Federated Meta-Learning for Emotion and Sentiment Aware Multi-modal Complaint Identification

Apoorva Singh[1], **Siddarth Chandrasekar**[2], **Sriparna Saha**[1] and **Tanmay Sen**[3]
[1]Department of Computer Science and Engineering, IIT Patna, India
[2]Department of Electronics and Communication Engineering, IIITDM, Kancheepuram, India
[3]Ericsson,Kolkata, India
{apoorva_1921cs19,sriparna}@iitp.ac.in, {siddarthc2000, sentanmay518}@gmail.com

Abstract

Automatic detection of consumers' complaints about items or services they buy can be critical for organizations and online merchants. Previous studies on complaint identification are limited to text. Images along with the reviews can provide cues to identify complaints better, thus emphasizing the importance of incorporating multi-modal inputs into the process. Generally, the customer's emotional state significantly impacts the complaint expression; thus, the effect of emotion and sentiment on complaint identification must also be investigated. Furthermore, different organizations are usually not allowed to share their privacy-sensitive records due to data security and privacy concerns. Due to these issues, traditional models find it hard to understand and identify complaint patterns, particularly in the financial and healthcare sectors. In this work, we extend a benchmark multi-modal complaint dataset, a collection of reviews and images of the products posted on the e-commerce website Amazon. We propose a federated meta-learning-based multi-modal multi-task framework for identifying complaints considering emotion recognition and sentiment analysis as two auxiliary tasks. Experimental results indicate that the proposed approach outperforms the baselines and the state-of-the-art approaches in centralized and federated meta-learning settings[1].

1 Introduction

Nowadays, social media platforms and online e-commerce websites provide users with the freedom to express their opinions and observations towards a product, an organization, or an event. Customers who plan to buy a product usually base their decisions on customer reviews (Preotiuc-Pietro et al., 2019). As a result, commercial and retail firms regard product reviews as a significant source of knowledge, which they can use to design their advertising strategies and resolve any product-related concerns. This could also benefit customers by providing recommendations on the quality of goods or services they intend to buy. Identifying complaint texts in natural language is critical for developers of downstream applications such as chatbots (Lailiyah et al., 2017), commercial organizations to strengthen their customer service capabilities by identifying and resolving product-related issues (Coussement and Van den Poel, 2008).
The emotional state and sentiment of an individual have a significant impact on the intended content (Lewis et al., 2010). However, emotion recognition is a far more subtle and fine-grained analysis than sentiment classification (Kumar et al., 2019). The correlation between emotion and sentiment motivates us to consider customers' sentiment and emotion while analyzing complaints. We learn the tasks of complaint identification, emotion recognition, and sentiment classification in a multi-task setting to examine further the relationship between complaint, emotion, and sentiment.

According to an analysis of relevant literature, text-based complaints have been previously analyzed based on semi-supervised strategies, different domains, degree of urgency, and feedback likelihood (Preotiuc-Pietro et al., 2019; Singh et al., 2021a; Tjandra et al., 2015; Yang et al., 2019b), (Jin and Aletras, 2020). Although multi-modal information sources (e.g., images in addition to text) could provide more information in identifying complaints, this has not been investigated to date, with one of the main reasons being the lack of multi-modal datasets. Deep-learning-based complaint detection techniques generally produce better outcomes. Nonetheless, the datasets available for training in such a particular application area are often unbalanced, with the class of interest (complaints) being significantly underrepresented compared to the others. This decreases the efficacy of binary

[1] The dataset and code are available at https://github.com/appy1608/EMNLP2023-Multimodal-Complaint-Detection

Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 16091–16103
December 6-10, 2023 ©2023 Association for Computational Linguistics

classifiers, thereby prejudicing the findings toward the dominant class (non-complaint) even though the minority class is of main interest. Using more labeled data from different sources or domains is one possible option. However, in practical applications, such data are almost always stored in separate geographical locations or with different organizations and are unavailable to others due to individual privacy or legal considerations. Federated Learning (FL) (Yang et al., 2019a) helps organizations break down such data-related barriers by offering an expansive range of data available to them.
The ability of *learning to learn* or meta-learning (Lake et al., 2011; Vinyals et al., 2016; Snell et al., 2017; Finn et al., 2017) could be beneficial in data-scarce scenarios. Organizations with recently launched products and a few training samples could benefit from using the meta-learning technique. In the healthcare domain, user health-record-related complaints often involve personal data gathered in providing healthcare services that may disclose details about the patient's medical history if made public; healthcare institutions are legally required to maintain privacy. In such cases, the proposed FL model safeguards end-user data and maintains privacy while providing healthcare services with a robust trained model.

In this work, we extend a benchmark multi-modal complaint dataset (Singh et al., 2022) with additional review instances and manually annotate each review with the complaint, emotion, and sentiment labels. Subsequently, we proposed a federated meta-learning-based framework for identifying complaints in a multi-modal multi-task framework. Even though the dataset developed and used in our work is based on publicly available reviews and images, the proposed model could benefit organizations interested in collaborative learning while keeping their local data private.

The key contributions of our proposed work are outlined as follows:

- We propose an attention-based joint learning framework for multi-modal complaint, emotion, and sentiment analysis. Complaint Identification (CI) is the primary task in our multi-task framework, whereas Emotion Recognition (ER) and Sentiment Analysis (SA) are considered auxiliary tasks.

- We extend the recently released multi-modal complaint dataset (Singh et al., 2022) with 3928 reviews and associated images with manual annotation of emotion, sentiment, and complaint classes.

- We develop a novel prototypical network (Snell et al., 2017) based complaint classifier. A prototypical network is a metric-based meta-learning technique in which computing distances between prototype representations can classify each class. We develop the meta-learning model in federated learning settings to handle data scarcity and locality issues.

- Experimental results indicate that the proposed approach outperforms the baselines and state-of-the-art approaches in centralized and federated meta-learning settings. We present the state-of-the-art for automatically identifying complaints in a multi-modal multi-task scenario in centralized and federated meta-learning settings.

2 Multi-modal Complaint Dataset (MCD)

Publicly available complaint datasets, such as (Preotiuc-Pietro et al., 2019; Singh et al., 2020), only consider text-based complaints and exclude the associated sentiment and emotion classes of the complaints. For building the proposed framework, we use the benchmark multi-modal complaint dataset *CESAMARD* (Singh et al., 2022) that consists of reviews and associated user-uploaded images from the e-commerce platform Amazon, labeled with emotion and sentiment classes. We extend the *CESAMARD* dataset with an additional 3928 review instances, each annotated with emotion and sentiment in addition to complaint labels. Here, we discuss the details of the extended multi-modal complaint (MCD) dataset.

