Adaptive Evidential Fusion for Hateful Meme Classification Utilizing the Dempster–Shafer Theory

Soukaina Fatimi
Machine Intelligence Laboratory (LIM), Faculty of Sciences and Techniques, Mohammedia, Hassan II University of Casablanca, Morocco
soukaina.fatimi1-etu@etu.univh2c.ma (corresponding author)

Wafae Sabbar
Machine Intelligence Laboratory (LIM), Faculty of Sciences and Techniques, Mohammedia, Hassan II University of Casablanca, Morocco
swafae@gmail.com

Abdelkrim Bekkhoucha
Machine Intelligence Laboratory (LIM), Faculty of Sciences and Techniques, Mohammedia, Hassan II University of Casablanca, Morocco
abekkhoucha@yahoo.fr

Received: 1 November 2025 | Revised: 18 December 2025 | Accepted: 6 January 2026

 DOI: https://doi.org/10.48084/etasr.15931

ABSTRACT

Younger generations now frequently communicate through memes, which combine images and text to express humor, emotions, or opinions. However, memes can become a serious problem on social media when they promote hateful, discriminatory, or offensive content. Detecting such hostile memes remains particularly challenging due to the complex interaction between visual and textual cues. In this work, we propose a simple yet effective multimodal fusion approach, named Multimodal Hateful Meme Classification via Dempster–Shafer Evidence Theory Fusion (MHM-DS), for hateful meme detection. Instead of training a large vision–language model, we perform late fusion of independent unimodal classifiers— Bidirectional Encoder Representations from Transformers (BERT) for text and Contrastive Language–Image Pretraining (CLIP) for images—by combining their probabilistic outputs using Dempster–Shafer Evidence Theory (DST) evidential reasoning. The proposed method explicitly models uncertainty and conflict between modalities through belief masses and an ignorance term. Experiments conducted on the Facebook AI Hateful Memes dataset (10,000 samples) show that the proposed DST-based fusion achieves 70.2% accuracy and a 70.8% Area Under the Receiver Operating Characteristic Curve (AUROC), outperforming standard late-fusion baselines and unimodal models, while remaining computationally efficient and interpretable. These results demonstrate that evidential fusion provides a robust and uncertainty-aware alternative to complex multimodal transformers for hateful meme classification.

Keywords-multimodal classification; Dempster–Shafer Evidence Theory (DST); late fusion; hateful meme detection

I. INTRODUCTION

With the rise of social media, hateful memes have emerged as a subtle yet harmful form of online toxicity. Unlike traditional hate speech, which is purely textual, hateful memes combine visual and textual elements to convey offensive, sarcastic, or discriminatory messages [1]. This multimodal interplay makes detection challenging, as neutral images may become hateful when paired with certain captions, and vice versa—rendering unimodal models insufficient.

To address this, researchers have explored multimodal learning strategies. Early approaches combined Convolutional Neural Networks (CNNs) and Bidirectional Encoder Representations from Transformers (BERT) via late fusion [1], whereas later models like Universal Image-Text Representation (UNITER) [2], Visual BERT (VisualBERT) [3], and Vision-

and-Language BERT (ViLBERT) [4], adopt early fusion through cross-modal attention mechanisms, and Contrastive Language–Image Pretraining (CLIP)-based approaches such as Hate-CLIPper [5] improve modality alignment using contrastive learning [6]. These models demonstrate strong performance but often require large-scale pretraining and substantial computational resources.

Multimodal fusion techniques have also been applied in other domains, including fake news detection [7], cybersecurity [8], emotion recognition [9], and sentiment analysis [10]. While effective, many of these approaches suffer from high computational cost, limited interpretability, and an inability to explicitly model uncertainty or disagreement between modalities—limitations that are particularly critical in socially sensitive tasks such as hateful meme detection.

In this work, we address these limitations by proposing a lightweight and interpretable multimodal fusion framework based on Dempster–Shafer Evidence Theory (DST), namely Multimodal Hateful Meme Classification via Dempster–Shafer Evidence Theory Fusion (MHM-DS). Unlike existing transformer-based architectures that rely on joint feature learning, our approach performs decision-level fusion of independent unimodal classifiers (BERT for text and CLIP for images). DST enables explicit modeling of belief, conflict, and ignorance, allowing the framework to handle ambiguous or contradictory multimodal evidence in a principled manner.

Although DST has been widely applied in domains such as sensor fusion [11], medical diagnosis [12], and ensemble learning [13], its application in hateful meme detection has remained limited and largely generic, typically relying on fixed or heuristic confidence assignments. In this work, DST is exploited as a structured decision-level reasoning framework tailored to the specific challenges of multimodal hateful meme classification. Our work bridges that gap by systematically integrating DST with modern pretrained vision and language models and by introducing adaptive modality reliability estimation mechanisms tailored to meme classification.

