Talk, Snap, Complain: Validation-Aware Multimodal Expert Framework for Fine-Grained Customer Grievances

Rishu Kumar Singh[*1], **Navneet Shreya**[*2], **Sarmistha Das**[*1], **Apoorva Singh**[*3], **Sriparna Saha**[1]

[1]Indian Institute of Technology Patna, India
[2]National Institute of Technology Patna, India
[3]Fondazione Bruno Kessler, Italy
{rishu_2301ee36, sarmistha_2221cs21, sriparna}@iitp.ac.in, navneets.ug23.cs@nitp.ac.in, apoorva0816@gmail.com

Abstract

Existing approaches to complaint analysis largely rely on unimodal, short-form content such as tweets or product reviews. This work advances the field by leveraging multimodal, multi-turn customer support dialogues—where users often share both textual complaints and visual evidence (e.g., screenshots, product photos)—to enable fine-grained classification of complaint aspects and severity. We introduce *VALOR*, a Validation-Aware Learner with Expert Routing, tailored for this multimodal setting. It employs a multi-expert reasoning setup using large-scale generative models with Chain-of-Thought (CoT) prompting for nuanced decision-making. To ensure coherence between modalities, a semantic alignment score is computed and integrated into the final classification through a meta-fusion strategy. In alignment with the United Nations Sustainable Development Goals (UN SDGs), the proposed framework supports SDG 9 (Industry, Innovation and Infrastructure) by advancing AI-driven tools for robust, scalable, and context-aware service infrastructure. Further, by enabling structured analysis of complaint narratives and visual context, it contributes to SDG 12 (Responsible Consumption and Production) by promoting more responsive product design and improved accountability in consumer services. We evaluate *VALOR* on a curated multimodal complaint dataset annotated with fine-grained aspect and severity labels, showing that it consistently outperforms baseline models, especially in complex complaint scenarios where information is distributed across text and images. This study underscores the value of multimodal interaction and expert validation in practical complaint understanding systems.

Resources — https://github.com/sarmistha-D/VALOR
Extended version — http://arxiv.org/abs/2511.14693

Introduction

> Your most unhappy customers are your greatest source of learning.
>
> *Bill Gates*

Understanding customer complaints is crucial for improving user experience, service reliability, and product quality, objectives that directly align with the United Nations Sustainable Development Goals (SDGs), particularly SDG 9 (Industry, Innovation, and Infrastructure) and SDG 12 (Responsible Consumption and Production). While prior work has advanced complaint analysis, most efforts focus on short-form, unimodal inputs such as tweets (Preotiuc-Pietro, Gaman, and Aletras 2019). These formats often lack the evolving context and emotional nuance found in real-world customer support scenarios. Moreover, modern users increasingly supplement textual complaints with visual content, such as screenshots of broken interfaces or damaged items shared across platforms (Singh et al. 2022a, 2023c). However, current research rarely addresses this multimodal and conversational nature of complaints, which introduces unique challenges in cross-modal alignment, contextual reasoning, and fine-grained classification.

Traditional approaches like aspect-based sentiment analy-

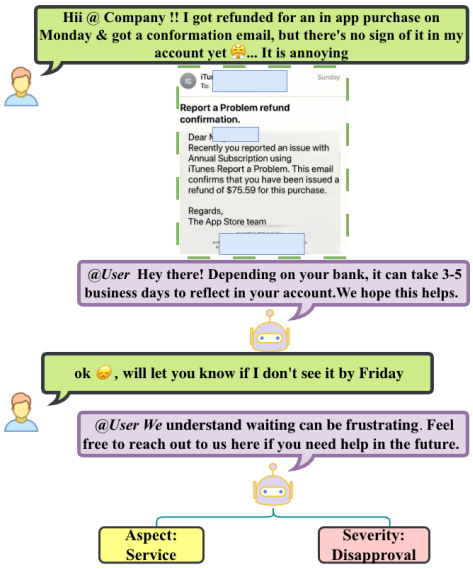

Figure 1: A conversation snippet from the *CIViL* dataset. Labels indicate the aspect-severity pairs identified from the conversation.

[*]These authors contributed equally.

sis primarily assign sentiment polarities (positive, negative, neutral) to specific aspects in isolated reviews. While helpful in gauging user opinion, these methods fall short when complaints evolve over multi-turn dialogues or involve multimodal evidence. Sentiment alone often lacks the specificity needed for actionable insights. For example, a user might report a late delivery and share a photo of a damaged product, each pointing to different aspects (logistics and packaging) and severity levels. Without disentangling such layered signals, automated systems fail to produce serviceable insights for resolution teams.

We redefine complaint analysis as a fine-grained multimodal classification task over multi-turn dialogues, jointly modeling conversational flow and aligned images to identify aspect categories and severity levels with contextual accuracy. These structured outputs enable downstream applications such as intent-based ticket routing, escalation prediction, and service analytics. Figure 1 presents an example of such a use case.

Building on this formulation, we propose *CIViL*: Customer Interactions with Visual and Linguistic signals, a curated dataset of multi-turn dialogues annotated with aspect and severity annotations and paired with topically aligned images sourced from social media websites. Motivated by the effectiveness of Large Language Models integrated with Mixture-of-Experts (MoE) architectures, we adopt a modular learning strategy that balances expert specialization with shared representation learning. To this end, we propose **Validation-Aware Learner with Expert Routing (*VALOR*)**, a multimodal framework that transforms complaint interactions into structured outputs using a robust MoE architecture enhanced with semantic alignment and Chain-of-Thought (CoT) reasoning. To assess the quality of modality integration and expert behavior, we incorporate a validation module that applies a three-part metric system, capturing *alignment* across modalities, *dominance* when one modality is more informative, and *complementarity* when both modalities contribute distinct yet coherent information. This work lays the foundation for real-world multimodal complaint understanding in dialogue systems by transforming unstructured complaints into structured, fine-grained insights; providing annotated resources, scalable methodologies, and actionable feedback to enhance service responsiveness and consumer engagement.

