electronics

MDPI

Article

Hybrid Uncertainty Calibration for Multimodal Sentiment Analysis

Qiuyu Pan and Zuqiang Meng *

School of Computer and Electronic Information, Guangxi University, Nanning 530004, China; 2113391052@st.gxu.edu.cn

* Correspondence: zqmeng@126.com or zqmeng@gxu.edu.cn

Abstract: In open environments, multimodal sentiment analysis (MSA) often suffers from low-quality data and can be disrupted by noise, inherent defects, and outliers. In some cases, unreasonable multimodal fusion methods can perform worse than unimodal methods. Another challenge of MSA is effectively enabling the model to provide accurate prediction when it is confident and to indicate high uncertainty when its prediction is likely to be inaccurate. In this paper, we propose an uncertain-aware late fusion based on hybrid uncertainty calibration (ULF-HUC). Firstly, we conduct in-depth research on the issue of sentiment polarity distribution in MSA datasets, establishing a foundation for an uncertain-aware late fusion method, which facilitates organic fusion of modalities. Then, we propose a hybrid uncertainty calibration method based on evidential deep learning (EDL) that balances accuracy and uncertainty, supporting the reduction of uncertainty in each modality of the model. Finally, we add two common types of noise to validate the effectiveness of our proposed method. We evaluate our model on three publicly available MSA datasets (MVSA-Single, MVSA-Multiple, and MVSA-Single-Small). Our method outperforms state-of-the-art approaches in terms of accuracy, weighted F1 score, and expected uncertainty calibration error (UCE) metrics, proving the effectiveness of the proposed method.

Keywords: hybrid uncertainty calibration; multimodal sentiment analysis; uncertainty-aware late fusion; expected uncertainty calibration error; noise

Citation: Pan, Q.; Meng, Z. Hybrid Uncertainty Calibration for Multimodal Sentiment Analysis. *Electronics* **2024**, *13*, 662. https://doi.org/10.3390/electronics13030662

Academic Editors: Tao Shen, Lei Zhang and John Wang

Received: 19 December 2023
Revised: 28 January 2024
Accepted: 2 February 2024
Published: 5 February 2024

1. Introduction

Sentiment analysis can be applied in various domains, including marketing, customer service, brand management, political analysis, and social listening [1]. It has emerged as a highly active research area due to the enormous volume of data generated daily on social media platforms and the World Wide Web. This abundance of data provides a rich source for sentiment analysis research and applications. The conventional sentiment analysis model primarily concentrates on analyzing text-based content [2]. Nonetheless, advancements in technology have provided individuals with the means to express their opinions and emotions through various channels, including text, images, and videos. Due to these developments, sentiment analysis is transitioning from a focus on a single modality to considering multiple modalities. This shift brings about novel possibilities in sentiment analysis, driven by the rapid growth of this field. The integration of complementary data streams facilitates enhanced sentiment detection, surpassing the limitations of text-based analysis [3].

Recent advancements in multimodal sentiment analysis architectures can be categorized into ten distinct categories [4]. Different fusion methods have various strengths and limitations. Although multimodal fusion can solve the limitations of a single modality, in real open environments, multimodal data are usually disturbed by noise, defects, and abnormal points, making it difficult to satisfy the complementarity and consistency of multimodality [5]. In the past few years, numerous researchers have conducted studies on sentiment analysis based on images and text. However, many existing approaches in this field either rely on a straightforward concatenation of features extracted from different

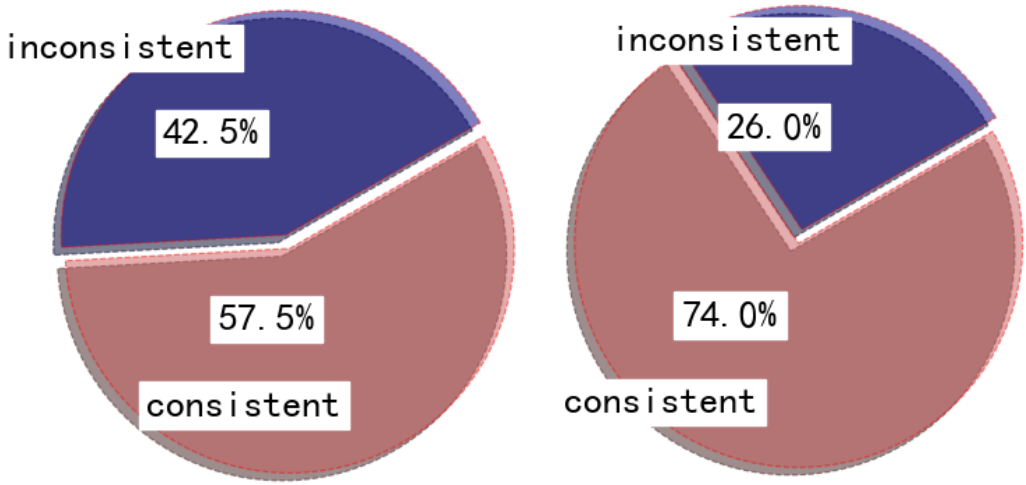
modalities [6] or only capturing coarse-level relationships between images and text [7]. Indeed, in the real open world, the sentiment polarity of text and visual content is not always completely aligned, which is one of the key challenges that need to be addressed for reliable multimodal learning.

In recent years, numerous studies have used MVSA datasets [8] (MVSA-single, MVSA-multiple) as benchmarks for exploring the sentiment analysis of images and text. These studies have pointed out that in the real open world, the sentiment polarity of text and visual content is not entirely consistent. Therefore, researchers often preprocess MVSA datasets by removing samples with opposite sentiment polarities. If one modality expresses a neutral sentiment, while the other modality is positive or negative, they are classified as positive or negative. Indeed, even after removing samples with opposite sentiment polarities, there are still a considerable number of inconsistent sentiment samples in the MVSA datasets. For convenience, we can classify samples where one modality expresses a neutral sentiment while the other modality is positive or negative as inconsistent samples, while samples where both modalities have consistent polarities can be categorized as polarity-consistent samples.

To conduct in-depth research, we classify the data in the filtered MVSA datasets, where one modality expresses a neutral sentiment while the other modality is positive or negative as inconsistent samples. On the other hand, samples where both modalities have consistent polarities are referred to as polarity-consistent samples. In Figure 1, detailed data analysis of the filtered MVSA datasets shows that a considerable proportion of samples exhibit inconsistent polarities. For example, MVSA-single contains 42.5% of samples with inconsistent polarities, while MVSA-multiple has 26.0% of samples with inconsistent polarities. We understand that within the samples, we define as having inconsistent polarities, one modality's data are neutral while the other is positive or negative. This means that we would need to incur an additional cost of 42.5% or 26.0% to inform our classifier that these originally neutral data points need to be classified as either positive or negative. This poses a significant challenge for any model. Indeed, looking at it from another perspective, when a modality's sentiment is initially neutral, the classifier needs to learn to associate it with a positive or negative sentiment in conjunction with the other modality. The classifier must effectively address the high uncertainty that arises during the polarity transformation process to achieve consistent and balanced learning between modalities. This entails capturing the nuanced relationships between modalities and understanding how they contribute to the overall sentiment analysis task. The classifier must strike a balance and minimize the ambiguity inherent in polarity conversion to achieve reliable and accurate results.

In particular, we should note that for the MVSA-multiple dataset, each pair is shown to three annotators, and each annotator independently judges the sentiments of the text and image. For the same text–image pair, the sentiment polarities given by different annotators are mostly different, indicating the widespread presence of high uncertainty in the model learning process. We must address the issue of high uncertainty in the model learning process to make our model's classification more robust.

However, it is regrettable that the current studies [9–11] on multimodal sentiment analysis rarely focuses on uncertainty calibration, thus overlooking the crucial significance of uncertainty calibration in improving model performance. Recently, there have been studies [12,13] focusing on improving model performance from the perspective of uncertainty calibration, whereas these methods often discuss the issue from an unimodal perspective, neglecting the challenge of inconsistent sentiment polarities across different modalities, which poses a new challenge to uncertainty calibration.



(**a**) (**b**)

Figure 1. Consistency analysis of sentimental polarity in the filtered MVSA-Single and MVSA-Multiple datasets. In our case, we consider samples where one modality expresses neutral sentiment while the other modality expresses positive or negative sentiment as instances of inconsistent sentiment polarity. (**a**) MVSA-Single. (**b**) MVSA-Multiple.