2.1 Data Collection

We gathered the reviews posted by customers from Amazon India[2] website; to collect the product reviews and the corresponding review image URLs, we used Scrapy[3], an open-source web-crawling framework. The reviews were further divided into seven different domains (Books, Edible, Electronics, Fashion, Health & Beauty, Home Essentials, and Miscellaneous) for a more fine-grained gold standard dataset, as shown in Figure 1a). To eliminate noise (HTML tags and special characters)

[2] https://www.amazon.in/

[3] https://scrapy.org/

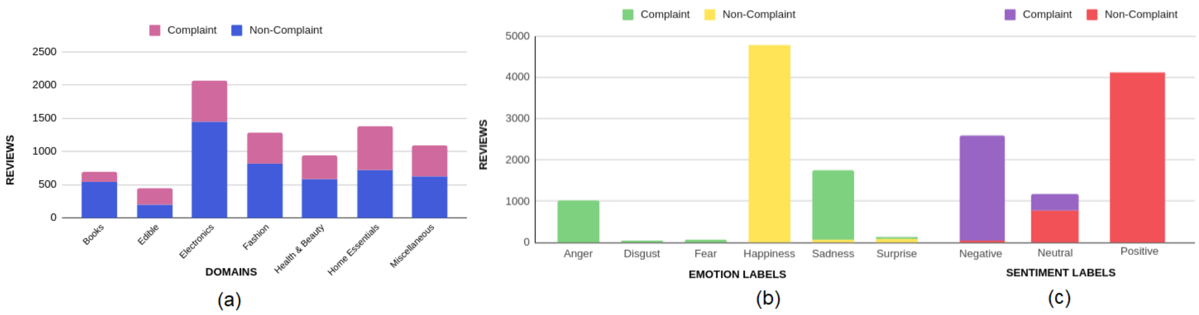


Figure 1: (a) Domain-wise distribution of the complaint and non-complaint classes in the MCD dataset. (b) Distribution of emotion labels across MCD dataset. (c) Distribution of sentiment labels across MCD dataset.

from the textual portion of the dataset, we performed some pre-processing operations on the corpus. The unicode emojis in the product reviews were converted to emoji short text with the Python module Emoji[4].

2.2 Data Annotation

We assigned three graduate students[5] who are fluent in English to annotate the reviews with appropriate complaint/non-complaint labels as well as emotion and sentiment tags. Before the annotation process began, the guidelines for annotation were provided, along with a few examples. If the review includes at least one complaint speech act, we consider the entire review to be a complaint. We utilize the complaint definition from linguistic research (Cohen and Olshtain, 1993) for complaint annotation: "A complaint presents a state of affairs that breaches the writer's favorable expectation." For the emotion annotation, we consider Ekman's six basic emotions[6] (Ekman et al., 1987) (*anger, disgust, fear, happiness, sadness, and surprise*). For the sentiment annotation, we consider three sentiment classes (*negative, neutral, positive*).

The majority voting technique was used for selecting the final complaint, emotion, and sentiment labels. Additionally, in reviews that contain terms such as "but", "however", "still", "although", etc., the statement following such terms was given more weightage while annotating the sentiment and emotion classes. For reviews of composite nature, the predominant sentiment/emotion class is determined based on the percentage of positive or negative sentences in the particular review instance. Reviews with contradictory annotations were eliminated from the dataset. We computed the Fleiss' (Fleiss, 1971) Kappa scores to assess inter-rater agreement among the three annotators and attained the agreement scores of 0.86, 0.68, and 0.82 on the complaint, emotion, and sentiment tasks, respectively, which are considered reliable.

The MCD dataset now comprises 7890 reviews with the corresponding complaint, emotion, and sentiment labels. Overall there are 4931 non-complaint reviews and 2959 complaint reviews in the dataset. Each record in the MCD dataset consists of the domain, review title, review text, image URL, and their corresponding annotated complaint, sentiment, and emotion labels. Please refer to Section A.1 in the **Appendix** where we show some sample annotations from the MCD dataset. Distributions of emotion and sentiment labels across the MCD dataset is shown in Figure 1b) and 1c), respectively. Please refer to Section A.2 in the **Appendix**, where we illustrate through some examples from the dataset the significance of multimodal analysis of complaints and incorporation of emotion and sentiment as auxiliary tasks.

3 Proposed Methodology

In this section, we define our problem and go over the details of the proposed approach. The overall framework is shown in Figure 2.

3.1 Problem Definition:

We intend to learn three closely related tasks at the same time, including complaint identification (main task), emotion recognition, and sentiment analysis (auxiliary tasks). Let $(x_{\mathrm{m}}, e_{\mathrm{m}}, s_{\mathrm{m}}, c_{\mathrm{m}})_{m=1}^{M}$ be a set of M reviews where e_{m}, s_{m}, and c_{m} represent the matching emotion, sentiment and com-

[4] https://pypi.org/project/emoji/

[5] Annotators were recruited from the author's institution.

[6] When an emotion conveyed in a review does not fall into one of the six categories, the annotators label it with the next closest emotion linked with the review. For example, "Optimistic" reviews can be directly mapped to the closest "Happiness" emotion class. We came across only 23 such reviews.

plaint labels for the $x_{\mathrm{m}}{}^{\mathrm{th}}$ instance, respectively. Here, $x_{\mathrm{m}} \epsilon X$ (title, review and image), $e_{\mathrm{m}} \epsilon E$ (emotion classes), $s_{\mathrm{m}} \epsilon S$ (sentiment classes) and $c_{\mathrm{m}} \epsilon C$ (complaint classes).

Our multi-task learning framework's objective is to maximize the posterior probability (1) across all three tasks, and is given as follows:

$$\Pi_{m=0}^{M} P(s_m, e_m, c_m | x_m; \theta) \tag{1}$$

where θ denotes the model's parameters we aim to optimize.

