Sentiment and Emotion-aware Multi-modal Complaint Identification

Apoorva Singh[1], **Soumyodeep Dey**[*2], **Anamitra Singha**[*2], **Sriparna Saha**[1]

[1] Indian Institute of Technology Patna, India
[2] National Institute of Technology Durgapur, India
(apoorva_1921cs19, sriparna)@iitp.ac.in, (sd18u10528, as18u10604)@btech.nitdgp.ac.in

Abstract

The expression of displeasure on a consumer's behalf towards an organization, product, or event is denoted via the speech act known as complaint. Customers typically post reviews on retail websites and various social media platforms about the products or services they purchase, and the reviews may include complaints about the products or services. Automatic detection of consumers' complaints about items or services they buy can be critical for organizations and online merchants. Since they can use this insight to meet the customers' requirements, including handling and addressing the complaints. Previous studies on Complaint Identification (CI) are limited to text. Images posted with the reviews can provide cues to better identify complaints, thus emphasizing the importance of incorporating multi-modal inputs into the process. Furthermore, the customer's emotional state has a significant impact on the complaint expression, since any speech act is generally influenced by emotions. As a result, the impact of emotion and sentiment on automatic complaint identification must also be investigated. One of the major contributions of this work is a new dataset- Complaint, Emotion, and Sentiment Annotated Multi-modal Amazon Reviews Dataset (CESAMARD) which is a collection of opinionated texts (reviews) and images of the products posted on the website of the retail giant Amazon. We present an attention-based multi-modal, adversarial multi-task deep neural network model for complaint detection to demonstrate the utility of the multi-modal dataset. Experimental results indicate that the multi-modality and multi-tasking complaint identification outperforms uni-modal and single-task variants.

Introduction

Nowadays, social media platforms and online e-commerce websites provide users the freedom to express their opinions and observations towards a product, an organization, or an event. Customers who plan to buy a product often base their decision on customer reviews (Preotiuc-Pietro, Gaman, and Aletras 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 also to resolve any product-related concerns. This could also benefit customer by providing recommendations on the quality of goods or services that they intend to buy. Identifying complaint texts in natural language is critical for developers of downstream applications such as chatbots (Lailiyah, Sumpeno, and Purnama 2017), commercial organizations or counsels 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, Haviland-Jones, and Barrett 2010). Generally, sentiment and emotion are seen as two different tasks (Majumder et al. 2019; Soleymani et al. 2017; Do et al. 2019) but sentiment and emotion are two closely related problems. However, emotion recognition is a far more subtle and fine-grained analysis than sentiment classification (Kumar et al. 2019). Emotion, together with sentiment, offers a greater insight into the customer's frame of mind. For example, a question or an inquiry would only have neutral sentiment but emotion could be anger or even anticipation. The correlation between emotion and sentiment motivates us to consider customer's sentiment and emotion while analyzing complaints. We learn the tasks of complaint identification, emotion recognition, and sentiment classification in a multi-task setting to further examine the relationship between complaint, emotion, and sentiment.

Due to technological advancements, people can now share their views or opinions in multiple modalities. The fact that nearly every content sharing and e-commerce platform allows users to accompany an opinion or review with multiple media forms, and that every mobile phone has that capability, exemplifies the superiority of a multi-modal form of communication in terms of ease of conveying and understanding information.

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

- We curate a new dataset called CESAMARD for aiding multi-modal complaint identification research with good quality annotations, including emotion and sentiment classes.
- We demonstrate using various instances, the usefulness of incorporating multi-modal sources of information as well as sentiment and emotion class of the review while identifying complaints.
- We propose a dual attention-based multi-task adversarial

[*] These authors contributed equally.

learning framework for multi-modal complaint, emotion, and sentiment analysis. The dual attention mechanism utilizes inter-segment inter-modal attention and contextual inter-modal attention. Complaint Identification (CI) is treated as the primary task in our multi-task framework, whereas Emotion Recognition (ER) and Sentiment Analysis (SA) are considered as supplementary (i.e., auxiliary) tasks. Multi-modal and multi-task CI surpasses uni-modal and single-task CI considerably.

- We present the state-of-the-art for automatically identifying complaints in the multi-modal scenario.

Related Work

Almost all e-commerce websites allow customers to publicly voice their opinions and thoughts concerning products on their websites and social media channels. E-commerce corporations and online retailers want to detect complaints based on product reviews for their own profit. Due to the plethora of information available online, companies and retailers find it challenging to identify complaints and address the issues directly. Additionally, detecting complaints on social media entails detecting complaints from unstructured and noisy text snippets with character limitations, usage of random abbreviations, ironical expressions, allegations (Pawar et al. 2015), making it a laborious and tedious task. According to an analysis of relevant literature, a multi-modal approach to complaint detection, as opposed to text-based classification, is a new approach. In this context, text-based complaints have been previously analyzed based on semi-supervised strategies, different domains, degree of urgency, and feedback likelihood (Preotiuc-Pietro, Gaman, and Aletras 2019; Singh et al. 2021; Tjandra, Warsito, and Sugiono 2015; Yang et al. 2019), (Jin and Aletras 2020).