Therefore, it is necessary to conduct further research to explore how to achieve effective uncertainty calibration in multimodal sentiment analysis. By accurately estimating and calibrating the uncertainty of models, we can enhance their reliability and robustness, thereby better addressing the differences in sentiment expression across different modalities and providing more accurate and consistent sentiment analysis results.

Based on the above analysis, we propose an uncertain-aware late fusion method based on hybrid uncertainty calibration (ULF-HUC) to enhance the calibration and classification of the model. The main contributions of this paper are summarized as follows:

- We propose a hybrid uncertainty calibration (HUC) method, which utilizes the labels of both modalities to impose uncertainty constraints on each modality separately, aiming to reduce the uncertainties in each modality and enhance the calibration ability of the model.
- We propose an uncertain-aware late fusion (ULF) method to enhance the classification ability of the model.
- We add common types of noise, such as Gaussian noise and salt-and-pepper noise, to the test set. Experimental results demonstrate that our proposed model exhibits greater generalization ability.

The rest of the paper is organized as follows. In Section 2, we put our approach in the context of relevant existing work. Then, in Section 3, we present a detailed description of our proposed method. In Section 4, we conduct an experimental evaluation and analysis of our approach. Finally, Section 5 provides a summary of our findings and concludes the paper.

2. Related Work

2.1. Multimodal Sentiment Analysis

With the transformation of social media, there has been an explosive emergence of internet information, such as images, voice, and videos, greatly enriching the content available on social media platforms. Multimodal sentiment analysis can help social media platforms better understand users' emotions and needs. Xu et al. [14] introduced an alternating co-attention mechanism in their work. The alternating co-attention mechanism allows for reciprocal interaction between the image and language modalities, enhancing the understanding and representation of both modalities in the joint modeling process. In [15],

they introduced a fusion framework that revolves around a Bidirectional Associative Memory (BAM). Unlike traditional data-level or score-level fusion strategies, this framework leverages a cognitive model of multisensory integration to enhance fusion performance. Kumar et al. [16] proposed a hybrid deep learning model that employed decision-level multimodal fusion to integrate these modalities in online content, enhancing the accuracy of sentiment prediction at a more detailed level. Jiang et al. [17] proposed a fusion extraction network model for multimodal sentiment analysis. The model utilizes an interactive information fusion technique to dynamically learn visual-specific textual representations and textual-specific visual representations. This approach aims to effectively combine and process multimodal information for improved sentiment analysis results. In [18], the researchers proposed a novel method called cross-modal Semantic Content Correlation (SCC) to capture complementary multimodal information for joint sentiment classification. This approach aims to leverage the semantic content shared between the two modalities to improve the accuracy of sentiment classification. Guo et al. [19] introduced a layout-driven multimodal attention network (LD-MAN) for end-to-end sentiment recognition in news articles. This approach allows for a comprehensive understanding of the news articles' sentiment by integrating both textual and visual information synergistically.

In [20], they introduced an image–text interactive graph neural network for sentiment analysis. The network utilized a graph structure where the node features were initially derived from text and image features. Ye et al. [21] proposed a sentiment-aware multimodal pre-training (SMP) framework for multimodal sentiment analysis. The SMP framework addresses the challenges in multimodal sentiment analysis by incorporating cross-modal contrastive learning, sentiment-aware pre-training objectives, and semantic information capture. In this article [22], the authors proposed a Gated Fusion Semantic Relation (GFSR) network, which aims to explore semantic relations for sentiment analysis in social media. This fusion process integrates both global and local information to capture the semantic relations between images and textual descriptions, leading to improved sentiment analysis results in the context of social media. Liu et al. [23] introduced the Scanning, Attention, and Reasoning (SAR) model for multimodal sentiment analysis to effectively comprehend and predict sentiment tendencies in multimodal content. The SAR model comprises several components designed to handle different aspects of the analysis process.

2.2. Multimodal Uncertainty Calibration

Over the past decade, neural networks have made significant strides and have found applications in a wide range of fields. However, as their use has expanded, the need for confidence in neural network predictions has become increasingly important. Traditional neural networks cannot often provide certainty estimates and can suffer from issues such as overconfidence or underconfidence, leading to poorly calibrated predictions [24].

In safety-critical applications, obtaining reliable and accurate uncertainty estimates from deep neural networks is crucial [25]. A well-calibrated model should provide accurate prediction when it is confident and indicate high uncertainty when its prediction is likely to be inaccurate [26]. However, uncertainty calibration is challenging since, there is no ground truth available for uncertainty estimates. To address this problem, Krishnan et al. [27] proposed an optimization method that leverages the relationship between accuracy and uncertainty as a reference for uncertainty calibration. They introduced a differentiable loss function called Accuracy versus Uncertainty Calibration (AvUC) that enables the model to learn to provide well-calibrated uncertainties while also improving accuracy.

While significant research efforts have focused on quantifying and reducing predictive uncertainty, most existing methods [28,29] are designed for unimodal data, neglecting the challenges posed by uncertainty calibration in multimodal data. To bridge this gap, Ma et al. [30] identified the issue of over-confidence on partial modalities in existing multimodal learning paradigms. They proposed a measure to evaluate confidence reliability and introduce a regularization strategy to enhance confidence calibration. In [31], they proposed a simple soft maximum distribution matching loss function that demonstrates how

to jointly learn well-calibrated and well-ranked unimodal uncertainty estimation, leading to a significant improvement in multi-modal classification performance. Kose et al. [32] proposed a novel error alignment uncertainty (EaU) optimization method and introduced the EaU calibration loss to guide the model in providing reliable uncertainty estimates that are correlated with model errors. In [33], the proposed method is based on uncertainty quantification techniques, which enable the adoption of a principled approach to reduce the number of patterns required to explain model predictions. The output classifier scores represent the probability of well-calibrated predictions. Wang et al. [34] proposed a non-parametric calibration method that aims to estimate uncertainty as accurately as possible, given an unknown data distribution. This approach enables multimodal uncertainty calibration without increasing model complexity or training costs.

However, despite the improved performance of multimodal models compared to unimodal ones, it has been observed that they still fail to fully harness the potential of multiple modalities [35]. For low-quality datasets, it is necessary to address the uncertainty within each modality to calibrate the overall uncertainty of the model and improve the performance of multi-modal models. To tackle this issue, in this paper, we propose an uncertain-aware late fusion method based on hybrid uncertainty calibration (ULF-HUC) that introduces a new approach to address multimodal uncertainty calibration. This approach aims to enhance the performance of multimodal models by reducing the uncertainties associated with individual modalities and improving their calibration.

3. Methodology

3.1. Framework Overview

To address the issue of uncertainty estimation in the fusion of different modalities, we propose an uncertain-aware late fusion method based on hybrid uncertainty calibration (ULF-HUC) to enhance the interpretability and robustness of a multimodal sentiment analysis model.

The architecture of this model, as shown in Figure 2, consists of four main components: unimodal backbones, ENN heads, hybrid uncertainty calibration (HUC), and uncertain-aware late fusion (ULF) method. Firstly, similar to most methods, we utilize BERT [36] to extract textual feature information and ResNet [37] to extract image feature information. To reduce the bias introduced by the heterogeneity between different modalities in the overall model, for each modality, we take the pooler output of the backbone of the unimodal model as input information for the ENN Head. The ENN Head primarily utilizes the method of Evidence Deep Learning (EDL) [38]. This approach overcomes the limitations of softmax-based Deep Neural Networks (DNNs) by incorporating the evidence framework, specifically Dempster–Shafer Theory (DST) [39], and Subjective Logic (SL) [40]. The model is trained using the Evidence Deep Learning (EDL) [38] loss function and regularized using our proposed Hybrid Uncertainty Calibration (HUC) module, which considers the outputs of EDL for both modalities. In contrast to traditional sentiment classification predictions, we employ a late fusion method based on uncertainty estimation to make more accurate sentiment analysis judgments. This approach allows for a more reliable and robust estimation of uncertainties and consequently improves the overall performance of the multimodal sentiment analysis model.

