning (MIL) problem. In a traditional binary supervised learning problem, a classifier receives an instance x as input and infers a target label value y. But in the MIL framework, instead of one single input instance, the classifier receives a bag of instances $B = \{x_1, ..., x_n\}$ and generates an inference for the bag [73].

The MIL paradigm aligns with the structure of our data, where a conversation consists of multiple utterances, and hostile behavior may not be consistently displayed throughout but might be prominent in certain utterances. In the hostility assessment task, an individual's hostility is evaluated based on the overall behaviors during the entire conversation rather than isolated from each utterance. Hence, we consider each utterance as an 'instance' and the conversation as a 'bag'. Particularly, we take the embedding u_i^c of each utterance from the interviewee (i.e., $u_i \in U_A$) as one instance and generate a bag representation. Utterances from the interviewer are not included for two reasons: (1) the hostility detection is for the interviewee; (2) the interactions between interviewer and interviewee have already been encoded through contextual embedding and graph-based fusion.

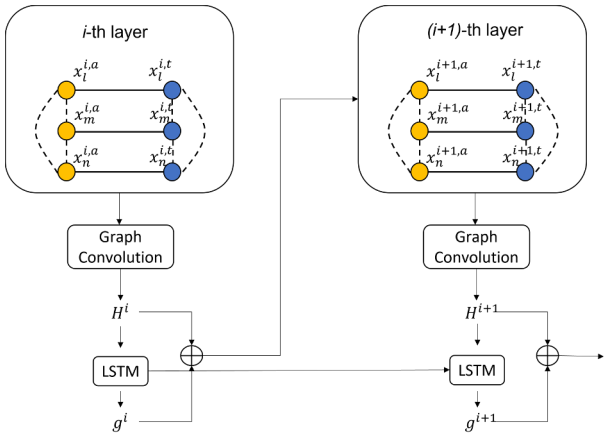

Fig. 3. In the graph, yellow nodes are from acoustic modality while the blue nodes are from textual modality. x_l, x_m, x_n denote different utterances. The existence of inter- and intra-modality edges helps better fuse the modalities as well as context information.

Then, we apply the embedding-based MIL approach to aggregate instance-level embeddings. A MIL pooling takes the embeddings and generates a bag representation z that is independent of the number of instances in the bag. This approach generates a joint representation of a bag from the instances; hence it does not introduce any additional bias to the bag-level classification [68], [74].

Following the attention-based MIL [68], we generate the bag representation z using a weighted sum of instance embeddings u_i^c where a neural network calculates the weights w_i. Additionally, the weights must sum to 1 to be invariant to the number of instances of a bag. Given a set of m utterance embeddings $[u_1^c, u_2^c, ..., u_m^c]$, the attention-based MIL pooling generate the bag representation z using the following:

$$z = \sum_{k=1}^{m} w_k u_k^c \tag{11}$$

$$w_k = \frac{exp\{w^T tanh(V(u_k^c)^T)\}}{\sum_{j=1}^{m} exp\{w^T tanh(V(u_j^c)^T)\}} \tag{12}$$

where w and V are learnable network parameters.

Attention-based MIL pooling can model an arbitrary permutation-invariant weight generation function that assigns different weights to instances within a bag. Hence, the bag representation is highly informative and capable of identifying the key utterances (i.e., instances) indicative of high hostility.

Finally, the bag-level representation z is passed to some decision-making linear layers, which generate the log-likelihoods of the positive and negative labels Y. The network is trained through backpropagation with a binary negative log-likelihood loss.

V. Experiments

Implementaion Details: Following the related ERC works [52], [58], [59], we extract utterance-level textual features using TextCNN [75], and project the raw acoustic features of an utterance into a lower dimension using linear layers. The number of layers of graph convolution K is set to 16; in Equation (8), α and β_k is set to 0.2 and $\log(0.5/k+1)$, as the previous studies [52], [70] suggest. In experiments, the dataset is divided into person-disjoint 5 folds for generalized results; each fold has a class-balanced testing set with data from 20 interviews. Reported results are averaged over the 5 folds.

The following sections discuss the model's performance in classifying high- and low-hostile individuals based on audio and text inputs.