The main contributions of this paper are as follows:

- A DST-based late fusion framework for multimodal hateful meme classification.
- Adaptive reliability estimation using validation accuracy, entropy-based weighting, and a learned reliability coefficient predicted directly using a lightweight neural regressor AlphaLearner.
- Improved calibration and uncertainty estimation compared to standard late fusion baselines.
- An interpretable decision-making process through belief and ignorance analysis.

II. MULTIMODAL HATEFUL MEME CLASSIFICATION VIA DEMPSTER–SHAFER EVIDENCE THEORY FUSION

The proposed MHM-DS framework integrates DST with multimodal learning to combine text and image evidence for hateful meme detection. The complete workflow presented in Figure 1 consists of three main stages: First, unimodal feature extraction and prediction using separate deep networks for text and image. Then, confidence estimation via modal reliability weights, and finally, evidence belief mass computation and fusion using Dempster's rule to obtain coherent decision, confidence range, and uncertainty measurement.

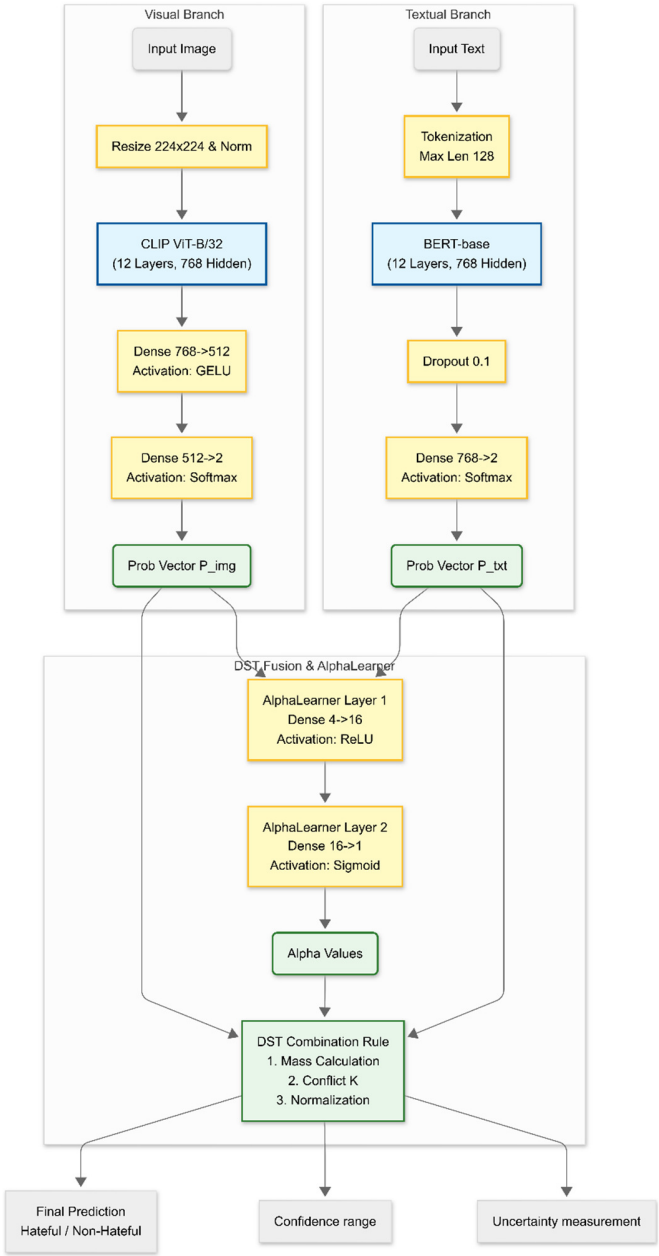

Fig. 1. Overall architecture of the DST-based fusion framework.

A. *Unimodal Modeling*

Each meme consists of an image I and a textual caption T. Two independent classifiers are trained. For the textual data classification, we fine-tune three different models: BERT-base [14], Robustly Optimized BERT Pretraining Approach

(RoBERTa)-base [15], and Distilled BERT (DistilBERT) [16]. Captions are lowercased and processed using the standard tokenization associated with each pretrained language model (BERT, RoBERTa, and DistilBERT). Sequences are padded to a maximum length of 128 tokens. No additional preprocessing such as stop-word removal, stemming, or lemmatization, is applied, in order to preserve linguistic cues related to sarcasm, implicit hate, and contextual meaning. After the benchmark study, the textual branch employs a BERT-base encoder (12 layers, 768 hidden size). The output embedding is passed through a dropout layer (p = 0.1) and a dense classification layer (768 to 2) with softmax activation to generate probability estimates.

$$P_t = \text{Softmax}(W_t h_t + b_t) = [p_t{}^{(0)}, p_t{}^{(1)}] \quad (1)$$

For image data, we also evaluate three unimodal models based on: ResNet-50 [17], EfficientNet-B0 [18], and CLIP ViT-B/32 [6]. Before processing images, they are resized to 224 × 224 to match the input requirements of ResNet-50, EfficientNet-B0, and CLIP ViT-B/32. Images are normalized using the mean and standard deviation values of the corresponding pretrained models. After the benchmarking of the three pretrained models, we selected CLIP (ViT-B/32) encoder for its superior results.