We aim to solve the above stated problem in federated setting. Each participating client, $K \epsilon [N]$, possess 2 datasets, namely support $D^{KSup} = \{X^{KSup}, E^{KSup}, S^{KSup}, C^{KSup}\}$ and query $D^{KQue} = \{X^{KQue}, E^{KQue}, S^{KQue}, C^{KQue}\}$ set. Employing meta-learning, each client K, aims to classify X^{KQue} with a prior knowledge imparted from D^{KSup}. So, given at any communication round $i \epsilon I$, client K tries to maximize the following:

$$\Pi_{m=0}^{M} P(e_m^{KQue}, s_m^{KQue}, c_m^{KQue} | r_m^{KQue}; D^{KSup}; \theta_i) \tag{2}$$

where r_m^{KQue} is the m^{th} query input whose complaint label (c_m^{KQue}), emotion label (e_m^{KQue}) and sentiment label (s_m^{KQue}) are to be predicted. D^{KSup} denotes the support set, which consists of both the input and corresponding labels, and θ_i denotes the global parameters communicated to the client at communication round i which we aim to optimize.

3.2 Multi-modal Feature Extraction

The process for extracting features across different modalities is detailed below.

Text Features: The word-wise textual features are obtained by averaging the last 4 hidden states of a pre-trained Bidirectional Encoder Representations from Transformers (BERT) model (Devlin et al., 2019). We use *bert-base-uncased* model (Uncased: 12-layer, 768-hidden, 12-heads, 110 M parameters). Note that the textual features represent the features extracted from both the title and review, but are processed separately[7].

Image Features: The images corresponding to each of the product reviews are first rescaled and normalized. The pre-processed images are sent as inputs to a Deep residual network ResNet-18 (He et al., 2016) pre-trained on ImageNet (Deng et al., 2009). The output from the image classification model is then passed through a Global Average Pool layer to extract the final image features. Each of the obtained image feature vector (I), $I \in \mathbb{R}^d$ where d = 256 is then reshaped to a size of (1 × d).

3.3 Modality Encoders

Textual Encoding: The title features (T) and review features (R) are passed through a stacked Bi-directional Long Short-Term Memory (Hochreiter and Schmidhuber, 1997) (BiLSTM$_{\mathrm{T}}$ and BiLSTM$_{\mathrm{R}}$), respectively. The hidden states of all elements of the sentence are extracted and represented by the shape $H_t \in \mathbb{R}^{n_t \times 2d_l \times h}$. Here, d_l represents the number of layers in each LSTM, n_t is the text length and h is the hidden size of each element.

Image Encoding: In a similar manner, the image features I are also passed through a stacked BiLSTM layer (BiLSTM$_{\mathrm{I}}$) to obtain a sequentially encoded context vector.

Shared Encoding: Finally, a shared BiLSTM (BiLSTM$_{\mathrm{Shared}}$) is employed to extract the complementary semantic dependencies between the two modalities. In order to concatenate the textual features T and R with the image features, T and R are passed through a dense layer (Dense$_{\mathrm{T}}$, Dense$_{\mathrm{R}}$) to contemporize the BERT embedding similar to I shape. Following which, they are concatenated and passed through the BiLSTM$_{\mathrm{Shared}}$.

3.4 Attention Mechanism

We employ the attention technique (Bahdanau et al., 2015) to concentrate on the words that contribute the most to the sentence meaning (Self-Att). In the case of textual modality (title, review), following each of the BiLSTM$_{\mathrm{T}}$, BiLSTM$_{\mathrm{R}}$ layers, we utilize Self-Att$_{\mathrm{T}}$ and Self-Att$_{\mathrm{R}}$ respectively. Similarly in the case of Image modality module, we have an image-specific attention layer Self-Att$_{\mathrm{I}}$ following the BiLSTM$_{\mathrm{I}}$. Alike, the BiLSTM$_{\mathrm{Shared}}$ is followed by Self-Att$_{\mathrm{Shared}}$.

Inter-segment Inter-modal Attention (A$_{\mathrm{S}}$) (Chauhan et al., 2020): A$_{\mathrm{S}}$ is applied to the outputs obtained from the dense layers (Dense$_2$, Dense$_{\mathrm{I}}$) of the text (title and review) and the image modalities, respectively. Both the modalities of our dataset are divided into some fixed number of segments beforehand as this can only be applied when both the modalities are divided into the same

[7] We also additionally experimented with Sentence Encoding with SBERT (Reimers and Gurevych, 2019), but the results were not satisfactory.

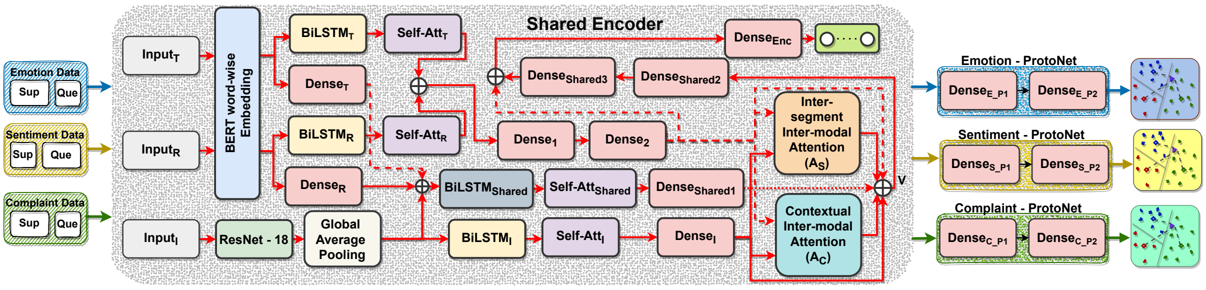


Figure 2: Architecture of the ProtoFed-MCI framework. The red path in the Shared Encoder is common across all three tasks. The Encoder's encoded output is projected via task-specific ProtoNet modules. T:Title, R:Review, I:Image

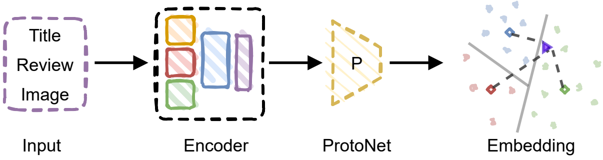


Figure 3: Prototypical network with the class-wise embeddings.

number of segments. A_S makes use of the relationship between distinct sentence segments across modalities. For each sentence, the feature vectors (i.e., $\in \mathbb{R}^d$) obtained from the two modalities i.e., $\in \mathbb{R}^{2 \times d}$ are concatenated and then split into n-segments (i.e.,$\in \mathbb{R}^{2 \times n}$).