A. Baselines

Because there is no previous work on multi-modal hostility assessment during interview conversation, we adapt some SOTA models designed for multi-modal ERC as the baselines:

(1) **BC-LSTM** [76] encodes contextual semantic information using Bi-directional LSTM. Modality fusion is performed by early concatenation.
(2) **EmoCaps** [77] utilizes a transformer encoder [78] to track emotional tendency in conversation. The embeddings from various modalities are concatenated with the original textual feature to make a capsule. Here, the context modeling is implemented via LSTM.
(3) **MM-DFN** [52] is one of the SOTA approaches that apply graph neural networks for multi-modal fusion. It models the contexts in conversations via GRU.
(4) **Trans-MM-DFN** is a recent variant of **MM-DFN**, by using a transformer encoder to extract contextual information instead of GRU, as transformer models have a better sequential context modeling ability [62].

Notably, the original baseline papers' approaches perform inference at the utterance level. In contrast, our task needs to aggregate information from multiple utterances to infer hostility. So, we apply mean pooling over the embeddings of the interviewee's utterances to generate a final embedding for individual-level hostility classification.

B. Experimental Results and Analysis

1) Overall Performance: The performance of baselines and GMGF-MIL is summarized in Table II. Presented evaluation metrics are: Accuracy, true positive rate (TPR), and true negative rate (TNR). According to the results, overall, GMGF-MIL outperforms all baselines. Details are discussed below:

- *GMGF-MIL vs. MM-DFN:* The results demonstrate that the GMGF-MIL outperforms the other baseline models overall. Utilizing attention-based MIL pooling, our model achieves a 3% improvement in accuracy compared to the second-best baseline model, MM-DFN which employs mean pooling. Compared to the mean pooling method, the MIL pooling method unevenly distributes weights to utterances according to their importance, demonstrating a better representation. The interpretation of MIL pooling is provided in Section V-B2.
- *GMGF-MIL vs. BC-LSTM and EmoCaps:* Compared to BC-LSTM and EmoCaps, which utilize simple concatenation for multi-modal fusion, GMGF-MIL's graph-based methods exhibit higher accuracy. This finding suggests that graphs offer a more effective multi-modal conversational information integration approach capable of capturing both inter- and intra-modality correlations.
- *GMGF-MIL vs. Trans-MM-DFN:* Trans-MM-DFN achieves a lower accuracy compared to MM-DFN when employing a transformer to extract contextual information instead of a GRU. This discrepancy may arise from the increased complexity of the transformer module, which potentially compromises the model's generalization ability. A similar observation is noted in the ablation of GMGF-MIL, discussed in Section V-C2.

All models demonstrate a higher TNR than TPR, indicating that classifying low-hostile individuals is generally easier than classifying high-hostile individuals. This observation could stem from the data distribution (Fig. 1), where some interviewees with IHAT scores slightly above the threshold are labeled as hostile (positive). High-hostile individuals situated near the label boundary pose a challenge for correct classification, leading models to produce more false negatives (e.g., misclassifying them as low-hostile).

TABLE II
Average accuracy and standard variation of different models on 5 folds. True Positive Rate (TPR) and True Negative Rate (TNR) indicate the models' capability to correctly identify hostile and non-hostile individuals, respectively.

Model	Accuracy	TPR	TNR
BC-LSTM	68% (±6.8%)	62% (±22.3%)	74% (±15.0%)
EmoCaps	70% (±7.1%)	66% (±17.4%)	74% (±10.2%)
MM-DFN	75% (±9.3%)	66% (±23.0%)	84% (±15.2%)
Trans-MM-DFN	73% (±5.7%)	70% (±20.0%)	76% (±15.2%)
GMGF-MIL	**78%** (±9.1%)	**74**% (±24.1%)	<u>82%</u> (±13.0%)

2) Interpretation of MIL Weights: The MIL weight, denoted as w_k in Equation (12), assigned to each instance (i.e., utterance) within a bag (i.e., conversation), indicates the significance of the instance in influencing the final inference. This section identifies if certain conversational topics are more related to hostility identification by visualizing the weights assigned to each question's corresponding responses.