The visual branch consists of 12 transformer layers with a hidden size of 768. This is followed by a Multi-Layer Perceptron (MLP) projection head containing a dense layer (768 to 512 neurons) with Gaussian Error Linear Unit (GELU) activation, and a final classification layer (512 to 2 neurons) with softmax activation. The encoder's output logits z_v are converted to probabilities:

$$P_v = \text{Softmax}(W_v z_v + b_v) = [p_v{}^{(0)}, p_v{}^{(1)}] \quad (2)$$

Both models are optimized using the categorical cross-entropy loss:

$$L = -\sum_k y^{(k)} \log p^{(k)} \quad (3)$$

where $y^{(k)}$ is the ground-truth one-hot label.

B. Dempster–Shafer Late Fusion

To effectively integrate multimodal predictions while accounting for uncertainty and conflicts between modalities, we employ DST as the core of our late fusion mechanism. Unlike conventional fusion methods that simply average or weight probabilities, DST enables reasoning over belief and plausibility, and allows representation of uncertainty when text and image provide inconsistent evidence.

Each unimodal classifier outputs a softmax probability vector over the two classes:

$$P_t = [p_t(\text{hateful}), p_t(\text{non} - \text{hateful})] \quad (4)$$

$$P_v = [p_v(\text{hateful}), p_v(\text{non} - \text{hateful})] \quad (5)$$

We interpret these probabilities as basic belief assignments**.** To account for modality reliability, each mass function is weighted by an adaptive confidence factor α_i, reflecting the credibility of each unimodal classifier:

$$m_t(\text{hateful}) = \alpha_t * p_t(\text{hateful}) \quad (6)$$

$$m_t(\text{non} - \text{hateful}) = \alpha_t * p_t(\text{non} - \text{hateful}) \quad (7)$$

$$m_v(\text{hateful}) = \alpha_v * p_v(\text{hateful}) \quad (8)$$

$$m_v(\text{non} - \text{hateful}) = \alpha_v * p_v(\text{non} - \text{hateful}) \quad (9)$$

where $\alpha_t, \alpha_v \in [0,1]$ represent the reliability factors of the text and visual modalities, respectively. Θ denotes total ignorance (uncertainty), and the uncommitted mass $m(\Theta)$ captures the model's uncertainty:

$$m_t(\Theta) = 1 - \alpha_t \sum_c p_t(\text{c}) \quad (10)$$

$$m_v(\Theta) = 1 - \alpha_v \sum_c p_v(\text{c}) \quad (11)$$

C. Reliability Factor Estimation

The contribution of each modality to the fusion is determined by the reliability coefficients α_t and α_v . We investigate three methods to quantify it: lightweight meta-learner-based alphas, entropy-based alphas, and fixed validation alphas.

1) Fixed Validation Alphas (Baseline)

To balance the influence of each modality, we define reliability factors as a constant weight based on unimodal validation accuracy Acc_t and Acc_v for text and image respectively:

$$\alpha_t = \frac{Acc_t}{(Acc_t + Acc_v)} \quad (12)$$

$$\alpha_v = 1 - \alpha_t \quad (13)$$

This ensures that the more reliable modality contributes more strongly to the final belief, while uncertainty is proportionally distributed when both are weak. This serves as a simple baseline but lacks adaptivity to individual samples.

2) Entropy-Based Alphas

Entropy-based alphas are dynamic adaptive weights computed per sample using Shannon entropy of the model's softmax distribution, where $H(p_i)$ is the Shannon entropy of the softmax distribution and C is the number of classes:

$$H(p_i) = -\sum_k p_i{}^{(k)} \log(p_i{}^{(k)}) \quad (14)$$

$$\alpha = 1 - \frac{H(p_i)}{\log C} \quad (15)$$

During fusion, this method down-weights uncertain or ambiguous outputs while favoring confident predictions with low uncertainty and guaranteeing that modalities that produce more confident predictions (lower entropy) obtain greater reliability weights. The model responds to shifting modality confidence rather than fixed or heuristic weighting by dynamically adjusting α on a per-sample basis. This keeps the fusion process both data-driven and uncertainty-aware.