Recent studies (Qureshi et al. 2019; Akhtar, Ekbal, and Cambria 2020; Ghosh, Ekbal, and Bhattacharyya 2021) have demonstrated the effectiveness of multi-task systems by learning numerous correlated tasks simultaneously. In the area of emotion and sentiment recognition in human conversations, multi-modality has become a popular research area (Poria et al. 2018; Akhtar et al. 2019; Chauhan et al. 2020). Despite the fact that 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 a lack of multi-modal datasets.

In this work, we initially collect the publicly posted reviews and images posted by customers from the Amazon India[1] website and then manually annotate each review with the complaint, emotion, and sentiment labels. Subsequently, we propose a deep learning-based framework with dual attention mechanisms to leverage information from images as well as complain texts for identifying complaints in a multi-modal multi-task framework. To the best of our knowledge, this is the first attempt to solve the multi-modal complaint detection problem in a deep multi-task framework.

Complaint, Emotion, and Sentiment Annotated Multi-modal Amazon Reviews Dataset (CESAMARD)

The existing complaint datasets (Preotiuc-Pietro, Gaman, and Aletras 2019; Singh et al. 2020) deal with text-based complaints only and do not consider the sentiment and emotion classes of the complaints. For building a multi-modal multi-task framework, we curate a multi-modal complaint dataset (CESAMARD) which has been labeled with emotion and sentiment classes, one of the contributions of our current work. Here, we discuss the details of the CESAMARD dataset.

Data Collection

To the best of our knowledge, there is no existing (freely available) annotated complaint dataset that comprises customer-posted reviews and images of the items purchased. To build an annotated multi-modal dataset, we initially gathered reviews from Amazon India's website. In order to collect the product reviews and the corresponding review image URLs, we used Scrapy[2], a free and open-source web-crawling framework. It accepts a product name as an input and returns product reviews and associated images. The reviews were collected based on five different domains (Books, Edible, Electronics, Fashion, and Miscellaneous) for a more fine-grained gold standard dataset. To eliminate noise (HTML tags and special characters) from the textual portion of the dataset, we had to perform some pre-processing operations on the corpus. The Unicode emojis in the product reviews were converted to emoji short text with the help of a Python module called Emoji[3]. The Python library Textblob[4], was used to correct the spellings of many improperly spelled words.

Data Annotation

We assigned three graduate students 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. To begin, we define an annotation task to determine whether or not a review contains a complaint. 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 earlier linguistic research (Cohen and Olshtain 1993) for complaint annotation: ”A complaint presents a state of affairs that breaches the writer's favourable expectation”. For the emotion annotation of the CESAMARD dataset, we consider Ekman's six basic emotions (Ekman et al. 1987) (*anger, disgust, fear, happiness, sadness, and surprise*). For the sentiment annotation, we consider three sentiment classes (*negative, neutral, positive*).

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

[2] https://scrapy.org/

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

[4] https://pypi.org/project/textblob/0.9.0/

Review	Domain	Label	Emotion	Sentiment
Received the book and started reading it after a few weeks. Many pages of the book are blank and not readable due due to poor printing quality. Disappointed!	Books	Complaint	Sadness	Negative
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.	Fashion	Non-Complaint	Happiness	Positive
It's worth buying but little big.	Misc	Non-Complaint	Happiness	Neutral

Table 1: Sample instances from the CESAMARD dataset.

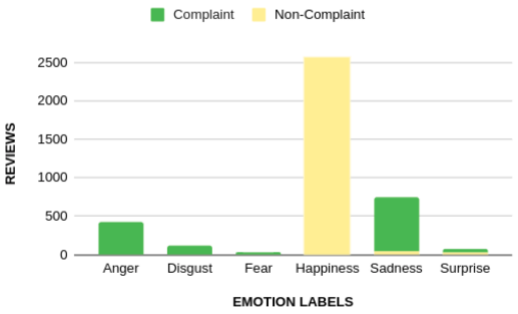


Figure 1: Distribution of emotion labels across CESAMARD dataset.