On average, interviewees provide approximately 20 utterances in response to the interviewer's questions listed in Table I, with 1 to 3 utterances typically allocated to answer each question. To have better insights into how specific questions prompt responses from interviewees, which bear important implications for hostility identification, we record the MIL weight for each question's corresponding utterances from

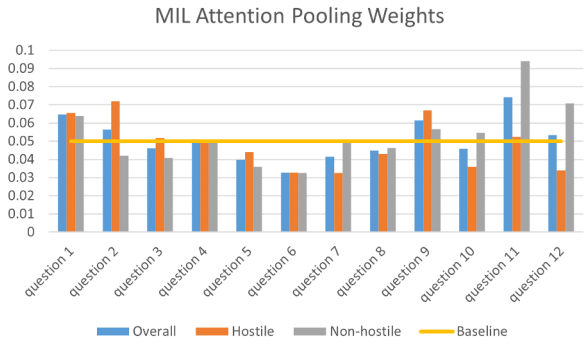

Fig. 4. MIL weights of interviewee's responses to each question. If all utterances were equally important for the final prediction, the weights would be evenly distributed, as depicted by the yellow baseline. The blue bar represents the overall average weights of the responses across all individuals; the orange and gray bars represent the weights of the responses among hostile and non-hostile individuals, respectively.

interviewees who are correctly classified. The average weight of the interviewee's utterance response to each question is shown in Fig. 4.

Fig. 4 illustrates that the weights assigned to the interviewee's responses to questions 1, 2, 9, and 11 exceed the baseline weight, indicating that interviewees are more likely to exhibit notable high vs. low hostility patterns when answering these questions. Particularly noteworthy is the observation that when responding to questions about work pressure (questions 2 and 9), high-hostile individuals exhibit a distinct marker, as evidenced by the much higher orange bars. Conversely, when addressing questions related to waiting in lines (questions 11 and 12), low-hostile individuals display a noticeable marker, perhaps indicating greater patience, as indicated by the elevated gray bar.

This visualization underscores the varying contributions of different questions to hostility identification, highlighting the importance for researchers to prioritize certain topics over others. It suggests potential improvements for the design of structured interviews in future studies related to hostility assessment, such as reducing trivial questions and creating a shorter interview.

C. Ablation Study

1) Comparison between Textual and Acoustic Modalities: This section investigates how acoustic and textual modality independently contribute to hostility identification.

In this evaluation, we exclusively utilize either text or audio as input to infer individual hostility. Consequently, the inter-modality edges depicted in Fig. 3 are absent, with only intra-modality relationships incorporated into the graph. We construct two classifiers employing identical architectures to classify hostility based on text and audio separately. The results presented in Table III reveal that the acoustic modality yields much higher accuracy than the textual modality when utilized independently. This observation aligns with findings

TABLE III
COMPARISON BETWEEN THE TWO MODALITIES IF USED SOLELY

Modality	Audio Only	Text Only
Accuracy	72% (± 7.6%)	64% (± 8.2%)

from the study [12], emphasizing the critical role of vocal stylistics in determining hostile behaviors.

The outcome suggests that hostility detection relies more heavily on the acoustic modality, contrasting with emotion detection, where textual modality often holds greater significance [52], [77]. Comparing these results with Table II, we observe that by leveraging the graph-based dynamic fusion mechanism to combine the two modalities, the rich textual semantics complement affective acoustic features, resulting in higher accuracy (78% accuracy achieved by our method).

2) Context Extraction Model: We experimented with utilizing a transformer encoder, known for its efficacy in encoding sequential data [78], as the contextual information extractor in our approach. However, the transformer-based model yielded only an average accuracy of 70%, much lower than the 78% accuracy of our GRU-based model. A similar trend is evident when comparing MM-DFN and Trans-MM-DFN in Table II. This discrepancy may stem from the transformer model's complexity, which might be excessive for our limited dataset compared to the GRU model, potentially resulting in overfitting during model training.

VI. Conclusion

This paper introduces GMGF-MIL, a pioneering automatic approach for distinguishing individuals with high vs. low transactional hostility from conversational speech, with significant applications in healthcare and beyond. GMGF-MIL's innovative design captures conversational context, integrates multiple modalities effectively, and identifies crucial utterances indicative of hostility, resulting in high efficacy. For the first time, GMGF-MIL enables unbiased and automated assessment of transactional hostility, making it highly impactful for practical applications. Additionally, through analysis of MIL weights of responses to different interview questions, this study identifies prioritized topics and suggests potential improvement for designing structured interviews.