3) AlphaLearner

The reliability coefficients are predicted directly using a lightweight neural regressor trained on the model's outputs, including probabilities, maximum probability, and entropy. Each modality i (text or image) has its own learner receiving the features and then outputs the corresponding reliability coefficients:

$$x_i = [p_i{}^{(0)}, p_i{}^{(1)}, H(p_i), \max(p_i)] \quad (16)$$

$$\hat{\alpha}^i = \sigma(W_2 ReLU(W_1 x_i + b_1) + b_2) \quad (17)$$

AlphaLearner is trained to minimize mean squared error against the entropy-derived pseudo-target:

$$L_\alpha = \frac{1}{N}\sum_{j=1}^{N}\left(\hat{\alpha}_i^{(j)} - \left(1 - \frac{H(p_i{}^{(1)})}{\log C}\right)\right)^2 \quad (18)$$

This design allows the α value to capture more nuanced reliability patterns that entropy alone cannot express, such as inter-modal consistency or overconfidence.

D. Dempster's Rule for Fusion

Dempster's combination rule is used to generate the fused belief. It integrates evidence from two independent sources: the textual belief mass m_t and the visual belief mass m_v. For a given hypothesis (A), the combined mass is computed as:

$$m_f(A) = \left(\frac{1}{(1-K)}\right)\sum_{B \cap C = A} m_{t(B)} m_{v(C)} \quad (19)$$

where K represents the conflict coefficient and it measures the degree of disagreement between modalities:

$$K = \sum_{B \cap C = \Theta} m_{t(B)} m_{v(C)} \quad (20)$$

When the value of K is high, it indicates strong disagreement between the textual and visual evidence, whereas consistent support is implied by a lower K. After normalization, the joint belief obtained from both modalities is shown in the fused mass $m_f(A)$. The final decision corresponds to the class hypothesis with the maximum belief:

$$\hat{y} = \text{argmax}_c\, m_{f(c)} \quad (21)$$

The residual doubt or ignorance is captured by the remaining unassigned belief, which is commonly represented as $m_{f(\Theta)}$:

$$m_{f(\Theta)} = 1 - \Sigma_c m_{f(c)} \quad (22)$$

Instead of imposing overconfident predictions, the residual term gives an interpretable measure of uncertainty in the fused judgment, enabling the framework to clearly acknowledge conflicting or incomplete evidence.

III. EXPERIMENTAL DETAILS AND RESULTS

A. Dataset and Setup

We conduct all experiments on the Facebook AI Hateful Memes dataset [1, 19], a publicly available benchmark designed for multimodal hate speech detection. The dataset contains 10,000 multimodal samples, where each sample consists of one image and one associated text caption, resulting in 10,000 images and 10,000 textual inputs. The official split provided by the dataset creators is used, comprising 8,500 training samples, 500 validation samples, and 1,000 test samples. Each meme is annotated with a binary label indicating whether it is hateful (1) or non-hateful (0). All experiments strictly follow the official dataset splits without resampling or relabeling. This ensures fair comparison with previously published results on the Hateful Memes benchmark.

B. Implementation Details and Training Configuration

This subsection details the implementation choices and training configuration of the proposed MHM-DS framework to ensure full reproducibility of the experiments. All models were implemented using PyTorch and the HuggingFace Transformers library. Unimodal classifiers were trained independently and their probabilistic outputs were later combined using the proposed DST late fusion framework. No joint multimodal fine-tuning was performed, ensuring modularity and fair comparison across fusion strategies.

Textual inputs were processed using a pretrained BERT-base encoder, selected based on validation performance among evaluated text models. Input captions were tokenized using the BERT tokenizer, truncated to a maximum length of 128 tokens, and no additional text normalization was applied in order to preserve linguistic cues such as sarcasm and implicit hate expressions. For the visual modality, CLIP ViT-B/32 was employed as the image encoder. Images were resized to 224 × 224 pixels, normalized using CLIP's default mean and standard deviation, and passed through the frozen visual backbone followed by a trainable classification head. All unimodal classifiers were optimized using the AdamW optimizer with identical hyperparameters to ensure comparability. Training was conducted for a fixed number of epochs with early stopping based on validation accuracy.

The reliability coefficients (α) used in the DST fusion stage were estimated using three strategies: fixed validation-based weighting, entropy-based dynamic weighting, and a lightweight AlphaLearner meta-network. The AlphaLearner consisted of a two-layer MLP trained separately using entropy-derived pseudo-targets. A summary of the full implementation and training configuration is provided in Table I.