Research Objectives: Following are the research objectives of the current study:
(1) To investigate how textual and visual signals enhance fine-grained complaint analysis, specifically aspect and severity detection in multi-turn customer support dialogues.
(2) To design and evaluate a two-phase MoE framework, *VALOR*, that combines cross-modal fusion, semantic alignment scoring, and CoT-based expert reasoning to produce accurate predictions from multimodal complaint data.
(3) To develop a domain-specific multimodal customer-support dialogue dataset, enabling systematic benchmarking and advancing research in multimodal dialogue-grounded grievance analysis.

Contributions: The primary contributions are as follows:
(1) We define and investigate the task of fine-grained multimodal complaint understanding within multi-turn dialogues, with a specific emphasis on identifying aspect categories (ACD) and assessing severity levels (SD).
(2) We introduce *CIViL*, a benchmark multimodal dataset of customer-support dialogues, built by extending a subset of the Kaggle customer support corpus with fine-grained aspect and severity annotations, and enriched with topically aligned images. This resource aims to advance research in multimodal conversational complaint understanding.
(3) We propose *VALOR*, a Validation-Aware Learner with Expert Routing tailored for robust identification of real-world multimodal complaints in conversational settings.
(4) The proposed framework sets a strong benchmark for fine-grained multimodal complaint analysis in dialogue-based contexts, consistently outperforming baselines across multiple metrics.

Related Works

Complaint Detection

Complaint detection from text has become a key area in computational linguistics. Early methods relied on rule-based systems and handcrafted features to identify dissatisfaction (Singh et al. 2021). With the rise of deep learning, particularly transformer models like BERT (Devlin et al. 2019) semantic and contextual modeling improved significantly. Recent multitask approaches further enhance generalization by incorporating related cues such as sentiment, emotion, and sarcasm (Singh et al. 2022b; Singh and Saha 2021; Singh, Nazir, and Saha 2022; Singh et al. 2023b). However, these studies are limited to short, single-turn texts like tweets which lack contextual depth.
Multimodal methods have begun addressing this gap by integrating visual and textual inputs (Singh et al. 2022a; Devanathan et al. 2024; Singh et al. 2024; Das et al. 2025a), but mostly rely on static features or simplistic fusion, often missing nuanced cross-modal interactions. Moreover, evaluations are typically conducted on product reviews, which lack the temporal structure and expressiveness of real-time, multi-turn conversations. Without multi-turn context, these models struggle to capture evolving emotions and layered complaint dimensions. In contrast, multi-turn dialogues offer richer signals, users elaborate issues over time, correct themselves, and discuss multiple concerns not evident in isolated posts.

Fine-grained Complaint Analysis

Recent advances in complaint analysis have focused on fine-grained tasks like severity classification (Jin and Aletras 2021; Singh, Bhatia, and Saha 2024; Das et al. 2024), typically using transformer-based models to estimate emotional intensity in social media posts. However, the limited context and ambiguity of short-form content often reduce prediction accuracy. To improve granularity, later work explored aspect-based modeling using attention mechanisms (Singh et al. 2023d,a; Jain et al. 2023; Das et al. 2025b). While effective, these methods underutilize the deeper reasoning and contextual capabilities of large language models—essential

S.No.	Annotation Guidelines
1	Annotations must be carried out independently, without outside influence.
2	Aspect categories should be assigned based on the customer's viewpoint.
3	Each identified aspect must be paired with an appropriate severity level.
4	Choose the aspect label that most accurately reflects the specific cause of dissatisfaction.
5	Ambiguous cases should be resolved through discussion among annotators and authors.

Table 1: Annotation guidelines for *CIViL* dataset.

for handling the complexity of aspect and severity classification in rich, multi-layered complaint narratives.

Mixture-of-Experts

Mixture-of-Experts (MoE) architectures enhance model performance by dynamically routing inputs to specialized, lightweight sub-models, each trained on distinct regions of the task space (Freund, Schapire et al. 1996). Unlike traditional ensembles that aggregate outputs post-hoc, modern MoE models employ learnable gating mechanisms to activate only relevant experts during inference, enabling efficient and scalable learning (He et al. 2021; Jiang et al. 2024). Integrated with large language models (LLMs), these methods have shown strong results across various NLP tasks (Shen et al. 2024; Li et al. 2023), making them suitable for fine-grained complaint analysis in dialogues. More recently, MoE frameworks have been extended to multimodal settings, where experts are trained to process specific modalities or modality combinations, allowing for flexible and fine-grained cross-modal reasoning (Yu et al. 2024; Li et al. 2025). Designing such systems requires attention to expert diversity and routing precision, as misallocation can degrade performance due to loss of critical information.

Research Gap

While automated complaint classification has advanced in recent years, most approaches remain limited to unimodal inputs like text-only tweets or reviews, missing the increasingly common use of visual evidence such as screenshots or product images. Despite this shift in user behavior, current research rarely incorporates visual context into the modeling of customer-support grievances. The interaction between text and images, especially in fine-grained tasks like aspect and severity classification remains underexplored. Multimodal complaint scenarios in dialogue settings introduce unique challenges, including modality alignment, ambiguity resolution, and cross-modal reasoning, which conventional LLMs and vision-language models are not explicitly designed to address. To address these gaps, we construct a multimodal complaint dataset grounded in customer-support dialogues and propose *VALOR*, a validation-aware multi-expert framework that fuses textual and visual cues via cross-attention, semantic alignment, and Chain-of-Thought reasoning for fine-grained complaint understanding.

Customer Interactions with Visual and Linguistic Signals Dataset (*CIViL*)

For this study, we build upon the publicly available *Kaggle Customer Support on Twitter dataset* (Axelbrooke 2017), the only publicly available large-scale collection of English-language customer-agent dialogues across domains such as airlines, tech, and retail. From over 20,000 conversations, we focused on interactions involving Apple Support, comprising 14% of the dataset and filtered for two-speaker dialogues ranging from 2 to 10 utterances[1] to reflect typical support exchanges. A subset of 2,004 conversations was randomly sampled and annotated with fine-grained aspect and severity levels.