Contextual Inter-modal Attention (A_C) (Ghosal et al., 2018): A single review often constitutes multiple sentences and can be a mixture of complaint and non-complaint sentences. The complaint information of such a review is dependent on the whole context. In a multi-modal framework, the interaction between modalities of the same sentence is crucial and so is the correlation between modalities across the contexts.

Attention Fusion: We linearly concatenate the outputs of A_S and A_C with the outputs of $Dense_2$, $Dense_I$ and $Dense_{Shared1}$. The obtained vector (V) is passed through two fully-connected layers ($Dense_{Shared2}$, $Dense_{Shared3}$) and finally concatenated with the outputs of the task specific $Dense_2$ layer from the text modality module.

3.5 Federated Meta-Learning Based Approach

Prototypical Networks (ProtoNet) Given a classification task Y ϵ [E, S, C], the data of each client K ϵ [N] is sampled into two mutually exclusive sets: Support set ($Sup^{KY} = X^{KSup}, Y^{KSup}$) for computing the prototypes and Query set ($Que^{KY} = X^{KQue}, Y^{KQue}$) for estimating the class-wise posteriors and computing the loss. The ProtoNet generates class-wise prototypes Z_c, via a mapping function $f_{\theta_P^{KY}} : \mathbb{R}^{768} \rightarrow \mathbb{R}^{128}$, and is given by

$$Z_c = \frac{1}{|X^{KSup_c}|} \sum f_{\theta_P^{KY}}(f_{\theta_E^K}(X^{KSup_c})) \qquad (3)$$

where X^{KSup_c} represents all inputs from the support set belonging to class c of task Y and $f_{\theta_E^K}$ represents the encoder weights. For every query sample $\{x, y\}$ ϵ Que^{KY}, the posterior probability $P_{f_{\theta_P^{KY}}}(y == c|x)$ for class c is given as:

$$\frac{exp(-d(f_{\theta_P^{KY}}(f_{\theta_E^K}(x)), Z_c))}{\sum_{c' \epsilon c} exp(-d(f_{\theta_K}(f_{\theta_E^K}(x)), Z_{c'}))} \qquad (4)$$

where d denotes Euclidean distance. Each client aims to minimize the negative-log of $P_{f_{\theta_P^{KY}}}(y = c|x)$ of the true class c for all three tasks Y ϵ [E, S, C] on their corresponding Query Set. Figure 3 shows the ProtoNet architecture with the class-wise embeddings.

Federated Multi-modal Complaint Identification (ProtoFed-MCI)

We develop our complaint identification model in FL setting to allow different companies to build an efficient complaint identification model cooperatively without having to deal with data sharing or privacy concerns[8]. Furthermore, by utilizing the meta-learning concept, even businesses with limited data can participate and benefit. The training process is as follows: an initial global model is shared with all the participating clients before

[8] We employ the horizontal federated learning framework, which is often used in cases where datasets have the same feature space but are sampled differently.

starting the training process. The communicated Proto-Fed MCI model consists of a encoder shared across all three tasks to encode the inputs and task-specific ProtoNet modules to project the encoded inputs in the projection space.

At each communication round (i = 1, 2, 3, . . , I), a random number of clients are selected. Each selected client K trains the communicated Proto-Fed MCI model, with weights $\theta_i = \{\theta_E^K, \theta_P^K\}$, on its dataset by minimizing the loss function across all three tasks and communicates back the locally updated weights. Following FedAvg (McMahan et al., 2017), the locally sourced weights are weighted-averaged based on the client's data count to obtain the updated global model. This is repeated for several communication rounds (I). The working of our proposed approach can be realized through the pseudo-code shown in Section B.1 in the **Appendix**.

4 Experiments, Results, and Analysis

4.1 Baselines

- **Single-task systems:** We develop a BERT-based single-task deep learning model for complaint detection with only text (STL$_\text{T}$). The BiLSTM output passes through the attention, dense and outer layer (task-specific). For the multi-modal single-task (complaint) model (STL$_\text{T+I}$), the image features are extracted using ResNet-18 model and the remaining architecture is similar to STL$_\text{T}$.

- **Multi-task systems:** We develop MTL$_\text{T}$ and MTL$_\text{T+I}$ models for multi-task baselines. The textual embeddings are generated from the pre-trained BERT model. For extracting the image features, ResNet-18 model is used. The system is composed of a fully-shared BiLSTM layer (256 units), followed by a shared attention layer. The output of the attention layer is passed through the three task-specific dense layers, which are then forwarded to respective output layers. The BERT-Shared Private Model (BSPMF$_\text{T}$) (Singh and Saha, 2021) is another suitable baseline for multi-task framework.

- **Centralized Model (MCI):** We have also compared the proposed Proto-Fed MCI with the centralized Multi-modal Complaint Identification (MCI) model. Please refer to Section B.2 in the **Appendix** where we show the centralized MCI architecture. In the centralized setting, there is no concept of data distribution across multiple clients as in the case of FL framework. For the complaint, emotion, and sentiment tasks, we compute the categorical-cross entropy losses.

- **Ablation models:** To understand the impact of emotion and sentiment prediction individually on complaint classification, we build dual-task variants of the centralized MCI model (MCI$_\text{CE}$, MCI$_\text{CS}$). The architectures are similar to the MCI system in other aspects. Furthermore, ablation studies are performed to analyze the importance of each of the special attention mechanisms used (A$_\text{S}$, A$_\text{C}$) in the MCI framework (MCI$_{\text{A}_\text{S}}$, MCI$_{\text{A}_\text{C}}$).

- **Federated Learning baseline:** Following the work of (Singh et al., 2021b), we implement their Fed-BMTL model on our dataset. Please note the Fed-BMTL model is a uni-modal architecture. They employ the FedAvg aggregation method for updating the global model parameters. We also developed the federated version of our MCI model (Fed-MCI) to depict the impact of FL in our proposed work. We also replaced the shared encoder module with a transformer-based encoder ALBERT[9] as one of the baselines (Fed-ALBERT).

4.2 Experimental Setup

We utilize PyTorch to implement our proposed framework and all baselines. We report the macro-F1 score and the accuracy of the models. Nvidia's GeForce RTX 2080 Ti was used for running the experiments.