TABLE I. IMPLEMENTATION AND TRAINING CONFIGURATION

Component	**Configuration**
Framework	PyTorch
Transformer library	HuggingFace Transformers
Text encoder	BERT-base (uncased)
Image encoder	CLIP ViT-B/32
Text token length	128 tokens
Image resolution	224 × 224
Text preprocessing	Tokenization, truncation, padding
Image preprocessing	Resize, normalization (CLIP defaults)
Loss function	Categorical cross-entropy
Optimizer	AdamW
Learning rate	2×10^{-5}
Batch size	32
Training epochs	4
Weight decay	0.01
Fusion strategy	Late fusion using DST
Reliability estimation	Fixed α, Entropy-based α, AlphaLearner α
AlphaLearner architecture	2-layer MLP with ReLU and Sigmoid
Evaluation metrics	Accuracy, Area Under the Receiver Operating Characteristic Curve (AUROC), conflict coefficient (K)
Hardware	Single NVIDIA GPU

C. Ablation Study

The ablation study evaluates the contribution of each component of the proposed DST late fusion framework to the overall performance on the Hateful Memes dataset. Specifically, we investigate first the effectiveness of individual unimodal encoders, second the impact of different image pretrained models, and third the influence of reliability weighting strategies (Fixed α , Entropy-based α , and AlphaLearner α). All models are trained and evaluated using the same data splits. Performance is reported in terms of accuracy and AUROC.

1) Unimodal Baselines

Table II presents the results of fine-tuning unimodal models. Given that text information conveys hatred more than image content, our findings support the idea that text modality dominates performance. Nonetheless, supplementary contextual cues are still provided by the image models. Based on these results, we select the most accurate pretrained model on our dataset for the unimodal analysis to enhance the multimodal fusion results.

TABLE II. PERFORMANCE OF UNIMODAL MODELS ON THE HATEFUL MEMES DATASET

Model	Modality	Validation accuracy	Test accuracy	AUROC
RoBERTa-base	Text	0.6416	0.6408	0.6431
DistilBERT	Text	0.3584	0.3592	0.6172
BERT-base	Text	0.6596	0.6612	0.6867
ResNet-50	Image	0.6078	0.6243	0.5778
EfficientNet-B0	Image	0.5937	0.6204	0.6034
CLIP ViT-B/32	Image	0.6353	0.6478	0.6273

2) DST Fusion with Adaptive Reliability

As already explained in the methodology section, we investigated three different methods to estimate the modality reliability weights (α_t, α_v) before applying Dempster's combination rule. In Table III, we present the values of the modality reliability weights for the three estimation methods and their corresponding metrics showing the impact on our model.

TABLE III. IMPACT OF RELIABILITY ESTIMATION METHODS ON DST FUSION PERFORMANCE

α estimation method	α_t (text)	α_v (image)	Accuracy (%)	AUROC (%)	Mean $K \downarrow$
Fixed α	0.5094	0.4906	70.35	68.66	0.1187
Entropy-based α	0.8107	0.8179	70.43	69.64	0.2481
AlphaLearner α	0.4850	0.4574	70.20	70.83	0.1088

These results highlight that the α estimation mechanism directly affects the quality of DST fusion:

- The Fixed α method, which is based on validation accuracies, provides limited adaptability but stability because all samples have the same modality trust weights.
- The entropy-based α approach introduces sample-level adaptivity, adjusting reliability according to model confidence. However, it can overemphasize overconfident but incorrect predictions, slightly harming F1-score consistency.
- The AlphaLearner α approach provides data-driven reliability estimation that generalizes across different meme contexts, achieving the best trade-off. It improves AUROC and resilience to contradictory evidence by learning subtle weighting between modalities.

Regardless of the α estimation approach, all variations preserve low conflict coefficients ($K < 0.2$), demonstrating that the DST framework successfully resolves cross-modal disagreement.

D. Comparison with State-of-the-Art

To evaluate the effectiveness of the proposed DST-based late fusion framework, we compared its performance against a range of baseline and state-of-the-art multimodal models designed for hateful meme detection. The results are presented in Table IV. This comparison includes unimodal models, BERT-base (text-only) and CLIP ViT-B/32 [6] (image-only), as well as a late fusion baseline using score averaging.

TABLE IV. COMPARISON OF DST FUSION FOR HATEFUL MEMES CLASSIFICATION WITH STATE-OF-THE-ART MULTIMODAL MODELS

Model	Accuracy (%)	AUROC (%)	Fusion type	Key Idea
Human performance	84.70	–	–	Upper bound with human annotators
Text only BERT-base	66.12	69.69	–	Text-only unimodal analysis
Image only CLIP ViT-B/32	60.71	62.73	–	Image-only unimodal model
Late fusion baseline	63.20	67.30	Late	Simple averaging of unimodal scores
ViLBERT [4]	62.30	70.45	Early	Joint text–image transformer
VisualBERT (COCO) [3]	63.20	71.33	Early	Vision-language BERT fine-tuned on COCO
Facebook best model (2020) [19]	76.50	81.08	Early	Multimodal BERT fusion from the Hateful Memes Challenge
TCAM (2024) [20]	76.40	83.60	Early	Transformer with cross-modality attention masks
MHM-DS (ours)	70.20	70.83	Late	DST-based belief combination

The Facebook AI Hateful Memes Challenge introduced the first large-scale benchmark for this task. The challenge provided ViLBERT [4] and VisualBERT [3] as official baselines, and published prize-winning work based on early fusion multimodal BERT [19]. Since then, several advanced transformer architectures, such as TCAM [20], have progressively improved performance through deep cross-modal attention mechanisms and large-scale pretraining.