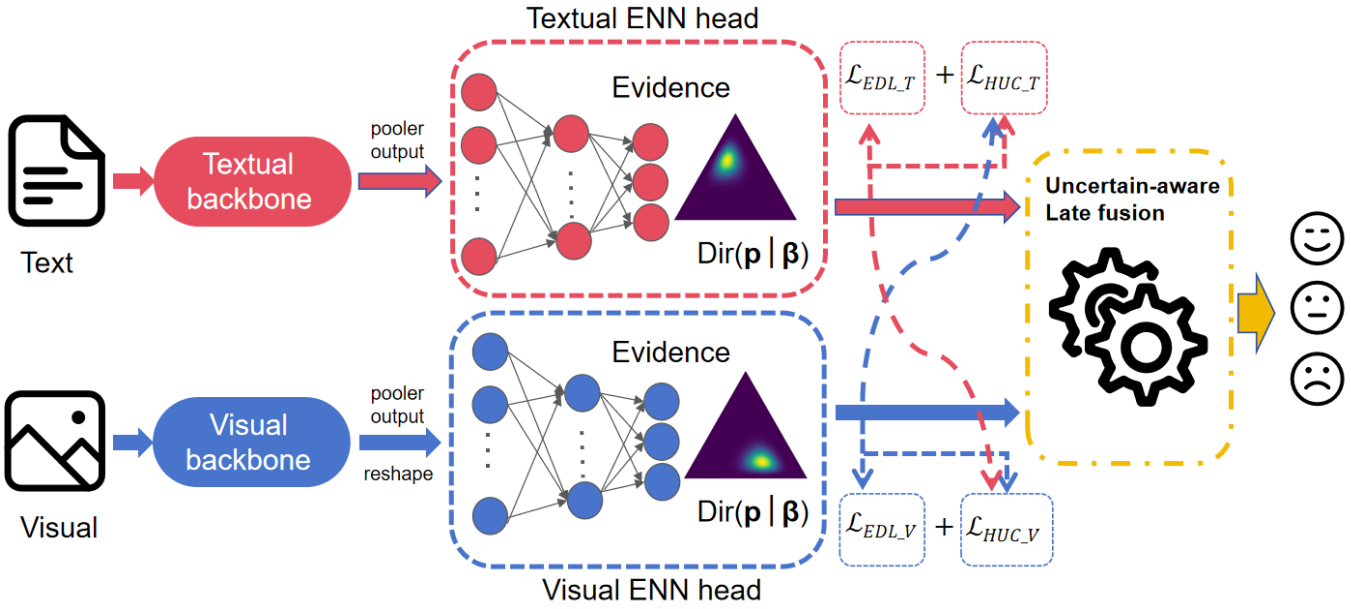


Figure 2. Illustration of our model's overall framework diagram. This model consists of four main components: unimodal backbones, ENN heads, hybrid uncertainty calibration (HUC), and uncertain-aware late fusion (ULF) method. The Evidential Neural Network (ENN) head predicts the evidence **e** to build the Dirichlet distribution of class probability **p**.

3.2. Unimodal Backbone Model

Traditional sentiment analysis methods have primarily focused on replacing different deep learning modules to improve performance. However, these approaches often lack interpretability and fail to effectively address the issue of uncertainty estimation caused by different modalities. Therefore, in contrast to previous methods, we will utilize mainstream unimodal baseline models and concentrate on reducing the uncertainty issues arising during the modality learning process. By leveraging well-established unimodal models as the foundation, our approach aims to enhance interpretability while effectively tackling uncertainty estimation problems associated with different modalities. This strategy allows us to build upon existing knowledge and expertise in unimodal sentiment analysis, leading to a more reliable and interpretable multimodal sentiment analysis model.

For convenience, let $\mathcal{T} = \{T_1, T_2, \ldots, T_i, \ldots, T_n\}$ represent the input of n text samples. For each sample T_i, we feed it into a textual backbone model, such as BERT. Then, we extract the pooler output of BERT as the output O_b for the text modality, as shown below:

$$O_b = f_b(T_i; \theta_b), O_b \in \mathbb{R}^Z, \tag{1}$$

where θ_b represents the parameters of the BERT, and Z is the hidden size of BERT. Similarly, let $\mathcal{I} = \{I_1, I_2, \ldots, I_i, \ldots, I_n\}$ represent the input of n image samples. For each image I_i, we utilize a visual baseline model, such as ResNet, to obtain the average pooling output O_r for the visual modality, as shown below:

$$O_r^{'} = f_r(I_i; \theta_r), O_r^{'} \in \mathbb{R}^{H \times W \times D}, \tag{2}$$

where θ_r represents the parameters of the ResNet, H, W, and D, respectively, represent the three dimensions of the pooled output. To obtain a pooled output similar to the text modality, we reshape the result of O_r to obtain the output for the image modality, denoted as O_r. This can be represented as:

$$O_r = reshape\left(O_r^{'}\right), O_r \in \mathbb{R}^L, \tag{3}$$

where $L = H \times W \times D$, and the function of *reshape* represents the transformation of the input tensor into a new shape format. It allows us to convert the three-dimensional output of the image modality into a format that aligns with the one-dimensional output.

3.3. ENN Head

Existing deep learning models typically employ a softmax layer on top of deep neural networks (DNNs) for classification tasks. However, these softmax-based DNNs are unable to estimate the prediction uncertainty of classification problems effectively. This is because softmax scores inherently provide point estimates of the predicted distribution [41], and softmax outputs tend to be overly confident even of mispredictions [42]. Therefore, to address the issues caused by softmax-based approaches, we introduce methods such as Dempster–Shafer Theory (DST) [39] and Subjective Logic (SL) [40], which form the core of the ENN Head.

For convenience, we will uniformly record the output O_b and O_r of the benchmark model of text mode and image mode as O_x. To better evaluate the evidence support information for each modality, we first apply a linear fully connected layer to the outputs of the unimodal baseline models. This can be expressed as follows:

$$O_c = Liner(O_x; \theta_x), O_c \in \mathbb{R}^M, \tag{4}$$

where θ_x represents the parameters of the *Liner*, and M is the number of categories in the multimodal classification task. Different from the traditional softmax-based method, according to the relevant theories of DST and SL, ENN Head generates evidence to support classification through a linear fully connected layer. The evidence refers to the metrics obtained from the input data to support classification. It represents the information or observations that contribute to the classification decision. We can obtain the representation of the evidence through the expression of the relevant non-negative activation function, such as exp:

$$\mathbf{e} = exp(clamp(O_c)), \mathbf{e} \in \mathbb{R}^M, \tag{5}$$

where the $clamp$ function controls the output of the fully connected layer within a range to prevent the impact of numerical offset on classification. Specifically, In the case of an M-class classification problem, Subjective Logic assigns belief mass b_m to each class based on the Dirichlet distribution and assigns overall uncertainty mass u to the entire class framework. The association between b_m and u can be represented by the following equation:

$$u + \sum_{m=1}^{M} b_m = 1, \tag{6}$$

where $u \geq 0$ and $b_m \geq 0$ for m = 1, . . . , M. The belief mass b_m of a single category m is calculated using the evidence e_m of a single category m, and the $e_m \geq 0$. In this way, belief mass b_m and uncertainty u can be easily calculated by setting the following equations:

$$b_m = \frac{e_m}{S} \; and \; u = \frac{M}{S}, \tag{7}$$

where $S = \sum_{m=1}^{M} \beta_m$ represents the Dirichlet strength. Based on the DST and SL theories, β_m is connected to the learned evidence through the equation $\beta_m = e_m + 1$. That is to say, by using the equation $b_m = (\beta_m - 1)/S$, we can easily obtain the subjective opinion from the corresponding parameters of the Dirichlet distribution.

3.4. Hybrid Uncertainty Calibration

3.4.1. EDL Loss

Evidence Deep Learning (EDL) plays a crucial role in enabling the sentiment analysis model to handle situations where the sentiment of a particular sample is ambiguous

or unknown. By incorporating uncertainty into the model, it can effectively "know the unknown", acknowledging cases where sentiment analysis may be challenging due to a lack of clear evidence or conflicting signals. In this paper, let $\mathbf{y}_i = \{y_{i1}, \ldots, y_{iM}\}$ be an one-hot M-dimensional label for sample i with $y_{ij} = 1$ and $y_{iM} = 0$ for all m $\neq$ j. Then, unlike the traditional cross-entropy loss, the Bayesian risk based on the Dirichlet distribution can be represented as the loss function of EDL [38]. The following is the representation of the EDL loss function:

$$\mathcal{L}_{EDL} = \sum_{j=1}^{M} y_{ij}\left(\psi(S_i) - \psi\left(\beta_{ij}\right)\right) + \gamma_t KL(\mathbf{p}, \boldsymbol{\beta}), \tag{8}$$

where $\psi(\cdot)$ is the digamma function, and the second term is Kullback–Leibler (KL) divergence [38]. γ_t is a balancing factor that determines the trade-off between the expected classification error and KL-regularization. $\mathbf{p}$ represents the likelihood of each class in an M-class classification problem, and $\boldsymbol{\beta}$ is the Dirichlet strength vector.