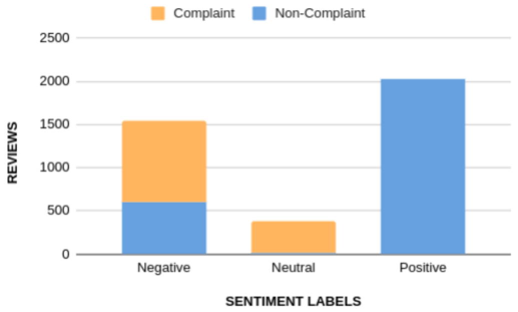


Figure 2: Distribution of sentiment labels across CESAMARD dataset.

The majority voting technique was used for selecting the final complaint, emotion, and sentiment labels. Reviews with no common emotion or sentiment label (as determined by the three annotators) were excluded from the final annotated dataset. On the annotated dataset, we computed the Fleiss' kappa (Fleiss 1971) scores to assess inter-rater agreement among the three annotators. We attain agreement scores of 0.83, 0.69, and 0.81 on the complaint, emotion, and sentiment task, respectively, which indicate that the annotations are of good quality[5].

The CESAMARD dataset now comprises of 3962 reviews with the corresponding complaint, emotion, and sentiment labels. It consists of 2641 reviews in the non-complaint category and 1321 reviews in the complaint category. Table 1 shows sample instances along with the corresponding emotion and sentiment labels. Each record in the CESAMARD dataset consists of the page URL, domain, image URL, review title, review text, and their corresponding annotated complaint, sentiment, and emotion labels. Distributions of emotion and sentiment labels across the CESAMARD dataset are shown in Figure 1 and 2.

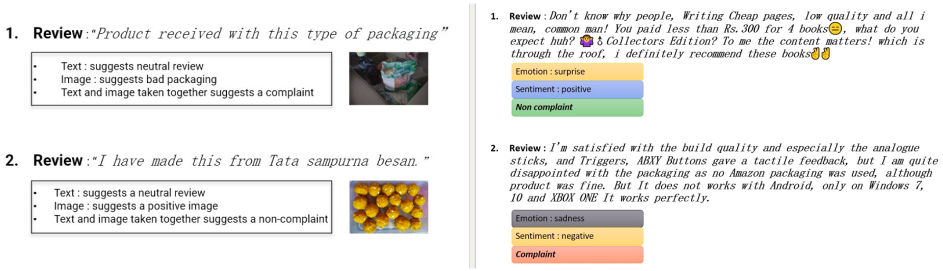
Theoretical Aspects

In this paper, we aim to analyze the usefulness of incorporating multi-modal sources of information as well as the sentiment and emotion class of the review while identifying complaints. We illustrate our point of CI benefiting from sentiment and emotion-aware multi-modal analysis using a few examples from our proposed dataset.

Significance of Multi-modality Figure 3(a) shows two instances where the complaint is articulated through the incorporation of both the modalities (text and image). In the first instance, the image implies bad packaging. The textual modality, on the other hand, lacks any obvious indicator of disapproval. So the textual review alone is not a correct indicator of breach of expectation. In the second instance, the textual modality suggests a neutral review whereas, the image modality implies a positive claim. In both cases, there is a mismatch between the two modalities which signifies 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 3(b) shows two sample instances from the CESAMARD dataset that justify the need of incorporating emotion and sentiment 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 give a 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.

[5] When an emotion is conveyed in a review does not fall into one of the six categories (anger, disgust, fear, happiness, sadness, and surprise), the annotators label it with the next closest emotion linked with the review.



1. **Review** : *"Product received with this type of packaging"*

- Text : suggests neutral review
- Image : suggests bad packaging
- Text and image taken together suggests a complaint

2. **Review** : *"I have made this from Tata sampurna besan."*

- Text : suggests a neutral review
- Image : suggests a positive image
- Text and image taken together suggests a non-complaint

1. **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

2. **Review** : *I'm satisfied with the build quality and especially the analogue sticks, and Triggers, ABXY Buttons gave a tactile feedback, but I am quite disappointed with the packaging as no Amazon packaging was used, although product was fine. But It does not works with Android, only on Windows 7, 10 and XBOX ONE It works perfectly.*

Emotion : sadness

Sentiment : negative

Complaint

Figure 3: (a) Significance of incorporating multi-modal information, (b) Significance of incorporating emotion and sentiment information

Proposed Methodology

In this section, we define our problem and go over the details of the multi-task multi-modal architecture for identifying complaints. The various components of the architecture are detailed in the sections that follow. The overall framework is shown in Figure 4.