Human performance is about 84.7%, which represents the upper bound. Text-only and image-only models perform around 60–66. The prize-winning Facebook challenge solution achieved 76.50% accuracy and 80.08% AUROC, whereas advanced multimodal transformers such as TCAM reach 76% accuracy and 83.6% AUROC.

Our proposed MHM-DS model achieves 70.20% accuracy and 70.83% AUROC, outperforming early baselines and even matching the performance range of some early fusion transformer models, despite not relying on massive pretraining or cross-modal transformers.

While recent transformer-based multimodal architectures achieve higher absolute accuracy through deep cross-modal attention and large-scale pretraining, such gains come at the cost of significant computational complexity, reduced interpretability, and limited uncertainty awareness. In contrast, the proposed DST-based late fusion framework is designed to prioritize interpretability, calibration, and robustness to modality disagreement, rather than solely maximizing accuracy.

This demonstrates that DST-based belief fusion effectively integrates multimodal evidence and provides a lightweight and interpretable alternative to complex deep fusion architectures.

E. Interpretability and Confidence Analysis

A key advantage of using DST for fusion is its ability to model uncertainty and provide interpretable decisions, avoiding overconfident outputs. In our framework, each modality (text and image) contributes an independent belief mass over the possible hypotheses (hate, non-hateful), along with a remaining ignorance mass $m_t(\Theta)$ that explicitly captures uncertainty.

1) Belief Distribution Visualization

Figure 2 shows belief distributions for the text, image, and fused outputs. While unimodal models show high confidence (beliefs near 0.9–1.0), DST fusion produces a flatter distribution, reflecting adaptive reliability balancing.

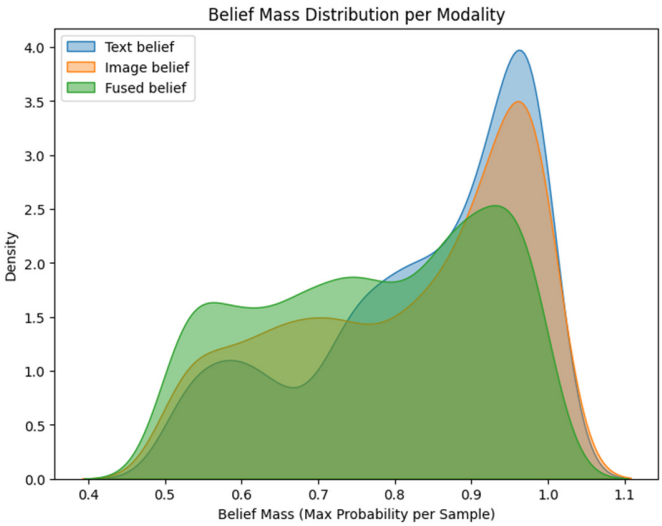

Fig. 2. Belief mass distribution across modalities.

DST modulates each modality's influence based on its reliability (α). When predictions align, belief is concentrated; when they diverge, DST shifts mass toward uncertainty, avoiding overconfident decisions.

2) Uncertainty Distribution

Figure 3 illustrates the residual ignorance $m_t(\Theta)$, averaging 0.23. This indicates well-calibrated confidence: low uncertainty for aligned modalities, and higher uncertainty when cues conflict. Unlike softmax-based fusion, DST explicitly quantifies ignorance, supporting downstream reliability estimation or human-in-the-loop moderation.

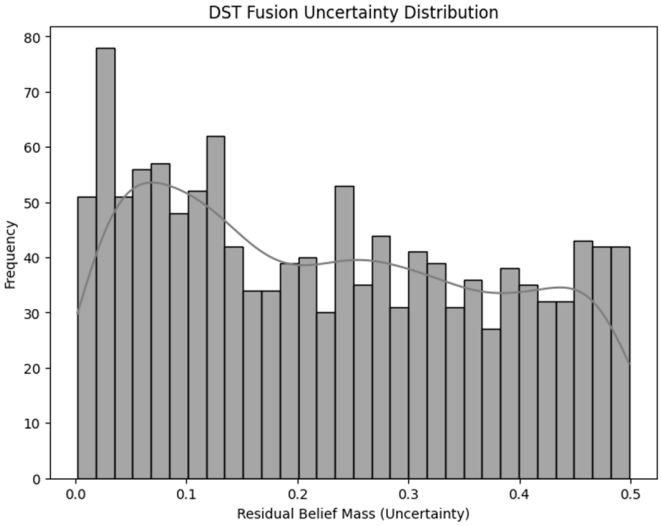

Fig. 3. Residual ignorance mass distribution.