3.4.2. HUC Loss

The calibration of uncertain estimates focuses primarily on calibrating the outputs of the ENN Head. While there have been numerous methods based on EDL that have been explored, accurate quantification of uncertainty estimation from DNNs remains an open research problem, despite recent progress in probabilistic deep learning for improving model robustness. A well-calibrated model should exhibit confidence in its predictions when accurate and display high uncertainty when making inaccurate predictions.

Currently, mainstream calibration methods involve differential approximations to the accuracy versus uncertainty (AvU) [27] defined in Equation (9) as a utility function. This utility function can be computed for a mini-batch of data samples during the model training process. The AvU utility function is optimized to achieve well-calibrated uncertainties, where the model provides lower uncertainty for accurate predictions and higher uncertainty for inaccurate predictions. To estimate the AvU metric during each training step, the outputs within a mini-batch can be grouped into four different categories: (1) accurate and certain (AC), (2) inaccurate and certain (IC), (3) accurate and uncertain (AU), and (4) inaccurate and uncertain (IU). This categorization allows for quantifying the relationship between accuracy and uncertainty, providing insights into the model's calibration performance. This utility function can be calculated using the following formula:

$$\text{AvU} = \frac{n_{\text{AC}} + n_{\text{IU}}}{n_{\text{AC}} + n_{\text{IC}} + n_{\text{AU}} + n_{\text{IU}}}, \tag{9}$$

where the n_{AC}, n_{IC}, n_{AU} and n_{IU} indicate the number of samples under the four prediction scenarios mentioned above.

However, excessive calibration using traditional methods can introduce new issues in uncertainty estimation. Inspired by this observation, we propose to focus on calibrating the accurate and uncertain (AU) and inaccurate and certain (IC) cases. In Figure 3, a toy example is presented to illustrate the four possible EDL outputs. The objective is to calibrate the predictive uncertainty of the EDL model. For accurate predictions, the model is encouraged to learn a skewed and sharp Dirichlet simplex, as shown in Figure 3a. On the other hand, for incorrect predictions, the model should provide an unskewed and flat Dirichlet simplex, as depicted in Figure 3d.

To achieve this calibration, we propose to regularize the EDL training process by minimizing the expectations of the IC and AU cases, as shown in Figure 3b,c, respectively. By minimizing these cases, we can encourage the other two cases (AC and IU) and promote the desired behavior in the model's uncertainty estimation. Consequently, if a sample is assigned a high EDL uncertainty, it is more likely to be incorrect, allowing for the identification of unknown sentiment. By prioritizing the calibration of these cases, we aim to address the specific challenges associated with uncertainty estimation in these scenarios.

This approach allows us to refine the model's uncertainty estimates and improve its overall calibration performance.

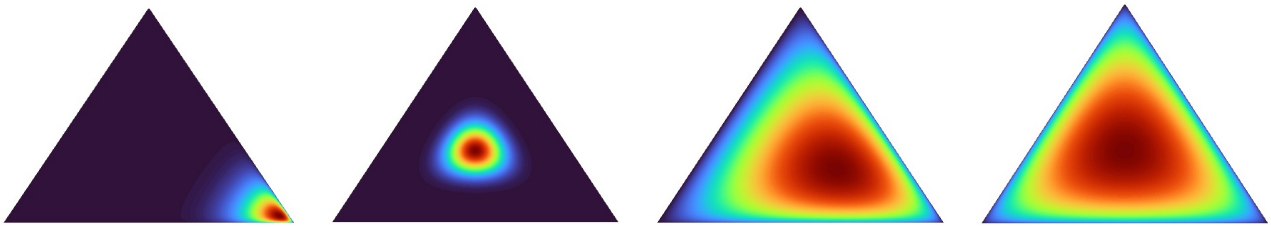


Figure 3. Typical examples of EDL outputs. We set the number of classes as three, and the second class as the correct output. For a well-calibrated model, the ideal prediction scenarios are AC (**a**) and IU (**b**). The IC (**c**) and AU (**d**) cases represent scenarios where the model's calibration needs improvement to reduce uncertainty in accurate predictions and reduce overconfidence in inaccurate predictions. (**a**) AC. (**b**) IC. (**c**) AU. (**d**) IU.

What sets our approach apart is that we propose a hybrid uncertainty calibration (HUC) method, where we calibrate the EDL for both modalities. In this method, we aim to achieve calibration across multiple modalities by considering the unique characteristics and challenges associated with each modality. By combining the calibration efforts for both modalities, we can potentially improve the overall calibration performance of the model. The HUC method takes into account the specific requirements and considerations of each modality to effectively calibrate the uncertainty estimates. Specifically, for text modality, we achieve the HUC method by taking into account the logarithm constraint between the confidence $p_T^{(i)}$ and uncertainty $u_T^{(i)}$:

$$\mathcal{L}_{HUC\_T} = -\zeta_q \sum_{i \in \{\hat{y}_T^{(i)} = y_T^{(i)} \text{ and } \hat{y}_V^{(i)} = y_V^{(i)}\}} p_T^{(i)} \log\left(1 - u_T^{(i)}\right) - (1 - \zeta_q) \sum_{i \in \{\hat{y}_T^{(i)} \neq y_T^{(i)} \text{ or } \hat{y}_V^{(i)} \neq y_V^{(i)}\}} \left(1 - p_T^{(i)}\right) \log\left(u_T^{(i)}\right), \quad (10)$$

where $p_T^{(i)}$ refers to the maximum class probability of an input text sample, and $u_T^{(i)}$ represents the associated evidential uncertainty for that particular text sample. Obviously, for image modalities or other modalities, we can easily obtain HUC calibration methods similar to text modalities, such as $\mathcal{L}_{HUC\_V}$. Here, T and V represent the text modal and image modal, respectively.

In the Hybrid Uncertainty Calibration (HUC) method, we can explain the basis for the two penalties using the example of the text modality. The first penalty aims to provide low uncertainty ($u_T^{(i)} \to 0$) when both modalities of the model make accurate predictions ($\hat{y}_T^{(i)} = y_T^{(i)}$ *and* $\hat{y}_V^{(i)} = y_V^{(i)}, p_T^{(i)} \to 1$). The second penalty aims to provide high uncertainty ($u_T^{(i)} \to 1$) when at least one modality of the model makes an inaccurate prediction ($\hat{y}_T^{(i)} \neq y_T^{(i)}$ *or* $\hat{y}_V^{(i)} \neq y_V^{(i)}, p_T^{(i)} \to 0$). The significance of these constraints is to make the uncertainty of both modalities decrease as much as possible.

To better balance the importance of accurate and inaccurate predictions during different stages of model training, we introduce an annealing factor [43] $\zeta_q = \zeta_0 \, exp\{-(ln\zeta_0 \, / Q)q\}$, where $\zeta_q \in [\zeta_0, 1]$. In the early stages of training, inaccurate predictions dominate, so we set ζ_0 to a small positive constant, and thus ζ_q is also a small value. As a result, the second penalty for IC loss receives more punishment. As the training progresses from the initial stage to the total number of training periods Q, the ζ_q factor increases exponentially from ζ_0 to 1.0. In other words, in the later stages of training, accurate predictions from both modalities become dominant, and therefore, the first penalty for AU loss should receive a larger punishment. This annealing factor helps to dynamically adjust the balance between

the penalties based on the training stage, allowing for effective uncertainty calibration across different modalities.

3.5. Uncertainty-Aware Late Fusion

Through feature extraction using a single-modality baseline model and the hybrid uncertainty calibration in the training process, the current objective is to provide reliable fusion, especially for low-quality data. As mentioned in the previous sections, the uncertainty exhibited by different modalities' data are inconsistent. Traditional fusion methods may overlook the problem of widely existing modality uncertainty, thus failing to effectively address the model's uncertainty.

The Uncertainty-aware Late Fusion (ULF) algorithm follows the flow described in Algorithm 1. Unlike previous methods, we combine uncertainty estimation methods to fuse the outputs of the ENN heads of the two modalities at the decision level. Specifically, considering that the quality of different modalities may vary, we introduce the uncertainty estimation of each modality to guide the dynamic fusion of information from different branches. Therefore, if both modalities make the same prediction, the decision fusion mechanism selects that prediction as the final one. If one modality predicts neutrality, the decision fusion mechanism selects the prediction from the other modality. Lastly, if neither of the above cases applies, it means the predictions from the two modalities are conflicting, i.e., one is positive and the other is negative. In this case, we utilize uncertainty estimation to guide the fusion mechanism, selecting the prediction from the modality with lower uncertainty as the final result.

Algorithm 1 Algorithm of Uncertainty-aware Late Fusion (ULF)

Input: The text sample T_i and the image sample I_i.
Output: The classification output Y_i for sample i.