To enrich the dataset with visual context, we scraped 4,478 relevant images from the same time period from X and Reddit websites using a two-phase pipeline. The scraping process utilized the PRAW library for Reddit and Scrapy for X. First, we curated images related to common complaint themes (e.g., broken screens, battery drain, camera quality) from targeted subreddits. Next, we employed a CLIP-based (Radford et al. 2021) semantic matching algorithm to assign topically aligned images to conversations based on textual content. Only high-confidence image–conversation pairs were retained through a multi-stage similarity and validation process.

Annotator Details

Three annotators independently labeled each dialogue with aspect and severity labels. Disagreements were resolved through collaborative review sessions, which also involved iterative refinements in the annotation guidelines. The team comprised one Ph.D. researcher and two postgraduate scholars, all with prior experience in supervised dataset creation and domain-specific annotation. Their strong command of English, supported by formal education in English-medium institutions, ensured consistency and accuracy in interpreting conversational content.

Statistic	Count
Total Conversations	2004
Total Utterances	7101
Total Images	4478
Total Customer Utterances	3825
Total Support agent Utterances	3276
Average User Utterance per conversation	2.74
Average Support agent Utterance per conversation	1.49

Table 2: *CIViL* dataset statistics

Annotation Phase & Dataset Analysis

To ensure consistency and reduce ambiguity, annotators were equipped with comprehensive guidelines (Table 1) and

[1] An utterance, also referred to as a turn, typically comprises multiple sentences.

a reference set of 50 sample conversations. The annotation process follows established protocols commonly used in aspect-based sentiment analysis (Liao et al. 2021; Nazir et al. 2020). The *CIViL* dataset comprises the following label distributions:
(a) Severity levels-Blame (799), Disapproval (486), Accusation (484), and No Explicit Reproach (235).
(b) Aspect categories-Software (1,662), Quality (117), Hardware (112), Service (77), Price (23), and Packaging (13). Detailed dataset statistics are summarized in Table 2.

Fleiss' Kappa scores (Fleiss 1971) for inter-annotator agreement were 0.68 for aspect categories and 0.75 for severity levels, indicating substantial consistency among annotators (McHugh 2012). As the dataset exclusively comprises complaint-driven customer-support dialogues, it contains no non-complaint instances. Each conversation is labeled with one or more aspect–severity pairs, capturing the distinct issues raised within a single interaction. In the interest of space, additional annotation details and dataset statistics for *CIViL* are provided in the supplementary resources linked with the paper.

Proposed Approach

Problem Statement: Given a multimodal input (T, I), where $T \in \mathbb{R}^{L \times d_T}$ denotes the tokenized textual embedding and $I \in \mathbb{R}^{3\times224\times224}$ represents the normalized image tensor, the task is to perform joint classification over two orthogonal dimensions: **ACD** with C_a discrete categories (e.g., Software, Hardware, Packaging, etc.) and **SD** with C_s ordinal levels (e.g., No Reproach, Disapproval, Blame, Accusation). A unified model $f_\theta : (T, I) \to (l_a, l_s)$ is learned, where $l_a \in \mathbb{R}^{C_a}$ and $l_s \in \mathbb{R}^{C_s}$ denote the raw logits for aspect and severity, respectively. The predictive distributions are obtained via softmax activation:

$$P(y_a \mid T, I) = \text{softmax}(l_a), \quad P(y_s \mid T, I) = \text{softmax}(l_s)$$

enabling efficient end-to-end optimization of the dual-target complaint classification objective within a multimodal learning framework. We present *VALOR*, a two-step multimodal architecture that combines Chain-of-Thought reasoning with expert validation for fine-grained complaint classification. The framework separates prediction and validation phases, enhancing both accuracy and transparency in multimodal understanding. Figure 2, outlines the proposed *VALOR* framework.