Ethical Impact Statement

It is essential to clarify that our research on classifying individuals' hostility levels aims at supporting healthcare services through enhancing CVD risk assessments, rather than passing judgment on any individual's personality. When using this tool in a medical setting, it is crucial to avoid making patients feel blamed or criticized. Practical implementation needs to frame the assessment of interview behaviors as reflecting enduring responses, like cardiovascular reactions to acute psychological stress [24], that are difficult to modify and yet might affect CVD risk. Given the constraints of our dataset, it is imperative to validate our methods and findings on larger samples and diverse demographic groups to ensure

fairness and reliability. Original interview recordings were collected by an IRB-approved and already published study, the Multiple Risk Factor Intervention Trial (MRFIT) [55]. Proper IRB approval was obtained to get access to the interview audio recordings that were collected as part of the original study. While we are unable to publicly share the sensitive raw data, we will release the processed features and code for reference.

REFERENCES

[1] J. C. Barefoot, "Developments in the measurement of hostility." 1992.

[2] R. C. Lorenzetti, C. M. Jacques, C. Donovan, S. Cottrell, and J. Buck, "Managing difficult encounters: understanding physician, patient, and situational factors," *American Family Physician*, vol. 87, no. 6, pp. 419–425, 2013.

[3] R. L. Street Jr, G. Makoul, N. K. Arora, and R. M. Epstein, "How does communication heal? pathways linking clinician–patient communication to health outcomes," *Patient education and counseling*, vol. 74, no. 3, pp. 295–301, 2009.

[4] M. D. Faay, J. Van Os, G. Risk, and O. of Psychosis (GROUP) Investigators, "Aggressive behavior, hostility, and associated care needs in patients with psychotic disorders: a 6-year follow-up study," *Frontiers in psychiatry*, vol. 10, p. 934, 2020.

[5] S. Wagner, R. Pasca, and J. Crosina, "Hostility in firefighters: Personality and mental health," *International Journal of Emergency Services*, vol. 5, no. 1, pp. 6–17, 2016.

[6] K. Kendall-Tackett, "Depression, hostility, posttraumatic stress disorder, and inflammation: The corrosive health effects of negative mental states." 2010.

[7] T. W. Smith, K. Glazer, J. M. Ruiz, and L. C. Gallo, "Hostility, anger, aggressiveness, and coronary heart disease: An interpersonal perspective on personality, emotion, and health," *Journal of personality*, vol. 72, no. 6, pp. 1217–1270, 2004.

[8] K. A. Matthews, B. B. Gump, K. F. Harris, T. L. Haney, and J. C. Barefoot, "Hostile behaviors predict cardiovascular mortality among men enrolled in the multiple risk factor intervention trial," *Circulation*, vol. 109, no. 1, pp. 66–70, 2004.

[9] S. Sahoo, S. K. Padhy, B. Padhee, N. Singla, and S. Sarkar, "Role of personality in cardiovascular diseases: An issue that needs to be focused too!" *Indian heart journal*, vol. 70, pp. S471–S477, 2018.

[10] R. B. Shekelle, M. Gale, A. M. Ostfeld, and O. Paul, "Hostility, risk of coronary heart disease, and mortality," *Psychosomatic medicine*, vol. 45, no. 2, pp. 109–114, 1983.

[11] J. C. Barefoot, K. A. Dodge, B. L. Peterson, W. G. Dahlstrom, and R. B. Williams Jr, "The cook-medley hostility scale: item content and ability to predict survival." *Psychosomatic medicine*, vol. 51, no. 1, pp. 46–57, 1989.

[12] T. L. Haney, K. E. Maynard, S. J. Houseworth, and L. W. Scherwitz, "Interpersonal hostility assessment technique: description and validation against the criterion of coronary artery disease," *Journal of Personality Assessment*, vol. 66, no. 2, pp. 386–401, 1996.

[13] E. J. Vella, T. W. Kamarck, J. D. Flory, and S. Manuck, "Hostile mood and social strain during daily life: A test of the transactional model," *Annals of Behavioral medicine*, vol. 44, no. 3, pp. 341–352, 2012.

[14] R. B. Williams Jr, T. L. Haney, K. L. Lee, J. A. Blumenthal, R. E. Whalen *et al.*, "Type a behavior, hostility, and coronary atherosclerosis," *psychosomatic Medicine*, vol. 42, no. 6, pp. 539–549, 1980.

[15] M. A. Chesney, J. R. Eagleston, and R. H. Rosenman, "The type a structured interview: A behavioral assessment in the rough," *Journal of Behavioral Assessment*, vol. 2, pp. 255–272, 1980.