1: Obtain the ${O_c}^{(T_i)}$ and ${O_c}^{(I_i)}$ for the text modality and image modality classifiers according to Equation (4);
2: Obtain the uncertainty estimates $u^{(T_i)}$ and $u^{(I_i)}$ for the text modality and image modality, respectively, according to Equation (7);
3: **if** ${O_c}^{(T_i)} == {O_c}^{(I_i)}$ **then**
4: $\quad Y_i = {O_c}^{(T_i)}$
5: **else if** ${O_c}^{(T_i)}$ == "neutral" **then**
6: $\quad Y_i = {O_c}^{(I_i)}$
7: **else if** ${O_c}^{(I_i)}$ == "neutral" **then**
8: $\quad Y_i = {O_c}^{(T_i)}$
9: **else if** $u^{(T_i)} < u^{(I_i)}$ **then**
10: $\quad Y_i = {O_c}^{(T_i)}$
11: **else**
12: $\quad Y_i = {O_c}^{(I_i)}$
13: **end if**
14: **return** Y_i

In Line 1, the outputs of the unimodal models are obtained based on Equation (4). In Line 2, the uncertain outputs of each modality are obtained based on Equation (7). Lines 3 to 13 describe the execution process of the newly proposed late fusion strategy. In Line 14, the final classification output Y_i is returned. It is evident that the time complexity of Algorithm 1 primarily depends on the classification output of the unimodal models, which is related to the model's parameters. It is worth noting that in the process of hybrid uncertainty calibration, we have focused on reducing the uncertainties of both modalities. Therefore, the uncertainty-aware late fusion method is theoretically feasible, and we will validate the effectiveness of this method in the experimental phase. Furthermore, our algorithm can serve as a guideline for future research on fusion problems based on uncertainty estimation.

4. Experiments

4.1. Experiment Setups

4.1.1. Datasets

We assess our model's performance using three publicly available multimodal sentiment datasets: MVSA-Single, MVSA-Multiple [8] and MVSA-Single-Small [44]. MVSA-Single comprises 5129 samples annotated by a single annotator, while MVSA-Multiple consists of 19,600 samples annotated by three annotators. The MVSA-Single-Small dataset is derived from the MVSA-Single dataset using the same method as [44]. Each sample in both datasets represents a tweet that includes a text–image pair collected from Twitter. To ensure a fair comparison, we preprocess the original MVSA datasets following the approach used in [9]. This involves removing noisy tweets where the textual label and visual label do not align. We randomly divide the datasets into training, development, and test sets using an 8:1:1 split ratio. For the MVSA-Single-Small dataset, we followed the same division strategy described in [44,45], resulting in the training set, validation set, and test set consisting of 1555, 518 and 519 image-text pairs, respectively. Therefore, we have obtained the sample counts for different sentiment categories in three MVSA datasets, as shown in Table 1.

Table 1. Sentiment categories of the processed MVSA datasets.

Dataset	Positive	Neutral	Negative	Total
MVSA-Single	2683	470	1358	4511
MVSA-Multiple	11,318	4408	1298	17,024
MVSA-Single-Small	1398	470	724	2592

4.1.2. Evaluation Metrics

In our research, to comprehensively evaluate the performance of our model, we utilize accuracy and weighted F1 as evaluation metrics, which are widely used in sentiment analysis tasks. Weighted F1 refers to a metric that calculates the F1 score by considering the class imbalance in the dataset [10,46]. It assigns higher importance to minority classes to address the potential bias caused by imbalanced class distribution. Additionally, to assess the calibration and uncertainty calibration of the model, we adopt the Expected Uncertainty Calibration Error (UCE) as a measure, following the approach used in previous studies [27,47]. UCE is introduced to measure the miscalibration of uncertainty and represents the expected difference between model error and uncertainty. These metrics provide insights into the calibration and uncertainty calibration performance of our model.

UCE is a metric that quantifies the expected difference between the model's error and its uncertainty. UCE aims to assess how well the model's predicted uncertainty aligns with its actual error. The predictions of the neural network are partitioned into J bins of equal width, where the jth $= \left(\frac{j-1}{J}, \frac{j}{J}\right]$ represents the interval. UCE is defined by Equation 11 in the referenced paper [47], where N represents the total number of samples, B_j represents the index set of samples with predicted confidence in the interval jth:

$$UCE = \sum_{j=1}^{J} \frac{|B_j|}{N} \left| err(B_j) - uncert(B_j) \right|, \tag{11}$$

where the model error and uncertainty for each bin are defined as follows:

$$err(B_j) = \frac{1}{|B_j|} \sum_{i \in B_j} 1(\hat{y}_i \neq y_i) \quad \text{and} \quad uncert(B_j) = \frac{1}{|B_j|} \sum_{i \in B_j} u_i, u_i \in [0,1]. \tag{12}$$

4.1.3. Implementation Details

To train our network and achieve better performance, we have chosen the Adam optimization algorithm as our optimizer. For this sentiment analysis task, we set the optimal learning rate for Adam to 5×10^{-5}. We trained our model for 50 epochs, with a batch size of 32. Additionally, we set the dropout rate to 0.5 and set the gradient accumulation steps to 8. All the experiments are performed on an AI Lab online server with the following specifications: three 24 G Nvidia GeForce RTX 3090 graphics cards with 12 vCPU Intel(R) Xeon(R) Platinum 8255C CPU @ 2.50 GHz, 43 GB of memory, Alibaba Cloud Computing, Hangzhou, China.

4.2. Comparison with Existing Methods

4.2.1. Comparative Methods

We compare our proposed method with different approaches, including single-modal models, multi-modal models, and uncertainty estimation methods. For the text modality, Bag-of-Words (BoW) [48] is a model that originated in the fields of Natural Language Processing and Information Retrieval. BERT [36], on the other hand, is a pre-trained text model that we have fine-tuned for our task. Regarding the image modality, ResNet [37] is a popular and strong single-modal baseline model for image classification tasks.

Similar to [44,45], ConcatBow and ConcatBert involve concatenating the outputs of the image-based model with the outputs of BoW and BERT, respectively. On the other hand, Late Fusion takes the average of the predictions from the image classifier and the BERT model. MultiSentiNet [9] is a model that extracts object and scene information from images as visual semantic features. HSAN is indeed a model based on image captions for multi-modal sentiment analysis [6]. Co-MN-Hop6 [7] proposes a co-memory network that models the mutual influences between images and text iteratively. CFF-ATT [11] introduces a multimodal cross-feature fusion model that is based on attention mechanisms. Sentiment Multi-Layer Neural Network (Se-MLNN) [10] combines multiple visual features and contextual text features to accurately predict overall sentiment. MLFC-SCSupCon [46] introduces the MLFC module, which combines a convolutional neural network (CNN) and a Transformer to address the redundancy problem and reduce irrelevant information in each modality's features. Moreover, MLFC-SCSupCon employs supervised contrastive learning to enhance further its ability to learn standard sentiment features from the data.

To demonstrate the effectiveness of our uncertainty fusion method, we compare it with several common uncertainty estimation methods. Trusted Multi-view Classification (TMC) [49] and ETMC [50] are multi-view classification methods based on the Dempster–Shafer Theory (DST). They provide reliable ensembles and decision interpretability. Quality-aware Multimodal Fusion (QMF) [44] is a quality-aware multimodal fusion framework that improves performance in terms of classification accuracy and model robustness. Furthermore, we also compare our approach with various common uncertainty calibration methods. Accuracy versus Uncertainty Calibration (AvUC) [27] enables a model to learn not only improved accuracy but also well-calibrated uncertainties. Evidential Uncertainty Calibration (EUC) [51] encourages the model to assign lower uncertainties to confident predictions and higher uncertainties to less confident predictions.

4.2.2. Results and Analysis

Table 2 presents a performance comparison between our HUC-ULF model and the baseline methods.

Table 2. The results of different methods on MVSA-Single and MVSA-Multiple datasets. For the data marked with an asterisk (*), it indicates that the method's data does not explicitly state whether weighted F1 is used or not.