1. **Phase 1 Prediction (Foundation of COT MoE):** The prediction phase transforms multimodal complaint data into structured outputs through a robust mixture-of-experts framework enhanced with semantic alignment and Chain-of-Thought (CoT) reasoning. Initially, raw text inputs T are tokenized using the BERT-base-uncased tokenizer (vocabulary size $V = 30{,}522$) and truncated to $L = 512$ tokens, yielding token tensors $\mathbf{T} \in \mathbb{R}^{\mathcal{B}\times L}$, while image inputs I are resized to 224×224 and passed through a ViT-patch16 (Dosovitskiy et al. 2021) embedding module to form $\mathbf{I} \in \mathbb{R}^{\mathcal{B}\times3\times224\times224}$ with 196 patches. The text is encoded using a 12-layer, 12-head BERT transformer (hidden size $d = 768$) to produce contextual embeddings $\mathbf{H}_t \in \mathbb{R}^{\mathcal{B}\times L\times d}$ and a [CLS] token $\mathbf{h}_t \in \mathbb{R}^{\mathcal{B}\times d}$, while the image passes through a ViT-base encoder yielding patch embeddings $\mathbf{H}_i \in \mathbb{R}^{\mathcal{B}\times196\times d}$ and CLS vector $\mathbf{h}_i \in \mathbb{R}^{\mathcal{B}\times d}$. These representations are fused using cross-modal multi-head attention with $H = 8$ heads, where for each head h, queries $\mathbf{Q}_h = \mathbf{H}_t\mathbf{W}_q^{(h)}$, keys $\mathbf{K}_h = \mathbf{H}_i\mathbf{W}_k^{(h)}$, and values $\mathbf{V}_h = \mathbf{H}_i\mathbf{W}_v^{(h)}$ are computed using projection matrices $\mathbf{W}_q^{(h)}, \mathbf{W}_k^{(h)}, \mathbf{W}_v^{(h)} \in \mathbb{R}^{d\times d/H}$. Attention is computed as $\text{softmax}(\mathbf{Q}_h\mathbf{K}_h^\top/\sqrt{d/H})\mathbf{V}_h$, and outputs are concatenated, projected, and passed through residual layers and feed-forward networks, followed by mean pooling to yield the unified multimodal embedding $\mathbf{x} \in \mathbb{R}^{\mathcal{B}\times d}$. Parallelly, a Semantic Alignment Score (SAS) is computed by projecting $\mathbf{h}_t$ and $\mathbf{h}_i$ into a shared 512-dimensional space using two-layer MLPs with GELU activation. The outputs are layer-normalized, concatenated, and passed through another MLP with $\tanh$ activation to yield the scalar alignment score $s \in [-1, 1]^{\mathcal{B}}$. The fused embedding $\mathbf{x}$ is then routed to $\mathcal{K} = 4$ Chain-of-Thought experts, each built on the DeepSeek-6.7B model (hidden size $d_t = 4096$). Each expert k transforms the input as $\mathbf{x}'_k = \mathbf{x} \odot \boldsymbol{\alpha}_k + \boldsymbol{\beta}_k$, with learnable parameters $\boldsymbol{\alpha}_k, \boldsymbol{\beta}_k \in \mathbb{R}^d$, followed by projection $\mathbf{x}''_k = \mathbf{x}'_k\mathbf{W}_{\text{in}}^{(k)} + \mathbf{b}_{\text{in}}^{(k)}$ and autoregressive reasoning using temperature $\tau = 0.5$, top-k $= 30$, and top-p $= 0.9$ sampling. This generates reasoning tokens and a final hidden state $\mathbf{h}_{\text{final}} \in \mathbb{R}^{\mathcal{B}\times d_t}$, projected as logits $\boldsymbol{\ell}^{(k)} = \mathbf{h}_{\text{final}}\mathbf{W}_{\text{out}}^{(k)} + \mathbf{b}_{\text{out}}^{(k)} \in \mathbb{R}^{\mathcal{B}\times\mathcal{C}}$, where $\mathcal{C} \in \{\mathcal{C}_a, \mathcal{C}_s\}$ denotes class counts for aspect or severity prediction. Expert routing is handled by a learned gating function $\mathbf{g} = \text{softmax}(\mathbf{x}\mathbf{W}_r + \mathbf{b}_r) \in \mathbb{R}^{\mathcal{B}\times\mathcal{K}}$, producing soft probabilities $g_{b,k}$ for expert relevance. Hard top-1 selection yields expert index $k_b^* = \arg\max_k g_{b,k}$, resulting in routing matrix $\mathbf{R} \in \{0, 1\}^{\mathcal{B}\times\mathcal{K}}$. To prevent expert collapse and encourage load balancing, the routing entropy $H(\mathbf{g}_b) = -\sum_{k=1}^{\mathcal{K}} g_{b,k} \log g_{b,k}$ is computed, and a regularization loss is introduced: $L_{\text{lb}} = \sum_{k=1}^{\mathcal{K}} \left(\frac{1}{\mathcal{K}} - \frac{1}{\mathcal{B}}\sum_{b=1}^{\mathcal{B}} g_{b,k}\right)^2$.

2. **Phase 2 (Validation MoE):** The validation phase introduces robust secondary reasoning using $\mathcal{L}_v = 2$ specialized validation experts and a multi-perspective evaluation pipeline to enhance prediction confidence and interpretability. Each validation expert is instantiated as a DeepSeek transformer stack with 32 layers and hidden dimension $d_t = 4096$, receiving the joint multimodal embedding $\mathbf{x} \in \mathbb{R}^{\mathcal{B}\times d}$, which is first linearly projected via $\mathbf{W}_{v,\text{in}}^{(l)} \in \mathbb{R}^{d\times d_t}$ and bias $\mathbf{b}_{v,\text{in}}^{(l)}$, yielding $\mathbf{x}'_l$. This is processed by the transformer block to obtain $\mathbf{h}_l$, then projected through $\mathbf{W}_{v,\text{out}}^{(l)}$ and passed through a two-layer MLP with ReLU activations to generate logits $\boldsymbol{\ell}_v^{(l)} \in \mathbb{R}^{\mathcal{B}\times\mathcal{C}}$. To analyze expert behavior, a threefold metric system is employed: (i) *Alingment* evaluates cosine similarity between logits of experts l and m, yielding $\text{Alingment}_{l,m} = \frac{\langle\boldsymbol{\ell}_v^l, \boldsymbol{\ell}_v^m\rangle}{\|\boldsymbol{\ell}_v^l\|\cdot\|\boldsymbol{\ell}_v^m\|}$, with mean score R_{avg};