Dataset: The complete dataset was segregated based on the product, and in order to replicate real-life data scarce situations, products only with more than 80 samples were selected as participating clients in the training process. Essentially the ProtoFed-MCI was trained on a total of 2371 samples distributed among 17 clients. The remaining 5519 samples were clubbed based upon the domains and were later divided into validation and test set. The model was validated on 1384 reviews forming the Electronics domain. The remaining 4135 reviews from six domains make up the six clients employed for testing the ProtoFed-MCI.

[9] https://huggingface.co/albert-base-v2

Proto-Fed MCI: All the BiLSTM modules used in the framework consists of 2 stacked BiLSTM of 128 units. Following the BiLSTM, Self-Att encodes BiLSTM's hidden state to a vector of length 64. For the Inter-segment Inter-modal Attention, the length of each segment is set to 4. Following the work in (Kumar et al., 2021), the Encoder encodes each input to a vector of length 768 and passes it to the ProtoNet. The ProtoNet consists of two dense layers with 128 units in each layer. We utilize $ReLU$ (Glorot et al., 2011) activation and a 20% *dropout* following the dense layers. At each communication round, K ϵ {5, 10, 15} clients were randomly selected and were shared the global model. Further, each client samples 5 support instances and 10 query instances from each class to train the ProtoFed-MCI model using SGD and $Adam$ (Kingma and Ba, 2015) with a learning rate of lr ϵ {1e-1, 1e-2, 1e-3, 1e-4}. For Fed-BMTL and Fed-MCI baselines the best results were achieved at communication round 25. The proposed framework, with 8.7M trainable parameters, takes 40 seconds to complete a single communication round involving 15 clients.

4.3 Results and Discussion

Figure 4 shows a sample t-SNE visualization (Van der Maaten and Hinton, 2008) of the embeddings learned by ProtoNet. To gain better insight, we visualize a sample of test reviews for the three different tasks (CI, ER, SA). It can be observed that the network is able to cluster the test reviews closely around the class prototypes[10].

Please note that the current work aims to improve the performance of CI with the help of the other two secondary tasks (ER and SA). Therefore, we state the results and analysis with CI strictly serving as the pivotal task in all the task combinations.

To evaluate the proposed approach, a series of experiments were carried out. Experiments were carried out in both centralized and federated learning settings. Thorough ablation research is conducted to assess the importance of each of the proposed architectural framework's attention mechanisms, as well as several variations of multi-modal and multi-task learning (e.g., STL_T, MCI_CE, etc.).

Table 1 depicts the classification results of all the baselines and the proposed framework. As can be observed, the ProtoFed-MCI outperforms the other baselines. The best results for the Proto-Fed MCI model is achieved in 25 communication rounds[11] with 15 randomly selected clients[12]. The learning rate for the best-performing model is 1e-2. The baselines (BSPMF, Fed-BMTL) are methodologies based on deep learning and naive federated learning techniques. The performances of these approaches are lower than the proposed model because they cannot effectively handle skewed datasets or generalize well for unknown data. Based on the results, it is clear that federated meta-learning can prove to be more effective than deep learning or simple federated learning methods in collaborative settings.

Model	**MCD Dataset**			
	Text		**Text+Image**	
	F1	**A**	**F1**	**A**
Centralized Baselines				
SOTA	85.58	86.39	-	-
BSPMF	85.43	86.17	-	-
Federated Learning Baseline				
Fed-BMTL	83.22	84.37	-	-
Fed-MCI	-	-	84.33	86.46
Fed-ALBERT	-	83.51	84.85	
Proposed approach				
Proto-Fed MCI	-	-	**89.00***	**89.06***

Table 1: Results of all the baselines and the proposed model in terms of macro F1-score(F1) and Accuracy(A) value. F1, A metrics are given in %. The maximum scores attained are represented by bold-faced values. The * signifies that these findings are statistically significant.

Ablation Experiment: Table 2 illustrates the results of the centralized framework (MCI) and the different ablation experiments performed. We design the ablation experiments to depict the reasonableness and usefulness of multi-task, multi-modal cues in the architecture. The multi-modal cues in the form of text and images significantly enhance the performance of single modality baselines (STL_T, MTL_T). This enhancement validates the proposed architecture's efficient usage of interaction among input modalities. This also emphasizes the significance of including multi-modal features for various opinionated text analysis tasks. In terms of varying ways of multi-tasking, the MCI model includes all three tasks (CI, ER, and SA), outper-

[10] Specifically for emotion, we plot the three major classes present in the test samples.

[11] We performed experiments with i ϵ {5, 10,15, 25}.

[12] The number of clients was varied with K ϵ {5, 10, 15}.

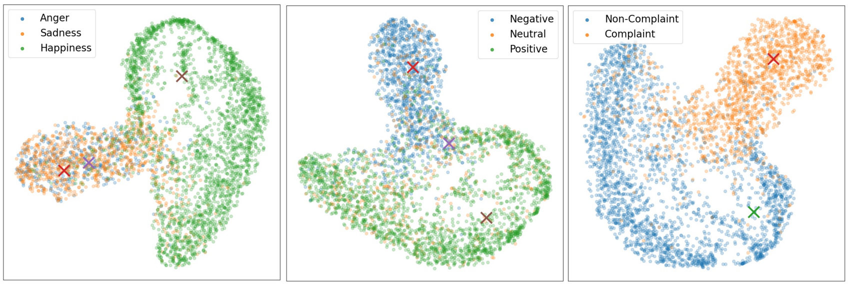


Figure 4: A t-SNE visualization of the embeddings learned by ProtoNet on the MCD dataset, for the three tasks. The symbol X denotes the respective class prototypes.

Model	**MCD Dataset**			
	Text		**Text+Image**	
	F1	**A**	**F1**	**A**
Single-task Baselines				
STL$_{T}$	86.78	86.96	-	-
STL$_{T+I}$	-	-	87.77	88.49
Multi-task Baselines				
MTL$_{T}$	88.32	89.58	-	-
MTL$_{T+I}$	-	-	90.05	90.71
Multi-modal Baselines				
MCI$_{CE}$	-	-	88.71	89.95
MCI$_{CS}$	-	-	87.43	88.65
MCI$_{A_S}$	-	-	89.38	90.21
MCI$_{A_C}$	-	-	89.20	90.35
Centralized Approach				
MCI	-	-	**91.20***	**91.95***

Table 2: Results of all the ablation studies performed on the proposed framework in terms of macro-F1 score (F1) and Accuracy(A) values. The maximum scores attained are represented by bold-faced values. The * signifies that these findings are statistically significant.

forms single-task variants and dual-task variants. Particularly in the dual-task variants, MCI$_{\text{CE}}$ outperforms MCI$_{\text{CS}}$. This can be driven by the fact that sentiment alone is often insufficient to convey entire information about the customer's mental state. For example, several emotions such as anger, contempt, fear, sadness, etc., can lead to negative sentiments about a product. As a result, sometimes, the discreteness or subtle differences in the state of mind cannot be properly determined and expressed by sentiment alone.