[16] J. Suls and G. S. Sanders, "Why do some behavioral styles place people at coronary risk?" in *In search of coronary-prone behavior*. Psychology Press, 2013, pp. 1–20.

[17] A. W. Siegman and T. M. Dembroski, *In search of coronary-prone behavior: Beyond Type A*. Psychology Press, 2013.

[18] A. Xenos, J. Pavlopoulos, and I. Androutsopoulos, "Context sensitivity estimation in toxicity detection," in *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, 2021, pp. 140–145.

[19] J. Pavlopoulos, L. Laugier, A. Xenos, J. Sorensen, and I. Androutsopoulos, "From the detection of toxic spans in online discussions to the analysis of toxic-to-civil transfer," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2022, pp. 3721–3734.

[20] F. T. Boishakhi, P. C. Shill, and M. G. R. Alam, "Multi-modal hate speech detection using machine learning," in *2021 IEEE International Conference on Big Data (Big Data)*. IEEE, 2021, pp. 4496–4499.

[21] T. Agrawal, M. Balazia, P. Müller, and F. Brémond, "Multimodal vision transformers with forced attention for behavior analysis," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 3392–3402.

[22] A. Guo, R. Hirai, A. Ohashi, Y. Chiba, Y. Tsunomori, and R. Higashinaka, "Personality prediction from task-oriented and open-domain human–machine dialogues," *Scientific Reports*, vol. 14, no. 1, p. 3868, 2024.

[23] D. Curto, A. Clapés, J. Selva, S. Smeureanu, J. Junior, C. Jacques, D. Gallardo-Pujol, G. Guilera, D. Leiva, T. B. Moeslund *et al.*, "Dyadformer: A multi-modal transformer for long-range modeling of dyadic interactions," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 2177–2188.

[24] T. W. Kamarck and W. R. Lovallo, "Cardiovascular reactivity to psychological challenge: conceptual and measurement considerations," *Psychosomatic medicine*, vol. 65, no. 1, pp. 9–21, 2003.

[25] T. W. Smith, "Hostility and health: current status of a psychosomatic hypothesis." *Health psychology*, vol. 11, no. 3, p. 139, 1992.

[26] Y. Birks and D. Roger, "Identifying components of type-a behaviour:"toxic" and "non-toxic" achieving," *Personality and Individual Differences*, vol. 28, no. 6, pp. 1093–1105, 2000.

[27] M. S. Hee, W.-H. Chong, and R. K.-W. Lee, "Decoding the underlying meaning of multimodal hateful memes," *arXiv preprint arXiv:2305.17678*, 2023.

[28] D. Kiela, H. Firooz, A. Mohan, V. Goswami, A. Singh, P. Ringshia, and D. Testuggine, "The hateful memes challenge: Detecting hate speech in multimodal memes," *Advances in neural information processing systems*, vol. 33, pp. 2611–2624, 2020.

[29] P. Badjatiya, S. Gupta, M. Gupta, and V. Varma, "Deep learning for hate speech detection in tweets," in *Proceedings of the 26th international conference on World Wide Web companion*, 2017, pp. 759–760.

[30] M. Das, R. Raj, P. Saha, B. Mathew, M. Gupta, and A. Mukherjee, "Hatemm: A multi-modal dataset for hate video classification," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 17, 2023, pp. 1014–1023.

[31] L. R. Goldberg, "The structure of phenotypic personality traits." *American psychologist*, vol. 48, no. 1, p. 26, 1993.

[32] S. Song, Z. Shao, S. Jaiswal, L. Shen, M. Valstar, and H. Gunes, "Learning person-specific cognition from facial reactions for automatic personality recognition," *IEEE Transactions on Affective Computing*, 2022.

[33] R. Liao, S. Song, and H. Gunes, "An open-source benchmark of deep learning models for audio-visual apparent and self-reported personality recognition," *IEEE Transactions on Affective Computing*, 2024.

[34] Z. Lian, L. Chen, L. Sun, B. Liu, and J. Tao, "Gcnet: Graph completion network for incomplete multimodal learning in conversation," *IEEE Transactions on pattern analysis and machine intelligence*, 2023.

[35] J. Huang, Y. Li, J. Tao, Z. Lian, M. Niu, and M. Yang, "Multimodal continuous emotion recognition with data augmentation using recurrent neural networks," in *Proceedings of the 2018 on audio/visual emotion challenge and workshop*, 2018, pp. 57–64.