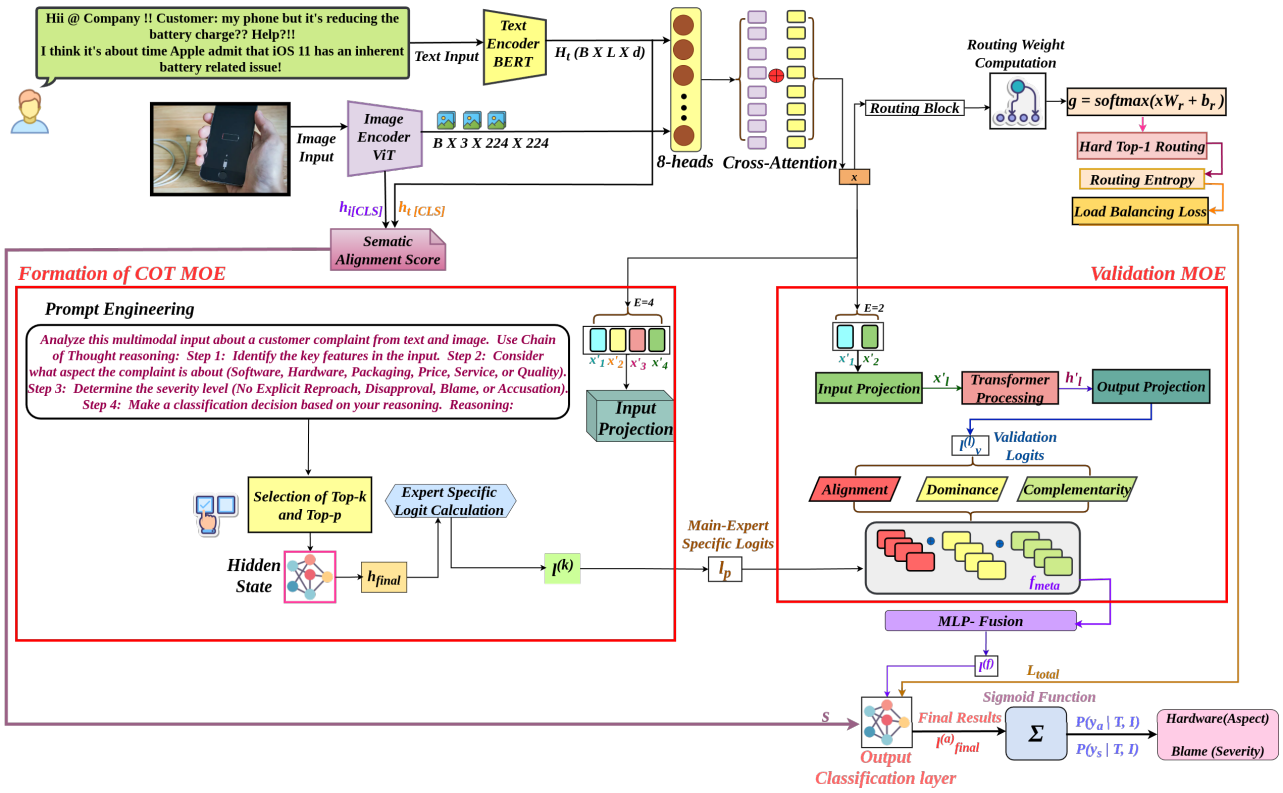

Figure 2: Architectural view of proposed *VALOR* framework

(ii) *dominance* quantifies predictive alignment between MoE outputs $\boldsymbol{\ell}_p$ and validation logits $\boldsymbol{\ell}_v$ via correlation: $\text{dominance}^{(a)} = \frac{\text{Cov}(\boldsymbol{\ell}_p^{(a)}, \boldsymbol{\ell}_v^{(a)})}{\sqrt{\text{Var}(\boldsymbol{\ell}_p^{(a)}) \cdot \text{Var}(\boldsymbol{\ell}_v^{(a)})}}$; and (iii) *complementarity*, a diversity measure, is captured via entropy over softmax-normalized logits: $\text{complementarity}^{(l)} = -\sum_{c=1}^{\mathcal{C}} p_v^{(l,c)} \log p_v^{(l,c)}$, with $p_v^{(l,c)} = \text{softmax}(\boldsymbol{\ell}_v^{(l)})_c$ and average U_{avg}. A meta-fusion network then aggregates predictions through routing-aware combination: $\boldsymbol{\ell}_p = \sum_k \mathbf{R}_{\cdot,k} \cdot \boldsymbol{\ell}^{(k)}$ and $\boldsymbol{\ell}_v = \sum_l g_{v,l} \cdot \boldsymbol{\ell}_v^l$, where $g_{v,l}$ are soft routing weights. The combined feature vector $\mathbf{f}_{\text{meta}} = [\boldsymbol{\ell}_p; \boldsymbol{\ell}_v; s; \bar{H}; R_{\text{avg}}; \text{dominance}; U_{\text{avg}}] \in \mathbb{R}^{\mathcal{B} \times \mathcal{M}}$ (with $\mathcal{M} = 2\mathcal{C}_a + 5$) is passed through a 3-layer MLP with hidden sizes $(768, 384, \mathcal{C}_a)$, ReLU activations, and dropout rate 0.1 to generate fused logits $\boldsymbol{\ell}_f$. These are adjusted by SAS-based alignment: $\boldsymbol{\ell}_{\text{final}} = \boldsymbol{\ell}_f + \lambda_s \cdot s \cdot \mathbf{1}_{\mathcal{C}_a}$, with $\lambda_s = 0.1$, and final predictions computed as $P(y|\cdot) = \text{softmax}(\boldsymbol{\ell}_{\text{final}})$. The overall training objective integrates aspect and severity classification losses using label-smoothed cross-entropy ($\epsilon_{\text{ls}} = 0.15$), validation loss L_{val}, semantic alignment margin loss $L_{\text{sas}} = \frac{1}{\mathcal{B}} \sum_b \max(0, \mu - s_b)$ with $\mu = 0.3$, and metric-driven regularizers: $L_{\text{Alingment}} = \max(0, R_{\text{avg}} - \tau_R)$, $L_{\text{dominance}} = \max(0, \tau_S - \text{dominance}^{(a)})$, and $L_{\text{complementarity}} = \max(0, \tau_U - U_{\text{avg}})$, with thresholds $\tau_R = 0.3$, $\tau_S = 0.5$, $\tau_U = 1.5$. The final total loss is computed as:

$$L_{\text{total}} = L_{\text{aspect}} + L_{\text{severity}} + \lambda_{\text{lb}} L_{\text{lb}} + \lambda_{\text{val}} L_{\text{val}} + \lambda_s L_{\text{sas}} + \lambda_R L_{\text{Alingment}} + \lambda_S L_{\text{dominance}} + \lambda_U L_{\text{complementarity}},$$

The final predictions for both aspect and severity are computed as $P(y_a \mid T, I) = \text{softmax}(l_{final}^{(a)})$ and $P(y_s \mid T, I) = \text{softmax}(l_{final}^{(s)})$, respectively.

Experiments and Results

This section describes the evaluation setup used to ensure a fair and rigorous comparison with strong state-of-the-art baselines. Our analysis is driven by three research questions:
RQ1: How does *VALOR* perform relative to state-of-the-art models?
RQ2: What is the individual contribution of each architectural component in *VALOR*?
RQ3: Which expert configuration delivers optimal task performance?

Evaluation Protocol

We evaluate the proposed *VALOR* framework on the *CIViL* dataset. The dataset is split into 70% training, 10% validation, and 20% testing to ensure statistical robustness and generalization fidelity. All experiments are conducted on an NVIDIA RTX 3090 GPU, using the AdamW optimizer (Loshchilov and Hutter 2019) with an initial learning rate of 2×10^{-5}, linear warm-up, and cosine decay scheduling.