We also illustrate the significance of different attention mechanisms for the proposed MCI framework by conducting ablation studies (MCI$_{\text{A}_\text{S}}$, MCI$_{\text{A}_\text{C}}$). The results suggest that each of these factors considerably boosted the performance of the proposed Proto-Fed MCI and centralized MCI frameworks.

It should be observed that the federated meta-learning framework's performance is impacted by the distribution of data across multiple clients, which is not the issue for models built on the server in centralized settings. All of the results presented here are statistically significant[13] (Welch, 1947).

Comparison with State-of-the-art Technique (SOTA): We also compare the Proto-Fed MCI model with the existing state-of-the-art technique (Jin and Aletras, 2020) for single-task uni-modal CI as we are unaware of any other multi-modal complaint identification framework. Please note the SOTA model has been developed in a centralized setting. SOTA utilizes an array of neural language models boosted by the use of transformer networks. We re-implement it on the MCD dataset and report the results in Table 1. Our centralized model (MCI) achieves better results as compared to the SOTA technique. It can also be observed that the proposed ProtoFed-MCI outperforms the SOTA technique.

Please refer to Section C.1 in the **Appendix**, where we discuss possible explanations for the errors in the complaint prediction.

5 Conclusion and Future Work

In this work, we present a multi-task framework based on federated meta-learning for identifying complaints in a multi-modal context. The underlying system is a dual attention-based multi-modal multi-task framework for simultaneous optimization of complaint, emotion, and sentiment tasks. To encourage the study of multi-modal complaints, we extend the publically available multi-modal complaint dataset with 3928 reviews and user-uploaded images collected from the Amazon website and

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

annotated with the complaint, emotion, and sentiment classes. Experimental results indicate that the proposed approach outperforms the baselines and the state-of-the-art approaches in centralized and federated meta-learning settings.

In the future, we aim to extend this work with more fine-grained complaint severity annotation and identify complaints at the sentence level. We plan to work on the security aspect and resource usage in future research concerning federated meta-learning.

6 Limitations

We attempt to develop a federated meta-learning-based multi-modal multi-task framework for identifying complaints in data-scarce and distributed scenarios. Even though our model is able to outperform the baselines in centralized and federated meta-learning settings, there are some possible limitations to our approach as discussed below:

- The proposed model works in a multi-modal setup where text and images for every sample are necessary for training. The model will not work with incomplete modalities, that is, if only text or image of the review is available.

- Additionally, people frequently use sarcasm to critique products bought online. But the specific class of sarcasm could not be considered as only a few curated samples were sarcastic. Hence, complaints with sarcastic remarks were often misclassified by the proposed model.

- In the current setup, we work with product reviews from the public domain; the model is not explicitly trained for any specific domain such as financial or healthcare services.

- To show the efficacy of the proposed federated meta-learning model, we work in a simulated environment by using a product review dataset available in the public domain. Since the proposed model is a privacy-conscious federated meta-learning model, it needs further training in practical scenarios with distributed setup.

- Furthermore, the model was trained and evaluated on only English language reviews. To accommodate other languages, further training in other languages would be necessary.

Acknowledgement

This publication is an outcome of the R&D work undertaken in the project under the Visvesvaraya Ph.D. Scheme of Ministry of Electronics & Information Technology, Government of India, being implemented by Digital India Corporation (Formerly Media Lab Asia).

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Dushyant Singh Chauhan, SR Dhanush, Asif Ekbal, and Pushpak Bhattacharyya. 2020. Sentiment and emotion help sarcasm? a multi-task learning framework for multi-modal sarcasm, sentiment and emotion analysis. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4351–4360.

Andrew D Cohen and Elite Olshtain. 1993. The production of speech acts by efl learners. *Tesol Quarterly*, 27(1):33–56.

Kristof Coussement and Dirk Van den Poel. 2008. Improving customer complaint management by automatic email classification using linguistic style features as predictors. *Decision Support Systems*, 44(4):870–882.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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)*, pages 4171–4186. Association for Computational Linguistics.

Paul Ekman, Wallace V Friesen, Maureen O'sullivan, Anthony Chan, Irene Diacoyanni-Tarlatzis, Karl Heider, Rainer Krause, William Ayhan LeCompte, Tom Pitcairn, Pio E Ricci-Bitti, et al. 1987. Universals and cultural differences in the judgments of facial expressions of emotion. *Journal of personality and social psychology*, 53(4):712.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning*, pages 1126–1135. PMLR.

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

Deepanway Ghosal, Md Shad Akhtar, Dushyant Chauhan, Soujanya Poria, Asif Ekbal, and Pushpak Bhattacharyya. 2018. Contextual inter-modal attention for multi-modal sentiment analysis. In *proceedings of the 2018 conference on empirical methods in natural language processing*, pages 3454–3466.

Xavier Glorot, Antoine Bordes, and Yoshua Bengio. 2011. Deep sparse rectifier neural networks. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 315–323.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Mali Jin and Nikolaos Aletras. 2020. Complaint identification in social media with transformer networks. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 1765–1771. International Committee on Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Abhishek Kumar, Asif Ekbal, Daisuke Kawahra, and Sadao Kurohashi. 2019. Emotion helps sentiment: A multi-task model for sentiment and emotion analysis. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

Manoj Kumar, Varun Kumar, Hadrien Glaude, Cyprien de Lichy, Aman Alok, and Rahul Gupta. 2021. Protoda: Efficient transfer learning for few-shot intent classification. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 966–972. IEEE.

M Lailiyah, S Sumpeno, and IK E Purnama. 2017. Sentiment analysis of public complaints using lexical resources between indonesian sentiment lexicon and sentiwordnet. In *2017 International Seminar on Intelligent Technology and Its Applications (ISITIA)*, pages 307–312. IEEE.