Modality	Model	MVSA-Single		MVSA-Multiple	
		Acc (%)	F1 (%)	Acc (%)	F1 (%)
Unimodal	BoW	51.00	48.13	66.00	59.88
	BERT	71.11	69.70	67.59	66.24
	ResNet	66.08	64.32	67.88	61.30
Multimodal	MultiSentiNet	69.84	69.84 *	68.86	68.11 *
	HSAN	69.88	66.90 *	67.96	67.76 *
	Co-MN-Hop6	70.51	70.01 *	68.92	68.83 *
	CFF-ATT	71.44	71.06 *	69.62	69.35 *
	ConcatBow	61.64	60.81	68.06	63.32
	ConcatBert	68.51	67.84	69.65	65.75
	Late fusion	74.28	73.16	69.29	65.81
	Se-MLNN	75.33	73.76	66.35	61.89
	MLFC-SCSupCon	76.44	75.61	70.53	67.97
	Ours	**77.61**	**76.59**	**72.06**	**68.83**

We use weighted F1 and ACC as evaluation metrics, following [46] in MVSA-Single and MVSA-Multiple. We made the following observations. Firstly, our model is competitive with other strong baseline models on all three datasets. Secondly, the multimodal model outperforms the unimodal models on all three datasets. Our model outperforms the current state-of-the-art (SOTA) model on the MVSA-Single dataset, with an improvement of 1.17% in ACC and 0.98% in weighted F1. Similarly, our model surpasses the SOTA model on the MVSA-Multiple dataset, with an improvement of 1.53% in ACC and 0.86% in weighted F1.

To further demonstrate the effectiveness of our proposed model compared to other uncertainty estimation methods, we conducted experiments on the same dataset splits as [44], as shown in Table 3. The results show that our model outperforms the current SOTA model on the MVSA-Single-Small dataset, with an improvement of 1.31% in ACC and 1.2% in weighted F1. Overall, our model is comparable to the SOTA models. Specifically, the F1 values of the methods marked with an asterisk (*) should be considered for reference only. This is because some of these methods do not provide a specific formula for calculating F1, and others may not use the weighted F1 measure.

Table 3. For the MVSA-Single-Small dataset, the table displays the ACC and F1 performance (%) of the designed model evaluation metrics.

Modality	Model	MVSA-Single-Small	
		Acc (%)	F1 (%)
Unimodal	BoW	66.67	64.59
	BERT	74.53	73.15
	ResNet	66.08	62.21
Multimodal	ConcatBow	65.32	64.32
	ConcatBert	66.15	65.02
	Late fusion	74.92	74.07
	TMC	75.98	75.21
	ETMC	76.11	75.34
	QMF	78.07	76.86
	Ours	**79.38**	**78.06**

We also observed in the experiments of Tables 2 and 3 that the ConcatBow and ConcatBert fusion methods yield even lower accuracy compared to the unimodal models.

This indicates that improper multimodal fusion methods may introduce additional noise to the model, leading to a decrease in accuracy, whereas our model proves to be effective.

To validate the robustness of HUC, we evaluated HUC and comparative methods based on Uncertainty Calibration Error (UCE) under Gaussian noise (for image modality) and blank noise (for text modality), following previous works [52–55]. The experimental results are shown in Figure 4.

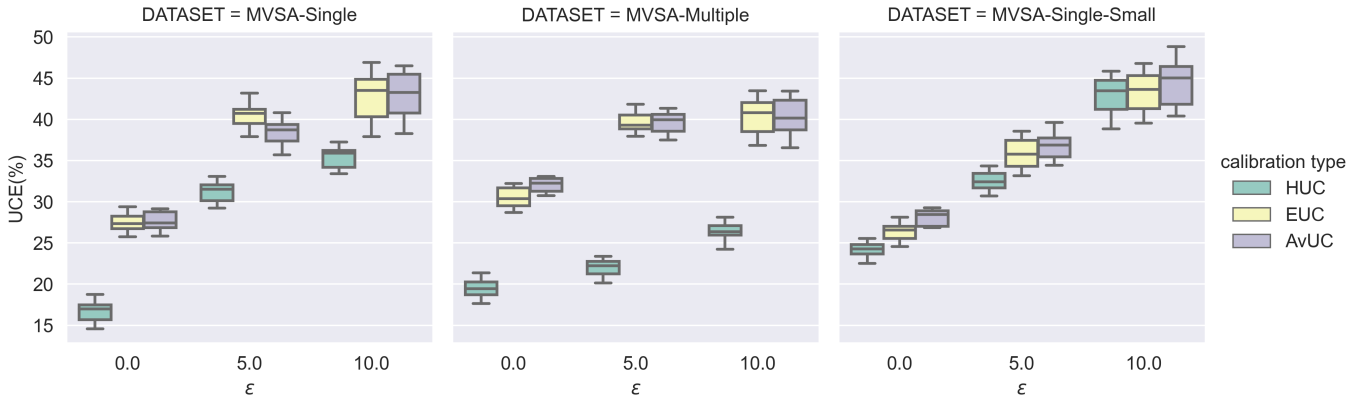


Figure 4. The comparison of Expected Uncertainty Calibration Error (UCE) for three MVSA datasets at different Gaussian noise, i.e., zero mean with variance of ε.

UCE represents the calibration error of model predictions, and a reliable and well-calibrated model should provide lower calibration error, even with increased intensity of data shift, although accuracy might decrease with data shift. From Figure 4, we observe that HUC achieves lower UCE calibration error than all the methods under different noise intensities. This indicates that HUC exhibits better generalization in the experiments.

It is worth noting that we also conducted experiments with different types of noise, such as Salt–Pepper noise, as shown in Figure 5. The results demonstrate that HUC outperforms the existing state-of-the-art methods (i.e., EUC and AvUC), highlighting the superiority of our approach. Please note that when ε is 0.0, it means that no noise was added to the test set. When ε takes values of 5.0 and 10.0, our method UCE exhibits lower average values and biases on all three MVSA datasets compared to the contrastive methods. This is especially evident in the low-quality MVSA-Multiple dataset, demonstrating the superior generalization ability of our proposed method even in the presence of low-quality multimodal data.

Figure 5. The comparison of UCE for three MVSA datasets at Salt–Pepper noise with varying noise rate ε.

4.3. Ablation Study

To further validate the effectiveness of the proposed ULF-HUC model, we conducted two ablation experiments on the three MVSA datasets in this section. Firstly, we compared different component combinations (ULF and HUC) as shown in Table 4.

Table 4. Ablation results of our ULF-HUC.

ULF	$\mathcal{L}_{HUC}$	MVSA-Single		MVSA-Multiple		MVSA-Single-Small	
		Acc (%)	F1 (%)	Acc (%)	F1 (%)	Acc (%)	F1 (%)
✗	✗	74.28	73.16	69.29	65.81	74.92	74.07
✗	✔	75.52	74.71	70.35	66.25	75.26	74.66
✔	✗	75.39	74.80	70.29	66.08	75.14	74.21
✔	✔	**77.61**	**76.59**	**72.06**	**68.83**	**79.38**	**78.06**

The symbol (✗) indicates that the component is not used, while (✔) indicates that the component is used. When neither component is used, it represents the late fusion method. The ablation experiments demonstrate that the classification ability of the model improves when ULF and HUC are added separately. The model achieves the best-expected performance when ULF and HUC are combined in their entirety. The results demonstrate that the fusion method of ULF-HUC makes the model more valuable, and the combination of HUC and ULF helps increase the reliability of model fusion, reduce model uncertainty, and improve overall model performance.

Next, we conducted experiments on different noise patterns by comparing ULF-HUC with models without HUC. The results of the ablation experiments are shown in Table 5. Our model with HUC exhibits reduced UCE calibration error under different noise intensities. All experimental metrics show improvement, validating the effectiveness of the model.

Table 5. Expected Uncertainty Calibration Error (UCE) results at different noise types with varying ε. Small UCE(%) indicates the model is better calibrated.

Noise Type	Model	MVSA-Single			MVSA-Multiple			MVSA-Single-Small		
		ε = 0.0	ε = 5.0	ε = 10.0	ε = 0.0	ε = 5.0	ε = 10.0	ε = 0.0	ε = 5.0	ε = 10.0
Gaussian	ULF-HUC (w/o $\mathcal{L}_{HUC}$)	27.03	42.57	45.38	31.55	38.87	41.72	26.46	35.56	46.47
	ULF-HUC (full)	**16.69**	**31.21**	**35.38**	**19.48**	**21.95**	**26.14**	**24.16**	**32.51**	**42.82**
Salt-Pepper	ULF-HUC (w/o $\mathcal{L}_{HUC}$)	27.03	38.74	47.62	31.55	36.71	42.09	26.46	32.71	46.42
	ULF-HUC (full)	**16.69**	**30.01**	**35.32**	**19.48**	**21.70**	**25.19**	**24.16**	**30.00**	**41.23**

4.4. Case Study

To further demonstrate the effectiveness of our model, we provide a case study where we compare the predicted sentiment labels based on the ULF-HUC model and the model without HUC. As shown in Figure 6, we can observe that it is not easy to accurately analyze the user's sentiment tendency in sentiment analysis tasks when the sentiment polarity of the text and image is inconsistent.