3) Modality Contribution Analysis

The adaptive reliability coefficients α_t and α_v vary dynamically across samples. On average, the learned weights (α_t ≈0.48, $\alpha_v \approx 0.46$) show that both modalities contribute comparably. However, certain meme types exhibit dominance of one modality:

- Text-dominant memes: sarcasm or coded hate phrases.
- Image-dominant memes: implicit visual hate.

This dynamic weighting enables the model to adaptively trust the more informative modality, unlike fixed fusion gates in standard transformer architectures.

IV. DISCUSSION

The proposed MHM-DS framework shows promising results, improving interpretability, decision transparency, and confidence calibration. The framework offers clear probabilistic reasoning and explicit uncertainty modeling, which is valuable for real-world moderation tasks where trust and explainability matter. DST is not used as a direct reapplication of a standard fusion rule, but rather as a structured decision-level reasoning framework tailored to the specific challenges of multimodal hateful meme classification. Our approach introduces adaptive, data-driven modality reliability estimation through entropy-based weighting and a learned AlphaLearner, enabling sample-level adjustment of belief masses according to uncertainty and modality disagreement. Unlike prior DST-based fusion methods, which assume static or uniform source credibility, the proposed

framework explicitly models conflict, ignorance, and reliability variability between text and image modalities, which are critical characteristics of hateful memes involving sarcasm, implicit meaning, or contradictory cues. As a result, the contribution of this work lies not in the theoretical reformulation of DST itself, but in its principled integration with modern pretrained unimodal classifiers and adaptive reliability mechanisms, yielding an interpretable, uncertainty-aware, and computationally efficient fusion strategy specifically designed for socially sensitive multimodal content moderation tasks.

We can summarize the key strengths of our approach as follows:

- Efficiency: The framework is lightweight and easy to deploy, requiring minimal resources compared to large multimodal transformers.
- Transparency: It outputs belief, ignorance, and conflict measures, offering interpretable decisions often missing in deep models.
- Modularity: Any unimodal encoder can be integrated without retraining joint models.
- Practicality: The approach can complement more complex systems as an uncertainty-aware decision layer.
- Good performance: Despite its simplicity, it matches early fusion baselines like VisualBERT [3] and ViLBERT [4].

However, our approach has its limitations that suggest future investigation and enhancements:

- Modality independence: DST assumes that text and image provide independent evidence, which may not hold in all meme contexts.
- Lack of cross-modal interaction: The current design fuses decisions, not features; lightweight cross-modal reasoning could further enhance performance.

In summary, our model balances interpretability and performance, offering a reliable decision layer that complements complex multimodal systems.

V. DATASET USAGE AND COPYRIGHT COMPLIANCE

The Hateful Memes dataset is publicly released by Facebook AI Research for non-commercial research purposes under its official usage license. All data used in this study were obtained through the authorized dataset distribution channels, and usage strictly complies with the dataset's terms and conditions.

Images and text captions were used exclusively for model training and evaluation and are not redistributed as part of this publication. The authors do not claim ownership of any dataset content, and all rights remain with the original dataset providers. References to the dataset and its creators are properly cited in accordance with academic standards.

VI. CONCLUSION

This work presents Multimodal Hateful Meme Classification via Dempster–Shafer Evidence Theory Fusion (MHM-DS), a Dempster–Shafer Evidence Theory (DST)-based late fusion framework for multimodal hateful meme detection, offering an efficient alternative to conventional probabilistic fusion. The approach enhances interpretability and decision reliability by explicitly modeling uncertainty and conflict between modalities. This contribution is particularly important for socially sensitive applications, such as content moderation. The modular design and low computational footprint of MHM-DS make it well-suited for low-resource environments, whereas its strong performance demonstrates the value of evidential reasoning in multimodal tasks. Despite limitations in cross-modal interaction, DST effectively handles ambiguous inputs and can complement more complex models. Future research will explore differentiable DST variants and hybrid architectures that integrate transformer-based embeddings, aiming to combine interpretability with state-of-the-art representational power.

REFERENCES

[1] D. Kiela *et al.*, "The hateful memes challenge: detecting hate speech in multimodal memes," in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, Vancouver, Canada, 2020, pp. 2611–2624.