Brenden Lake, Ruslan Salakhutdinov, Jason Gross, and Joshua Tenenbaum. 2011. One shot learning of simple visual concepts. In *Proceedings of the annual meeting of the cognitive science society*, volume 33.

Michael Lewis, Jeannette M Haviland-Jones, and Lisa Feldman Barrett. 2010. *Handbook of emotions*. Guilford Press.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. 2017. Communication-efficient learning of deep networks from decentralized data. In *Artificial Intelligence and Statistics*, pages 1273–1282. PMLR.

Daniel Preotiuc-Pietro, Mihaela Gaman, and Nikolaos Aletras. 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*, pages 5008–5019. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *CoRR*, abs/1908.10084.

Apoorva Singh, Soumyodeep Dey, Anamitra Singha, and Sriparna Saha. 2022. 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*, pages 12163–12171. AAAI Press.

Apoorva Singh and Sriparna Saha. 2021. Are you really complaining? a multi-task framework for complaint identification, emotion, and sentiment classification. In *International Conference on Document Analysis and Recognition*, pages 715–731. Springer.

Apoorva Singh, Sriparna Saha, Mohammed Hasanuzzaman, and Anubhav Jangra. 2021a. Identifying complaints based on semi-supervised mincuts. *Expert Systems with Applications*, page 115668.

Apoorva Singh, Tanmay Sen, Sriparna Saha, and Mohammed Hasanuzzaman. 2021b. Federated multi-task learning for complaint identification from social media data. In *Proceedings of the 32nd ACM Conference on Hypertext and Social Media*, pages 201–210.

Raghvendra Pratap Singh, Rejwanul Haque, Mohammed Hasanuzzaman, and Andy Way. 2020. Identifying complaints from product reviews: A case study on hindi. In *Proceedings of The 28th Irish Conference on Artificial Intelligence and Cognitive Science, Dublin, Republic of Ireland, December 7-8, 2020*, volume 2771 of *CEUR Workshop Proceedings*, pages 217–228. CEUR-WS.org.

Jake Snell, Kevin Swersky, and Richard S Zemel. 2017. Prototypical networks for few-shot learning. *arXiv preprint arXiv:1703.05175*.

Suhatati Tjandra, Amelia Alexandra Putri Warsito, and Judi Prajetno Sugiono. 2015. Determining citizen complaints to the appropriate government departments using knn algorithm. In *2015 13th International Conference on ICT and Knowledge Engineering (ICT & Knowledge Engineering 2015)*, pages 1–4. IEEE.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. 2016. Matching networks for one shot learning. *Advances in neural information processing systems*, 29:3630–3638.

Bernard L Welch. 1947. The generalization of 'student's'problem when several different population variances are involved. *Biometrika*, 34(1-2):28–35.

Qiang Yang, Yang Liu, Tianjian Chen, and Yongxin Tong. 2019a. Federated machine learning: Concept and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(2):1–19.

Wei Yang, Luchen Tan, Chunwei Lu, Anqi Cui, Han Li, Xi Chen, Kun Xiong, Muzi Wang, Ming Li, Jian Pei, et al. 2019b. Detecting customer complaint escalation with recurrent neural networks and manually-engineered features. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Industry Papers)*, pages 56–63.

A Multi-modal Complaint Dataset

A.1 Sample Instances from the dataset

Table 3 shows sample instances along with the corresponding emotion and sentiment labels from the Multi-modal Complaint Dataset.

A.2 Qualitative Perspective

We examine a few samples from the dataset that illustrate the requirement of sentiment-emotion tasks and multi-modal analysis of complaints in the sections below.

Significance of Multi-modality: Figure 5(a) shows two instances where the complaint is articulated through the incorporation of both the modalities (text and image). In the first instance, the textual modality implies a substandard product. The image affirms the reason for disapproval. So the breach of expectation is amplified by the usage of both text and image modalities. In the second instance, the textual modality suggests a neutral review whereas, the image modality implies a deceptive claim. Both the cases signify that multiple sources of information could provide supplementary indicators for CI. The presence of contrasting input from various modalities increases the model's capacity to learn the selective patterns that underpin this complex process.

Significance of Emotion and Sentiment: Figure 5(b) shows two sample instances from the MCD dataset that justify the need to incorporate emotion and sentiment information into the complaint identification framework. In the first example, the mixed emotions of the customer could be confusing, but the emotion and sentiment labels provide clarity about the customer's state of mind. Similarly, in the second example, the emotion and sentiment labels also provide better insight regarding the customer's negative review. Our dataset's inclusion of emotion and sentiment information enables the models to employ additional information when reasoning about complaints.

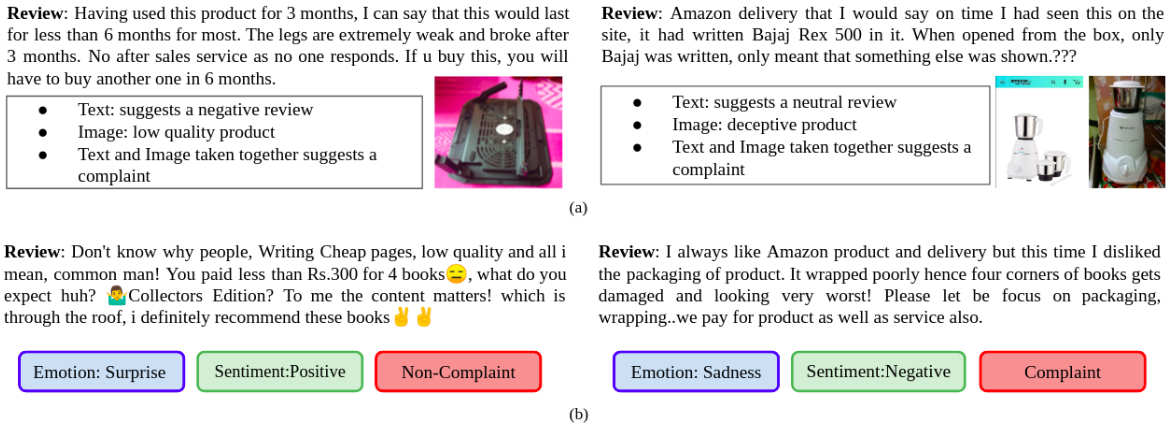
B Proposed Methodology

B.1 Proto-Fed MCI algorithm

The pseudo-code of Proto-Fed MCI model is shown in Algorithm 1.