For example, in the first data example in Figure 6, the image depicts a neutral sentiment, while the text expresses a positive sentiment. Our model with HUC can correctly fuse the modalities and make accurate predictions. In the second data example, we find that the image is negative while the text is neutral, and when HUC is removed, the model's prediction becomes incorrect. Similarly, our model handles the problem of inconsistent sentiment polarity between the two modalities in the third data example and makes the correct judgment.

The case study in this section demonstrates that our model is better able to calibrate uncertainty estimation, thereby improving the accuracy of predictions.

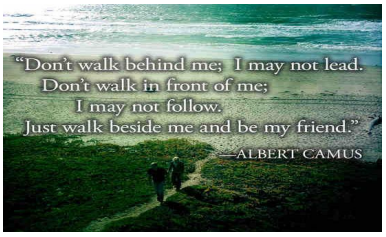
Image	Text	ULF-HUC (Full)	ULF-HUC (w/o $\mathcal{L}_{HUC}$)
	"Empty looks good enough :P"	positive	negative
	"Why do you get upset so much? Take the quiz: "	negative	positive
	"Let's just say the other team got wrecked "	negative	positive

Figure 6. Examples of misclassified by ULF-HUC (w/o $\mathcal{L}_{HUC}$) and correctly classified by ULF-HUC (full).

5. Conclusions

To address the issue of traditional multimodal sentiment analysis methods being unable to effectively solve the uncertainty estimation problem among different modalities, we propose an uncertain-aware late fusion method based on hybrid uncertainty calibration (ULF-HUC). The core idea of this paper is to introduce a late fusion strategy based on uncertainty estimation and then use hybrid uncertainty calibration to learn the sentimental features of the two modalities. To successfully implement this core idea, we propose a series of methods. Firstly, we conduct an in-depth analysis of the sentiment polarity distribution in sentiment analysis datasets. Secondly, to minimize the high uncertainty caused by inconsistent sentiment polarities in different modalities, we propose a fusion strategy based on uncertainty estimation. Next, to achieve a balance between model accuracy and uncertainty, we use a learning method with hybrid uncertainty calibration, effectively reducing uncertainty when the model is accurate and reducing certainty when the model is inaccurate. Finally, we add different types of noise (namely Gaussian noise and Salt–Pepper noise) to verify the model's classification and calibration capabilities. Experimental results show that our proposed ULF-HUC method overcomes the limitations of unimodal models and improves performance after fusion. Additionally, our method outperforms the comparison methods in terms of classification performance and calibration performance on three MVSA datasets, improving evaluation metrics such as accuracy, weighted F1, and expected uncertainty calibration error (UCE).

This research work has the following limitations: (1) The study focuses on multimodal sentiment analysis. (2) The impact of noise on the model's performance and how to mitigate its effects is a relevant and worthy topic for further exploration.

In the future, to address the issue of disparate learning capabilities among different modalities, we will consider methods that are more suitable for calibrating modality learning capabilities in existing multimodal fusion strategies. Additionally, we will explore

new methods for uncertainty calibration and consider the challenges of accuracy and uncertainty estimation calibration brought by more complex multimodal fusion.

Author Contributions: Conceptualization, Q.P. and Z.M.; methodology, Q.P.; software, Q.P.; validation, Q.P.; formal analysis, Q.P. and Z.M.; investigation, Q.P.; resources, Q.P.; data curation, Q.P.; writing—original draft preparation, Q.P.; writing—review and editing, Q.P. and Z.M.; visualization, Q.P.; supervision, Z.M.; project administration, Q.P.; funding acquisition, Z.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China under Grant 62266004 and 61762009.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding authors. The data are not publicly available due to privacy.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Mercha, E.M.; Benbrahim, H. Machine learning and deep learning for sentiment analysis across languages: A survey. *Neurocomputing* **2023**, *531*, 195–216. [CrossRef]
2. Zad, S.; Heidari, M.; Jones, J.H.; Uzuner, O. A survey on concept-level sentiment analysis techniques of textual data. In Proceedings of the 2021 IEEE World AI IoT Congress (AIIoT), Seattle, WA, USA, 10–13 May 2021; pp. 0285–0291.
3. Das, R.; Singh, T.D. Multimodal Sentiment Analysis: A Survey of Methods, Trends, and Challenges. *ACM Comput. Surv.* **2023**, *55*, 1–38. [CrossRef]
4. Gandhi, A.; Adhvaryu, K.; Poria, S.; Cambria, E.; Hussain, A. Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions. *Inform. Fusion* **2023**, *91*, 424–444. [CrossRef]
5. Amrani, E.; Ben-Ari, R.; Rotman, D.; Bronstein, A. Noise Estimation Using Density Estimation for Self-Supervised Multimodal Learning. *Proc. AAAI Conf. Artif. Intell.* **2021**, *35*, 6644–6652. [CrossRef]
6. Xu, N. Analyzing multimodal public sentiment based on hierarchical semantic attentional network. In Proceedings of the 2017 IEEE International Conference on Intelligence and Security Informatics (ISI), Beijing, China, 22–24 July 2017; pp. 152–154.
7. Xu, N.; Mao, W.; Chen, G. A co-memory network for multimodal sentiment analysis. In Proceedings of the The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, Ann Arbor, MI, USA, 8–12 July 2018; pp. 929–932.
8. Niu, T.; Zhu, S.; Pang, L.; El Saddik, A. Sentiment analysis on multi-view social data. In Proceedings of the MultiMedia Modeling: 22nd International Conference, MMM 2016, Miami, FL, USA, 4–6 January 2016; Proceedings, Part II 22; Springer: Berlin/Heidelberg, Germany, 2016; pp. 15–27.
9. Xu, N.; Mao, W. Multisentinet: A deep semantic network for multimodal sentiment analysis. In Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, Singapore, 6–10 November 2017; pp. 2399–2402.
10. Cheema, G.S.; Hakimov, S.; Müller-Budack, E.; Ewerth, R. A fair and comprehensive comparison of multimodal tweet sentiment analysis methods. In Proceedings of the 2021 Workshop on Multi-Modal Pre-Training for Multimedia Understanding, Taipei, China, 16–19 November 2021; pp. 37–45.
11. Zhang, K.; Geng, Y.; Zhao, J.; Liu, J.; Li, W. Sentiment Analysis of Social Media via Multimodal Feature Fusion. *Symmetry* **2020**, *12*, 2010. [CrossRef]
12. Tomani, C.; Cremers, D.; Buettner, F. Parameterized temperature scaling for boosting the expressive power in post-hoc uncertainty calibration. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 24–28 October 2022 ; Springer: Berlin/Heidelberg, Germany, 2022; pp. 555–569.
13. Zhuang, D.; Bu, Y.; Wang, G.; Wang, S.; Zhao, J. SAUC: Sparsity-Aware Uncertainty Calibration for Spatiotemporal Prediction with Graph Neural Networks. In Proceedings of the Temporal Graph Learning Workshop@ NeurIPS 2023, New Orleans, LA, USA, 10–16 December 2023.
14. Xu, J.; Huang, F.; Zhang, X.; Wang, S.; Li, C.; Li, Z.; He, Y. Visual-textual sentiment classification with bi-directional multi-level attention networks. *Knowl.-Based Syst.* **2019**, *178*, 61–73. [CrossRef]
15. Cholet, S.; Paugam-Moisy, H.; Regis, S. Bidirectional Associative Memory for Multimodal Fusion: A Depression Evaluation Case Study. In Proceedings of the 2019 International Joint Conference on Neural Networks (IJCNN), Budapest, Hungary, 14–19 July 2019. [CrossRef]
16. Kumar, A.; Srinivasan, K.; Cheng, W.H.; Zomaya, A.Y. Hybrid context enriched deep learning model for fine-grained sentiment analysis in textual and visual semiotic modality social data. *Inf. Process. Manag.* **2020**, *57*, 102141. [CrossRef]