Model	ACD		SD	
	A	F1	A	F1
DeepSeek-VL	0.66	0.65	0.66	0.65
Gemma-3 (9B)	0.69	0.66	0.65	0.66
Flash Gemini (1.6B)	0.66	0.65	0.66	0.65
ImageBind	0.66	0.65	0.64	0.63
Paligemma (3B)	0.65	0.66	0.65	0.64
SMOL-VLM	0.65	0.64	0.63	0.62
GIT (300M)	0.65	0.64	0.63	0.62
FLAVA	0.62	0.61	0.60	0.59
ALBEF	0.61	0.60	0.59	0.56
UNITER	0.60	0.59	0.56	0.55
CLIP ViT-B/32	0.59	0.56	0.55	0.56
VisualBERT	0.56	0.55	0.56	0.55
ViLT	0.55	0.56	0.55	0.54

Table 3: Performance comparison of baseline models on the *CIViL* dataset after 20 epochs of fine-tuning. Models are ranked by their overall F1-score. A: Accuracy, F1: macro F1-score

Optimization employs *binary cross-entropy loss* for multi-label classification, with a batch size of 16, dropout rate of 0.5, and gradient clipping (max norm = 1.0) to stabilize training. Models are trained for up to 20 epochs with early stopping (patience = 5) based on validation loss, and a fixed random seed (42) ensures full reproducibility. Evaluation is performed using Accuracy and macro F1-score, computed independently for both ACD and SD tasks to provide a comprehensive assessment of model performance across the two fine-grained complaint dimensions.

Baseline Comparison

To rigorously benchmark the performance of our *VALOR* framework, we evaluate it against a broad suite of state-of-the-art multimodal models across three learning paradigms, *zero-shot*, *few-shot*, and *fully fine-tuned*, to assess adaptability under varying supervision levels. The baselines include leading **vision-aligned language models** such as DeepSeek-VL, Gemma-3 (9B), Flash Gemini (1.6B), and Paligemma (3B), as well as prominent **vision-language pretraining architectures** like ImageBind, SMOL-VLM, GIT (300M), FLAVA, ALBEF, UNITER, CLIP ViT-B/32, VisualBERT, and ViLT. Together, these models represent the state-of-the-art in multimodal representation learning. As shown in Table 3, we report and rank their performance after 20 epochs of fine-tuning on the *CIViL* dataset, providing a comprehensive comparison across both aspect and severity prediction tasks.

The results indicate that larger models such as Gemma-3 (9B) tend to perform better, suggesting that model scale is a significant factor in this complex task. Models specifically designed for vision-language integration, like DeepSeek-VL, also show strong performance. However, there is considerable variance across the board, underscoring the challenges of fine-grained multimodal analysis.

Ablation Study

To quantify the impact of key design choices in the *VALOR* framework, we conduct a focused ablation study across four core components. First, we compare our Chain-of-Thought (CoT) experts with standard MLP and Transformer-based experts. Second, we assess the effect of including the *Validation MoE* module. Third, we evaluate our learnable *Semantic Alignment Score* (SAS) against cosine similarity and alignment-agnostic baselines. Lastly, we vary the *Top-K routing* parameter to analyze the influence of expert sparsity. Results in Table 4 provides proposed framework *VALOR* result and detailed insights into the role of each component.

Results Analysis

Our experimental results are guided by three research questions (RQs):

RQ1: How does *VALOR* perform relative to state-of-the-art models? In its full configuration, *VALOR* achieves significant performance gains over all competitive baselines, attaining 81.94% aspect classification accuracy and 72.51% severity accuracy, marking absolute improvements of 12.94% and 6.51%, respectively, over the strongest contender, *Gemma-3*. These results empirically validate the efficacy of our expert-driven, validation-aware architecture and underscore its superior capacity for multimodal complaint understanding.

RQ2: What is the contribution of each component in *VALOR*? The ablation analysis substantiates the additive efficacy of each architectural module within *VALOR*. Notably, the incorporation of the *Validation MoE* yields a substantial gain of 8.2% in aspect accuracy (from 73.74% to 81.94%), highlighting its pivotal role in expert quality control. Furthermore, the integration of a learnable *Semantic Alignment Score (SAS)* consistently outperforms static alignment baselines (e.g., cosine similarity or no alignment), affirming the importance of dynamic cross-modal alignment in enhancing representational fidelity and task-specific reasoning.

RQ3: Which expert type is most effective for this task? *Chain-of-Thought (CoT)* experts exhibit clear superiority over both Transformer and MLP-based counterparts, attributed to their explicit step-by-step reasoning capabilities that are essential for capturing the nuanced semantics of customer complaints. While Transformer experts offer competitive performance due to their expressive capacity, they fall short in interpretability and sequential inference. MLP experts, limited by their architectural simplicity and lack of contextual modeling, yield the weakest results, underscoring the critical advantage of structured reasoning in this domain. *All reported results are statistically significant*[2] *(Welch 1947).*

Human Evaluation: To provide a fine-grained and context-aware assessment of *VALOR*'s real-world effectiveness beyond conventional automated metrics, we conducted a rigorous human evaluation (Figure 3) involving 200 randomly selected test samples from the *CIViL* dataset, encompassing diverse complaint categories and severity levels. Using a win-loss-draw protocol, expert evaluators com-

[2]We performed Student's t-test for the test of significance. The results are statistically significant when testing the null hypothesis (p-value < 0.05).