B.2 MCI Architecture

Figure 6 depicts the centralized MCI framework.



Review: Having used this product for 3 months, I can say that this would last for less than 6 months for most. The legs are extremely weak and broke after 3 months. No after sales service as no one responds. If u buy this, you will have to buy another one in 6 months.

- Text: suggests a negative review
- Image: low quality product
- Text and Image taken together suggests a complaint

Review: Amazon delivery that I would say on time I had seen this on the site, it had written Bajaj Rex 500 in it. When opened from the box, only Bajaj was written, only meant that something else was shown.???

- Text: suggests a neutral review
- Image: deceptive product
- Text and Image taken together suggests a complaint

(a)

Review: Don't know why people, Writing Cheap pages, low quality and all i mean, common man! You paid less than Rs.300 for 4 books😑, what do you expect huh? 🤷‍♂️Collectors Edition? To me the content matters! which is through the roof, i definitely recommend these books✌✌

Emotion: Surprise | Sentiment:Positive | Non-Complaint

Review: I always like Amazon product and delivery but this time I disliked the packaging of product. It wrapped poorly hence four corners of books gets damaged and looking very worst! Please let be focus on packaging, wrapping..we pay for product as well as service also.

Emotion: Sadness | Sentiment:Negative | Complaint

(b)

Figure 5: (a) Significance of multi-modality, (b) Significance of emotion and sentiment

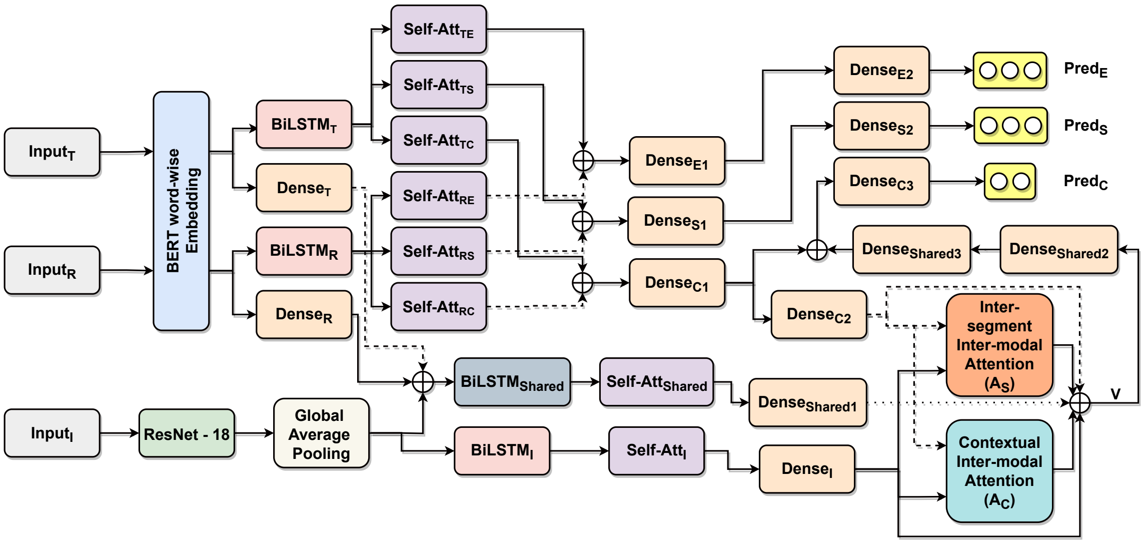


Figure 6: Architecture of the Multi-modal Complaint Identification (MCI) framework. T:Title, R:Review, I:Image, E:Emotion, S:Sentiment, C:Complaint.

Title	Review	Label	Emotion	Sentiment
Better then expectation	I ordered it few days ago and when I receive it. I almost surprised to see the quality and finish. It looks classy and stylish. It fits well on me. I recommend it to everyone to buy and wear. This is such a great deal	Complaint	Surprise	Positive
Trustable and comfortable	All in all this product was satisfying and I'm happy with my purchase from Urbano. Product looks new and nice and stretchable and comfortable too. Fitted perfectly you can trust this product and brand.	Non-Complaint	Happiness	Positive
Cable and packaging is worse	Cooling pad is good but the cable that they gave is worse and the packaging is also worse. Cooling pad is okay!	Complaint	Sadness	Negative
Be careful while using this product.	Be careful when using this heater. This heater is suitable for small rooms.	Complaint	Fear	Neutral
Average book	Over hyped book, not recommended for people who already are aware of all these basic financial principles.	Non-Complaint	Sadness	Negative

Table 3: Sample instances from the MCD dataset.

Algorithm 1: Proto-Fed MCI

```
for communication round i $\epsilon$ I do
    Sample K clients
    for each Client k $\epsilon$ K in parallel do
        Receive Global model weights
        for task Y $\epsilon$ {E, S, C} do
            Sample Support and Query Set
            Update weights by maximizing 3
        end for
        Return locally updated weights
    end for
    Update Global model via FedAvg
end for
```

C Results and Analysis

C.1 Error Analysis

The following are possible explanations for the errors in the complaint prediction:

- Ironical Instances: Instances having ironic or comments where the underlying tone is positive or neutral, but the instance is of complaint type, the MCI model inaccurately predicts such instances as a complaint. For example, *'Biscuits with oil might be a rare combination of Amazon nowadays'*. For the above sentence, the predicted class is non-complaint, but the actual class is complaint. One of the reasons could be neutral undertone and usage of less explicit words to signify complaint.
- Multifold Sentences: The majority of the sentences in the MCD dataset are lengthy and heterogeneous, including diverse emotions in a single instance. In such scenarios, learning specific complaint features becomes challenging. For example, *'Although it's not a Microsoft genuine product, it's good quality and comfortable to use. Price is reasonable too when compared to its build quality and features.'*; predicted class: complaint. The correct class for the preceding example is non-complaint, but because of the composite nature and contrasting context of the statement, the MCI model misclassifies it as a complaint.
- Skewness of Dataset: The MCD dataset's imbalanced class distribution influences the proposed MCI model's predictions. The complaint class (37.5%) is under-represented as compared to the non-complaint class due to which the model is biased towards the non-complaint class. This conforms with the practical scenarios where complaints occur less frequently compared to non-complaints.