17. Jiang, T.; Wang, J.; Liu, Z.; Ling, Y. Fusion-extraction network for multimodal sentiment analysis. In Proceedings of the Advances in Knowledge Discovery and Data Mining: 24th Pacific-Asia Conference, PAKDD 2020, Singapore, 11–14 May 2020; Proceedings, Part II 24; Springer: Berlin/Heidelberg, Germany, 2020; pp. 785–797.
18. Zhang, K.; Zhu, Y.; Zhang, W.; Zhu, Y. Cross-modal image sentiment analysis via deep correlation of textual semantic. *Knowl.-Based Syst.* **2021**, *216*, 106803. [CrossRef]
19. Guo, W.; Zhang, Y.; Cai, X.; Meng, L.; Yang, J.; Yuan, X. LD-MAN: Layout-Driven Multimodal Attention Network for Online News Sentiment Recognition. *IEEE Trans. Multimed.* **2021**, *23*, 1785–1798. [CrossRef]
20. Liao, W.; Zeng, B.; Liu, J.; Wei, P.; Fang, J. Image-text interaction graph neural network for image-text sentiment analysis. *Appl. Intell.* **2022**, *52*, 11184–11198. [CrossRef]
21. Ye, J.; Zhou, J.; Tian, J.; Wang, R.; Zhou, J.; Gui, T.; Zhang, Q.; Huang, X. Sentiment-aware multimodal pre-training for multimodal sentiment analysis. *Knowl.-Based Syst.* **2022**, *258*, 110021. [CrossRef]
22. Zeng, J.; Zhou, J.; Huang, C. Exploring Semantic Relations for Social Media Sentiment Analysis. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2023**, *31*, 2382–2394. [CrossRef]
23. Liu, Y.; Li, Z.; Zhou, K.; Zhang, L.; Li, L.; Tian, P.; Shen, S. Scanning, attention, and reasoning multimodal content for sentiment analysis. *Knowl.-Based Syst.* **2023**, *268*, 110467. [CrossRef]
24. Gawlikowski, J.; Tassi, C.R.N.; Ali, M.; Lee, J.; Humt, M.; Feng, J.; Kruspe, A.; Triebel, R.; Jung, P.; Roscher, R.; et al. A survey of uncertainty in deep neural networks. *Artif. Intell. Rev.* **2023**, *56*, 1513–1589. [CrossRef]
25. Minderer, M.; Djolonga, J.; Romijnders, R.; Hubis, F.; Zhai, X.; Houlsby, N.; Tran, D.; Lucic, M. Revisiting the calibration of modern neural networks. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 15682–15694.
26. Pakdaman Naeini, M.; Cooper, G.; Hauskrecht, M. Obtaining Well Calibrated Probabilities Using Bayesian Binning. *Proc. AAAI Conf. Artif. Intell.* **2015**, *29*, 2901–2907.
27. Krishnan, R.; Tickoo, O. Improving model calibration with accuracy versus uncertainty optimization. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 18237–18248.
28. Tomani, C.; Gruber, S.; Erdem, M.E.; Cremers, D.; Buettner, F. Post-hoc Uncertainty Calibration for Domain Drift Scenarios. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021. [CrossRef]
29. Hubschneider, C.; Hutmacher, R.; Zollner, J.M. Calibrating Uncertainty Models for Steering Angle Estimation. In Proceedings of the 2019 IEEE Intelligent Transportation Systems Conference (ITSC), Auckland, New Zealand, 27–30 October 2019. [CrossRef]
30. Zhang, H.M.Q.; Zhang, C.; Wu, B.; Fu, H.; Zhou, J.T.; Hu, Q. Calibrating Multimodal Learning. *arXiv* **2023**, arXiv:2306.01265.
31. Tellamekala, M.K.; Amiriparian, S.; Schuller, B.W.; André, E.; Giesbrecht, T.; Valstar, M. COLD Fusion: Calibrated and Ordinal Latent Distribution Fusion for Uncertainty-Aware Multimodal Emotion Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *46*, 805–822. [CrossRef] [PubMed]
32. Kose, N.; Krishnan, R.; Dhamasia, A.; Tickoo, O.; Paulitsch, M. Reliable Multimodal Trajectory Prediction via Error Aligned Uncertainty Optimization. In Proceedings of the Computer Vision—ECCV 2022 Workshops, Tel Aviv, Israel, 24–28 October 2022 ; Springer Nature: Cham, Switzerland, 2023; pp. 443–458. [CrossRef]
33. Folgado, D.; Barandas, M.; Famiglini, L.; Santos, R.; Cabitza, F.; Gamboa, H. Explainability meets uncertainty quantification: Insights from feature-based model fusion on multimodal time series. *Inform. Fusion* **2023**, *100*, 101955. [CrossRef]
34. Wang, R.; Liu, X.; Hao, F.; Chen, X.; Li, X.; Wang, C.; Niu, D.; Li, M.; Wu, Y. Ada-CCFNet: Classification of multimodal direct immunofluorescence images for membranous nephropathy via adaptive weighted confidence calibration fusion network. *Eng. Appl. Artif. Intel.* **2023**, *117*, 105637. [CrossRef]
35. Peng, X.; Wei, Y.; Deng, A.; Wang, D.; Hu, D. Balanced Multimodal Learning via On-the-fly Gradient Modulation. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 21–24 June 2022.
36. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
37. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016 ; pp. 770–778.
38. Sensoy, M.; Kaplan, L.; Kandemir, M. Evidential deep learning to quantify classification uncertainty. *Adv. Neural Inf. Process. Syst.* **2018**, *31*, 1–31.
39. Sentz, K.; Ferson, S. *Combination of Evidence in Dempster-Shafer Theory*; Sandia Nat. Lab.: Albuquerque, NM, USA, 2002. [CrossRef]
40. Jøsang, A. *Subjective Logic*; Springer: Berlin/Heidelberg, Germany, 2016; Volume 3.
41. Gal, Y. Uncertainty in Deep Learning. Ph.D. Thesis, Department of Engineering, University of Cambridge, Cambridge, UK, 2016.
42. Guo, C.; Pleiss, G.; Sun, Y.; Weinberger, K.Q. On calibration of modern neural networks. In *Proceedings of the International Conference on Machine Learning*; PMLR: New York, NY, USA, 2017; pp. 1321–1330.
43. Rere, L.R.; Fanany, M.I.; Arymurthy, A.M. Simulated Annealing Algorithm for Deep Learning. *Procedia Comput. Sci.* **2015**, *72*, 137–144. [CrossRef]
44. Zhang, Q.; Wu, H.; Zhang, C.; Hu, Q.; Fu, H.; Zhou, J.T.; Peng, X. Provable Dynamic Fusion for Low-Quality Multimodal Data. *arXiv* **2023**, arXiv:2306.02050.
45. Kiela, D.; Bhooshan, S.; Firooz, H.; Perez, E.; Testuggine, D. Supervised multimodal bitransformers for classifying images and text. *arXiv* **2019**, arXiv:1909.02950.

46. Wang, H.; Li, X.; Ren, Z.; Wang, M.; Ma, C. Multimodal Sentiment Analysis Representations Learning via Contrastive Learning with Condense Attention Fusion. *Sensors* **2023**, *23*, 2679. [CrossRef]
47. Laves, M.H.; Ihler, S.; Kortmann, K.P.; Ortmaier, T. Well-calibrated model uncertainty with temperature scaling for dropout variational inference. *arXiv* **2019**, arXiv:1909.13550.
48. Zhang, Y.; Jin, R.; Zhou, Z.H. Understanding bag-of-words model: A statistical framework. *Int. J. Mach. Learn Cyb.* **2010**, *1*, 43–52. [CrossRef]
49. Han, Z.; Zhang, C.; Fu, H.; Zhou, J.T. Trusted multi-view classification. *arXiv* **2021**, arXiv:2102.02051.
50. Han, Z.; Zhang, C.; Fu, H.; Zhou, J.T. Trusted Multi-View Classification With Dynamic Evidential Fusion. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 2551–2566. [CrossRef]
51. Bao, W.; Yu, Q.; Kong, Y. Evidential deep learning for open set action recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 13349–13358.
52. Ma, H.; Han, Z.; Zhang, C.; Fu, H.; Zhou, J.T.; Hu, Q. Trustworthy multimodal regression with mixture of normal-inverse gamma distributions. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 6881–6893.
53. Verma, V.; Qu, M.; Kawaguchi, K.; Lamb, A.; Bengio, Y.; Kannala, J.; Tang, J. Graphmix: Improved training of gnns for semi-supervised learning. In Proceedings of the AAAI Conference on Artificial Intelligence, Online, 2–9 February 2021 ; Volume 35, pp. 10024–10032.
54. Hu, Z.; Tan, B.; Salakhutdinov, R.R.; Mitchell, T.M.; Xing, E.P. Learning data manipulation for augmentation and weighting. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 1–12.
55. Xie, Z.; Wang, S.I.; Li, J.; Lévy, D.; Nie, A.; Jurafsky, D.; Ng, A.Y. Data noising as smoothing in neural network language models. *arXiv* **2017**, arXiv:1703.02573.