Configuration	Expert	Validation	MoE	SAS	Top-K	ACD (A)	SD (A)	ACD (F1)	SD (F1)
CoT (No Validation, Learnable SAS, Top-2)	cot	False		learnable	2	73.74	62.62	70.44	52.84
CoT (No Validation, Learnable SAS, Top-4)	cot	False		learnable	4	75.14	64.64	70.46	59.47
VALOR	**cot**	**True**		**learnable**	**2**	**81.94**	**72.51**	**76.96**	**67.91**
MLP (No Validation, Learnable SAS, Top-2)	mlp	False		learnable	2	70.43	57.35	63.82	48.55
MLP (No Validation, Learnable SAS, Top-4)	mlp	False		learnable	4	71.97	58.98	65.04	54.45
MLP (Validation, Cosine SAS, Top-2)	mlp	True		cosine	2	68.81	65.24	61.20	57.58
MLP (Validation, Cosine SAS, Top-4)	mlp	True		cosine	4	69.06	65.14	66.62	56.14
MLP (Validation, No SAS, Top-2)	mlp	True		none	2	71.19	64.39	68.18	60.69
Transformer (No Validation, Learnable SAS, Top-2)	transf	False		learnable	2	77.51	67.95	70.62	61.70
Transformer (No Validation, No SAS, Top-2)	transf	False		none	2	72.86	52.67	68.22	43.84
Transformer (No Validation, No SAS, Top-4)	transf	False		none	4	65.51	59.48	59.79	55.17
Transformer (Validation, Learnable SAS, Top-2)	transf	True		learnable	2	77.08	63.98	70.24	60.24
Transformer (Validation, No SAS, Top-2)	transf	True		none	2	74.84	63.55	71.30	58.05

Table 4: Results for *VALOR* framework and ablation study of its different components. The table shows the performance impact of different expert types, validation strategies, and SAS settings. Best scores are in bold face. A: Accuracy, F1: macro F1-score, transf: transformer

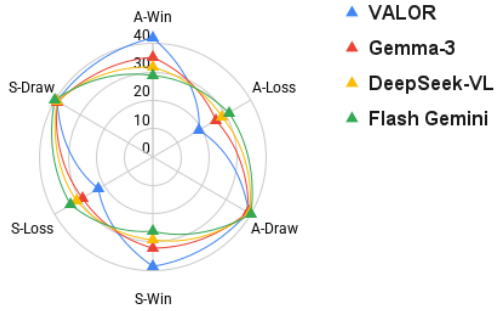

Figure 3: Human Evaluation on win-loss-draw performance criteria between popular baselines against *CIViL*; Here A- Stands for Aspect and S-stands for Severity

pared *VALOR*'s predictions against those of state-of-the-art baselines, Gemma-3 (a reasoning-aligned encoder-decoder LLM), DeepSeek-VL, and Flash Gemini, across two core dimensions: aspect identification and severity classification. *VALOR* demonstrated superior judgmental fidelity, achieving the highest win rates at 42.3% for aspect identification and 38.5% for severity classification, while simultaneously maintaining the lowest loss rates of 18.7% and 22.1%, respectively. These empirical findings underscore *VALOR*'s architectural advantage in integrating validation-aware multimodal experts, Chain-of-Thought-enabled expert modules, and semantic alignment mechanisms, which collectively enhance its interpretability and robustness in nuanced complaint understanding tasks.

Qualitative Analysis: To gain deeper insight into the behavior of *VALOR* framework, we conducted a focused qualitative analysis of its predictions. Key observations include: (1) Multi-aspect recognition: *VALOR* effectively handles complex complaints involving multiple issues. For example, a user reporting a "battery issue and slow software," accompanied by an image indicating the battery problem, is accurately classified into distinct aspect–severity pairs: hardware–disapproval and software–accusation.
(2) Expert fallback reliability: The validation-aware MoE layer successfully intervenes in low-confidence cases from the primary experts, offering a corrective secondary inference. In the interest of space, additional qualitative analyses and representative examples are provided in the supplementary resources linked with the paper.

Error Analysis: The error analysis of *VALOR* highlights following key challenges impacting aspect–severity classification in conversational settings:
(1) *Subjective severity interpretation:* Variability in user tone or emotionally neutral expressions can lead the model to underestimate or misclassify severity levels.
(2) *Class imbalance:* Over representation of dominant aspects (e.g., "software") and underrepresentation of others (e.g., "price") skew predictions and reduce generalization to low-frequency categories. This distribution, while challenging, mirrors real-world complaint frequency patterns.

Conclusion and Future Work

This work introduces *CIViL*, a benchmark multimodal dialogue dataset, and *VALOR*, a modular Mixture-of-Experts framework for fine-grained multimodal complaint analysis in customer-support conversations. By integrating semantic alignment, Chain-of-Thought reasoning, and expert validation, *VALOR* produces structured, interaction-aware predictions that consistently outperform strong baselines. Our experiments highlight the effectiveness of dynamic expert routing in disentangling complex, multimodal signals across dialogue turns. Future work will focus on extending the framework to support multilingual scenarios and additional multimodal signals, while incorporating speaker roles and temporal dependencies to enhance its applicability across diverse service contexts and user populations.

Ethics Statement

This work is conducted solely for the research community and is not intended for commercial use. Neither the authors nor the annotators intend to defame any company. Authors and annotators refrained from expressing personal views during the dataset creation process.

References

Axelbrooke, S. 2017. Customer Support on Twitter.

Das, S.; Lyngkhoi, R. Z. M.; Jain, K.; Goyal, V.; Saha, S.; and Gupta, M. 2025a. When Words Can't Capture It All: Towards Video-Based User Complaint Text Generation with Multimodal Video Complaint Dataset. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, 5634–5641.

Das, S.; Mujavarsheik, B.; Lyngkhoi, R. Z.; Saha, S.; and Maurya, A. 2025b. Deciphering the complaint aspects: Towards an aspect-based complaint identification model with video complaint dataset in finance. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 7195–7204. IEEE.

Das, S.; Singh, A.; Saha, S.; and Maurya, A. 2024. Negative review or complaint? exploring interpretability in financial complaints. *IEEE Transactions on Computational Social Systems*, 11(3): 3606–3615.

Devanathan, R.; Singh, A.; Poornash, A. S.; and Saha, S. 2024. Seeing Beyond Words: Multimodal Aspect-Level Complaint Detection in Ecommerce Videos. In Cai, J.; Kankanhalli, M. S.; Prabhakaran, B.; Boll, S.; Subramanian, R.; Zheng, L.; Singh, V. K.; César, P.; Xie, L.; and Xu, D., eds., *Proceedings of the 32nd ACM International Conference on Multimedia, MM 2024, Melbourne, VIC, Australia, 28 October 2024 - 1 November 2024*, 243–252. ACM.

Devlin, J.; Chang, M.; Lee, K.; and Toutanova, K. 2019. 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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, 4171–4186. Association for Computational Linguistics.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Fleiss, J. L. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5): 378.