[2] Y.-C. Chen *et al.*, "UNITER: UNiversal Image-TExt Representation Learning," in *16th European Conference on Computer Vision*, Virtual Event, 2020, pp. 104–120, https://doi.org/10.1007/978-3-030-58577-8_7.

[3] L. H. Li, M. Yatskar, D. Yin, C.-J. Hsieh, and K.-W. Chang, "VisualBERT: A Simple and Performant Baseline for Vision and Language." arXiv, Aug. 09, 2019, https://doi.org/10.48550/arXiv.1908.03557.

[4] J. Lu, D. Batra, D. Parikh, and S. Lee, "ViLBERT: pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," in *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, Vancouver, Canada, 2019, pp. 13–23.

[5] G. K. Kumar and K. Nandakumar, "Hate-CLIPper: Multimodal Hateful Meme Classification based on Cross-modal Interaction of CLIP Features," in *Proceedings of the Second Workshop on NLP for Positive Impact*, Abu Dhabi, United Arab Emirates (Hybrid), 2022, pp. 171–183, https://doi.org/10.18653/v1/2022.nlp4pi-1.20.

[6] A. Radford *et al.*, "Learning Transferable Visual Models From Natural Language Supervision," in *Proceedings of the 38th International Conference on Machine Learning*, Virtual Event, 2021, pp. 8748–8763.

[7] H. Chen *et al.*, "A self-learning multimodal approach for fake news detection," *Frontiers in Artificial Intelligence*, vol. 8, Nov. 2025, Art. no. 1665798, https://doi.org/10.3389/frai.2025.1665798.

[8] V. Ravi and A. S. Poornima, "SecMa: A Novel Multimodal Autoencoder Framework for Encrypted IoT Traffic Analysis and Attack Detection," *Engineering, Technology & Applied Science Research*, vol. 15, no. 3, pp. 23020–23026, June 2025, https://doi.org/10.48084/etasr.10336.

[9] H. Kumar and M. Aruldoss, "Gated Cross-Modal Fusion Mechanism for Audio-Video-based Emotion Recognition," *Engineering, Technology & Applied Science Research*, vol. 15, no. 2, pp. 20835–20841, Apr. 2025, https://doi.org/10.48084/etasr.9430.

[10] S. Fatimi, W. Sabbar, and A. Bekkhoucha, "Toward a Dual Attention Model for Image-Text Sentiment Classification," in *2023 IEEE Afro-Mediterranean Conference on Artificial Intelligence*, Hammamet, Tunisia, 2023, pp. 01–07, https://doi.org/10.1109/AMCAI59331.2023.10431498.

[11] T. Denœux, "Decision-making with belief functions: A review," *International Journal of Approximate Reasoning*, vol. 109, pp. 87–110, June 2019, https://doi.org/10.1016/j.ijar.2019.03.009.

[12] L. A. Zadeh, "A Simple View of the Dempster-Shafer Theory of Evidence and its Implication for the Rule of Combination," *AI Magazine*, vol. 7, no. 2, pp. 85–90, 1986, https://doi.org/10.1609/aimag.v7i2.542.

[13] Y. Wu, X. Liu, and L. Guo, "A New Ensemble Clustering Method Based on Dempster-Shafer Evidence Theory and Gaussian Mixture Modeling," in *21th International Conference on Neural Information Processing (Proceedings, Part II)*, Kuching, Malaysia, 2014, pp. 1–8, https://doi.org/10.1007/978-3-319-12640-1_1.

[14] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota, 2019, pp. 4171–4186, https://doi.org/10.18653/v1/N19-1423.

[15] Y. Liu *et al.*, "RoBERTa: A Robustly Optimized BERT Pretraining Approach." arXiv, July 26, 2019, https://doi.org/10.48550/arXiv.1907.11692.

[16] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter." arXiv, Mar. 01, 2020, https://doi.org/10.48550/arXiv.1910.01108.

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, 2016, pp. 770–778, https://doi.org/10.1109/CVPR.2016.90.

[18] M. Tan and Q. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," in *Proceedings of the 36th International Conference on Machine Learning*, Long Beach, CA, USA, 2019, pp. 6105–6114.

[19] R. Velioglu and J. Rose, "Detecting Hate Speech in Memes Using Multimodal Deep Learning Approaches: Prize-winning solution to Hateful Memes Challenge." arXiv, Dec. 23, 2020, https://doi.org/10.48550/arXiv.2012.12975.

[20] F. Wu, G. Chen, J. Cao, Y. Yan, and Z. Li, "Multimodal Hateful Meme Classification Based on Transfer Learning and a Cross-Mask Mechanism," *Electronics*, vol. 13, no. 14, July 2024, Art. no. 2780, https://doi.org/10.3390/electronics13142780.