[36] J. Williams, S. Kleinegesse, R. Comanescu, and O. Radu, "Recognizing emotions in video using multimodal dnn feature fusion," in *Proceedings of Grand Challenge and Workshop on Human Multimodal Language (Challenge-HML)*, 2018, pp. 11–19.

[37] T.-L. Nguyen, S. Kavuri, and M. Lee, "A multimodal convolutional neuro-fuzzy network for emotion understanding of movie clips," *Neural Networks*, vol. 118, pp. 208–219, 2019.

[38] J. D. Ortega, M. Senoussaoui, E. Granger, M. Pedersoli, P. Cardinal, and A. L. Koerich, "Multimodal fusion with deep neural networks for audio-video emotion recognition," *arXiv preprint arXiv:1907.03196*, 2019.

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

[40] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, "Tensor fusion network for multimodal sentiment analysis," *arXiv preprint arXiv:1707.07250*, 2017.

[41] T. Mittal, U. Bhattacharya, R. Chandra, A. Bera, and D. Manocha, "M3er: Multiplicative multimodal emotion recognition using facial, textual, and speech cues," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 02, 2020, pp. 1359–1367.

[42] J. Ma, H. Tang, W.-L. Zheng, and B.-L. Lu, "Emotion recognition using multimodal residual lstm network," in *Proceedings of the 27th ACM international conference on multimedia*, 2019, pp. 176–183.

[43] F. Tao and G. Liu, "Advanced lstm: A study about better time dependency modeling in emotion recognition," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 2906–2910.

[44] N. Majumder, S. Poria, D. Hazarika, R. Mihalcea, A. Gelbukh, and E. Cambria, "Dialoguernn: An attentive rnn for emotion detection in conversations," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, no. 01, 2019, pp. 6818–6825.

[45] J. Huang, J. Tao, B. Liu, Z. Lian, and M. Niu, "Multimodal transformer fusion for continuous emotion recognition," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 3507–3511.

[46] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in *Proceedings of the conference. Association for computational linguistics. Meeting*, vol. 2019. NIH Public Access, 2019, p. 6558.

[47] Z. Lin, B. Liang, Y. Long, Y. Dang, M. Yang, M. Zhang, and R. Xu, "Modeling intra-and inter-modal relations: Hierarchical graph contrastive learning for multimodal sentiment analysis," in *Proceedings of the 29th International Conference on Computational Linguistics*, vol. 29, no. 1. Association for Computational Linguistics, 2022, pp. 7124–7135.

[48] J. Li, X. Wang, G. Lv, and Z. Zeng, "Graphcfc: A directed graph based cross-modal feature complementation approach for multimodal conversational emotion recognition," *IEEE Transactions on Multimedia*, 2023.

[49] T. Ishiwatari, Y. Yasuda, T. Miyazaki, and J. Goto, "Relation-aware graph attention networks with relational position encodings for emotion recognition in conversations," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 7360–7370.

[50] Y. Zhang, A. Jia, B. Wang, P. Zhang, D. Zhao, P. Li, Y. Hou, X. Jin, D. Song, and J. Qin, "M3gat: A multi-modal, multi-task interactive graph attention network for conversational sentiment analysis and emotion recognition," *ACM Transactions on Information Systems*, vol. 42, no. 1, pp. 1–32, 2023.

[51] D. Zhang, L. Wu, C. Sun, S. Li, Q. Zhu, and G. Zhou, "Modeling both context-and speaker-sensitive dependence for emotion detection in multi-speaker conversations." in *IJCAI*, 2019, pp. 5415–5421.

[52] D. Hu, X. Hou, L. Wei, L. Jiang, and Y. Mo, "Mm-dfn: Multimodal dynamic fusion network for emotion recognition in conversations," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7037–7041.

[53] A. B. Zadeh, P. P. Liang, S. Poria, E. Cambria, and L.-P. Morency, "Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, pp. 2236–2246.

[54] D. Ghosal, N. Majumder, S. Poria, N. Chhaya, and A. Gelbukh, "Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation," *arXiv preprint arXiv:1908.11540*, 2019.

[55] J. D. Neaton, R. H. Grimm Jr, J. A. Cutler, M. R. Group *et al.*, "Recruitment of participants for the multiple risk factor intervention trial (mrfit)," *Controlled Clinical Trials*, vol. 8, no. 4, pp. 41–53, 1987.