Freund, Y.; Schapire, R. E.; et al. 1996. Experiments with a new boosting algorithm. In *icml*, volume 96, 148–156. Citeseer.

He, J.; Qiu, J.; Zeng, A.; Yang, Z.; Zhai, J.; and Tang, J. 2021. Fastmoe: A fast mixture-of-expert training system. *arXiv preprint arXiv:2103.13262*.

Jain, R.; Singh, A.; Gangwar, V.; and Saha, S. 2023. AbCoRD: Exploiting multimodal generative approach for Aspect-based Complaint and Rationale Detection. In *Proceedings of the 31st ACM International Conference on Multimedia*, 8571–8579.

Jiang, S.; Zheng, T.; Zhang, Y.; Jin, Y.; Yuan, L.; and Liu, Z. 2024. Med-MoE: Mixture of Domain-Specific Experts for Lightweight Medical Vision-Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, 3843–3860.

Jin, M.; and Aletras, N. 2021. Modeling the Severity of Complaints in Social Media. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, 2264–2274. Association for Computational Linguistics.

Li, J.; Su, Q.; Yang, Y.; Jiang, Y.; Wang, C.; and Xu, H. 2023. Adaptive Gating in Mixture-of-Experts based Language Models. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, 3577–3587. Association for Computational Linguistics.

Li, Y.; Jiang, S.; Hu, B.; Wang, L.; Zhong, W.; Luo, W.; Ma, L.; and Zhang, M. 2025. Uni-moe: Scaling unified multimodal llms with mixture of experts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Liao, W.; Zeng, B.; Yin, X.; and Wei, P. 2021. An improved aspect-category sentiment analysis model for text sentiment analysis based on RoBERTa. *Applied Intelligence*, 51(6): 3522–3533.

Loshchilov, I.; and Hutter, F. 2019. Decoupled Weight Decay Regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

McHugh, M. L. 2012. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3): 276–282.

Nazir, A.; Rao, Y.; Wu, L.; and Sun, L. 2020. Issues and challenges of aspect-based sentiment analysis: A comprehensive survey. *IEEE Transactions on Affective Computing*, 13(2): 845–863.

Preotiuc-Pietro, D.; Gaman, M.; and Aletras, N. 2019. Automatically Identifying Complaints in Social Media. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28-August 2, 2019, Volume 1: Long Papers*, 5008–5019. Association for Computational Linguistics.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 8748–8763. PMLR.

Shen, S.; Hou, L.; Zhou, Y.; Du, N.; Longpre, S.; Wei, J.; Chung, H. W.; Zoph, B.; Fedus, W.; Chen, X.; Vu, T.; Wu, Y.; Chen, W.; Webson, A.; Li, Y.; Zhao, V. Y.; Yu, H.; Keutzer, K.; Darrell, T.; and Zhou, D. 2024. Mixture-of-Experts Meets Instruction Tuning: A Winning Combination

for Large Language Models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Singh, A.; Bhatia, R.; and Saha, S. 2024. Complaint and Severity Identification From Online Financial Content. *IEEE Trans. Comput. Soc. Syst.*, 11(1): 660–670.

Singh, A.; Dey, S.; Singha, A.; and Saha, S. 2022a. Sentiment and Emotion-Aware Multi-Modal Complaint Identification. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, 12163–12171. AAAI Press.

Singh, A.; Gangwar, V.; Sharma, S.; and Saha, S. 2023a. Knowing What and How: A Multi-modal Aspect-Based Framework for Complaint Detection. In *Advances in Information Retrieval: 45th European Conference on Information Retrieval, ECIR 2023, Dublin, Ireland, April 2–6, 2023, Proceedings, Part II*, 125–140. Springer.

Singh, A.; Jain, R.; Jha, P.; and Saha, S. 2023b. Peeking inside the black box: A commonsense-aware generative framework for explainable complaint detection. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 7333–7347.

Singh, A.; Jha, P.; Bhatia, R.; and Saha, S. 2023c. What Is Your Cause for Concern? Towards Interpretable Complaint Cause Analysis. In *European Conference on Information Retrieval*, 141–155. Springer.

Singh, A.; Jha, P.; Das, S.; Jain, R.; and Saha, S. 2024. Toward Multimodal Complaint Severity Detection From Social Media. *IEEE Trans. Comput. Soc. Syst.*, 11(5): 5903–5912.

Singh, A.; Nazir, A.; and Saha, S. 2022. Adversarial Multitask Model for Emotion, Sentiment, and Sarcasm Aided Complaint Detection. In *Advances in Information Retrieval - 44th European Conference on IR Research, ECIR 2022, Stavanger, Norway, April 10-14, 2022, Proceedings, Part I*, volume 13185 of *Lecture Notes in Computer Science*, 428–442. Springer.

Singh, A.; and Saha, S. 2021. Are You Really Complaining? A Multi-task Framework for Complaint Identification, Emotion, and Sentiment Classification. In *International Conference on Document Analysis and Recognition*, 715–731. Springer.

Singh, A.; Saha, S.; Hasanuzzaman, M.; and Dey, K. 2022b. Multitask Learning for Complaint Identification and Sentiment Analysis. *Cogn. Comput.*, 14(1): 212–227.

Singh, A.; Saha, S.; Hasanuzzaman, M.; and Jangra, A. 2021. Identifying complaints based on semi-supervised min-cuts. *Expert Syst. Appl.*, 186: 115668.

Singh, A.; Verma, A.; Jain, R.; and Saha, S. 2023d. Investigating the Impact of Multimodality and External Knowledge in Aspect-level Complaint and Sentiment Analysis. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, 2291–2300.

Welch, B. L. 1947. The generalization of 'STUDENT'S'problem when several different population varlances are involved. *Biometrika*, 34(1-2): 28–35.

Yu, H.; Qi, Z.; Jang, L. K.; Salakhutdinov, R.; Morency, L.-P.; and Liang, P. P. 2024. Mmoe: Enhancing multimodal models with mixtures of multimodal interaction experts. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 10006–10030.