[56] Google, "Speech-to-text," https://cloud.google.com/speech-to-text?hl=endocumentation, 2024.

[57] S. Jin and R. Zafarani, "Emotions in social networks: Distributions, patterns, and models," in *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, 2017, pp. 1907–1916.

[58] J. Hu, Y. Liu, J. Zhao, and Q. Jin, "Mmgcn: Multimodal fusion via deep graph convolution network for emotion recognition in conversation," *arXiv preprint arXiv:2107.06779*, 2021.

[59] D. Hazarika, S. Poria, R. Mihalcea, E. Cambria, and R. Zimmermann, "Icon: Interactive conversational memory network for multimodal emotion detection," in *Proceedings of the 2018 conference on empirical methods in natural language processing*, 2018, pp. 2594–2604.

[60] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in opensmile, the munich open-source multimedia feature extractor," in *Proceedings of the 21st ACM international conference on Multimedia*, 2013, pp. 835–838.

[61] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller, and S. Narayanan, "The interspeech 2010 paralinguistic challenge," in *Proc. INTERSPEECH 2010, Makuhari, Japan*, 2010, pp. 2794–2797.

[62] Y. Shou, T. Meng, W. Ai, N. Yin, and K. Li, "A comprehensive survey on multi-modal conversational emotion recognition with deep learning," *arXiv preprint arXiv:2312.05735*, 2023.

[63] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

[64] V. Chudasama, P. Kar, A. Gudmalwar, N. Shah, P. Wasnik, and N. Onoe, "M2fnet: multi-modal fusion network for emotion recognition in conversation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 4652–4661.

[65] G. Hu, T.-E. Lin, Y. Zhao, G. Lu, Y. Wu, and Y. Li, "Unimse: Towards unified multimodal sentiment analysis and emotion recognition," *arXiv preprint arXiv:2211.11256*, 2022.

[66] J. Li, X. Wang, G. Lv, and Z. Zeng, "Ga2mif: Graph and attention based two-stage multi-source information fusion for conversational emotion detection," *IEEE Transactions on Affective Computing*, 2023.

[67] S. Jin and R. Zafarani, "Sentiment prediction in social networks," in *2018 IEEE International Conference on Data Mining Workshops (ICDMW)*. IEEE, 2018, pp. 1340–1347.

[68] M. Ilse, J. Tomczak, and M. Welling, "Attention-based deep multiple instance learning," in *International conference on machine learning*. PMLR, 2018, pp. 2127–2136.

[69] K. Skianis, F. Malliaros, and M. Vazirgiannis, "Fusing document, collection and label graph-based representations with word embeddings for text classification," in *NAACL-HLT Workshop on Graph-Based Natural Language Processing (TextGraphs)*, 2018.

[70] M. Chen, Z. Wei, Z. Huang, B. Ding, and Y. Li, "Simple and deep graph convolutional networks," in *International conference on machine learning*. PMLR, 2020, pp. 1725–1735.

[71] R. Li, S. Wang, F. Zhu, and J. Huang, "Adaptive graph convolutional neural networks," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018.

[72] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[73] O. Maron and T. Lozano-Pérez, "A framework for multiple-instance learning," *Advances in neural information processing systems*, vol. 10, 1997.

[74] X. Wang, Y. Yan, P. Tang, X. Bai, and W. Liu, "Revisiting multiple instance neural networks," *Pattern Recognition*, vol. 74, pp. 15–24, 2018.

[75] Y. Kim, "Convolutional neural networks for sentence classification," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, A. Moschitti, B. Pang, and W. Daelemans, Eds. Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 1746–1751. [Online]. Available: https://aclanthology.org/D14-1181

[76] S. Poria, E. Cambria, D. Hazarika, N. Majumder, A. Zadeh, and L.-P. Morency, "Context-dependent sentiment analysis in user-generated videos," in *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: Long papers)*, 2017, pp. 873–883.

[77] Z. Li, F. Tang, M. Zhao, and Y. Zhu, "EmoCaps: Emotion capsule based model for conversational emotion recognition," in *Findings of the Association for Computational Linguistics: ACL 2022*. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 1610–1618. [Online]. Available: https://aclanthology.org/2022.findings-acl.126

[78] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.