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 $(r_{\mathrm{k}}, e_{\mathrm{k}}, s_{\mathrm{k}}, c_{\mathrm{k}})_{k=1}^{R}$ be a set of R reviews where e_{k}, s_{k}, and c_{k} represent the matching emotion, sentiment and complaint labels for $r_{\mathrm{k}}{}^{\mathrm{th}}$ tweet, respectively. Here, $r_{\mathrm{k}} \epsilon R$, $e_{\mathrm{k}} \epsilon E$ (emotion classes), $s_{\mathrm{k}} \epsilon S$ (sentiment classes) and $c_{\mathrm{k}} \epsilon C$ (complaint classes).

Our multi-task learning framework's objective is to maximize the function f (Eqn 1) that draws a new instance r_{k} to its fitting emotion label e_{k}, sentiment label s_{k} and complaint label c_{k} simultaneously.

$$argmax(\Pi_{k=0}^{R} P(s_k, e_k, c_k | r_k; \theta)), \tag{1}$$

where r_k is the input sentence whose complaint label (c_k), emotion label (e_k) and sentiment label (s_k) are to be predicted. θ denotes the model's parameters we aim to optimize.

Multi-modal Complaint Identification (MCI)

The proposed MCI framework consists of five principal components: (i) Multi-modal Feature Extraction (MFE) extracts the features from the two modalities (text and image), (ii) Modality Encoders provide corresponding modality encodings by using the uni-modal features acquired from MFE as input, (iii) Attention Mechanism that incorporates two significant attention mechanisms namely, inter-segment inter-modal attention and contextual inter-modal attention, (iv) Adversarial Loss which ensures the shared layers and task-specific feature space remain mutually exclusive. (v) Output Layer which consists of output medium for the three tasks to obtain a generalized representation throughout all the tasks.

Feature Extraction

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

Text Features The textual features (T) were obtained using state-of-the-art Sentence Bidirectional Encoder Representations from Transformers (SBERT) model[6]. SBERT architecture is a multi-layer bidirectional Transformer encoder (Reimers and Gurevych 2019) which is a modification of the pre-trained BERT network illustrated in (Devlin et al. 2019). We employ *stsb-bert-base*, a *bert-base-uncased* model which was optimized for Semantic Textual Similarity (STS). SBERT encodes a single sentence into a vector of length 768.

Image Features The images corresponding to each of the product reviews are first rescaled and normalized. The preprocessed images are sent as inputs to an ImageNet (Deng et al. 2009) pre-trained ResNet-152 (He et al. 2016) image classification model. The output from the image classification model is then passed through a Global Average Pool layer thus extracting the image features. Each of the obtained image feature vector (I), $I \in \mathbb{R}^d$ where d = 2048 is then reshaped to a size of (1 X d).

Modality Encoders

Textual Encoding The features T and I derived from each of the modalities associated with a review are then passed through three Bi-directional Long Short-Term Memory (BiLSTMs) (Hochreiter and Schmidhuber 1997). For the textual modality, the final hidden state matrix of a review is obtained as $H_t \in \mathbb{R}^{n_t X 2d_l}$. Here, d_l represents the number of hidden units in each LSTM and n_t is the text length.

Image Encoding The image features (I) obtained are passed through a private BiLSTM layer as well as the shared BiLSTM layer$_2$. The image features are passed through the private BiLSTM layer (BiLSTM layer$_{\mathrm{P}}$) to obtain hidden states by sequentially encoding these representations. Both the image and textual feature vectors are passed through the shared BiLSTM layer$_2$ to obtain hidden states with complementary semantic dependencies between the two modalities. The final hidden state matrix of the image obtained from the private BiLSTM is $H_i \in \mathbb{R}^{n_i X 2d_l}$ where n_i is the dimension of the image representations and d_l is the number of hidden units in each LSTM. Similarly, the final hidden state matrix of the shared BiLSTM layer is $H_{sh} \in \mathbb{R}^{n_{sh} X 2d_l}$ where $n_{sh} \in \mathbb{R}^{(n_i + n_t)}$.

[6] https://github.com/UKPLab/sentence-transformers

Attention Mechanism

We employ the attention technique (Bahdanau, Cho, and Bengio 2015) to concentrate on the words that contribute the most to the sentence meaning (Att). In the case of textual modality, following the shared BiLSTM layer$_1$, we employ three independent task-specific attention layers (Att$_{\mathrm{COM}}$, Att$_{\mathrm{EMO}}$, Att$_{\mathrm{SENT}}$). After the shared BiLSTM layer$_2$ there is one attention layer (Att$_{\mathrm{Sh}}$) and another attention layer (Att$_{\mathrm{img}}$) after the task-specific BiLSTM layer (BiLSTM$_{\mathrm{p}}$) in the image modality module.

Inter-segment Inter-modal Attention (A$_{\mathrm{S}}$) Motivated by the work in (Chauhan et al. 2020), a special attention mechanism is applied to the outputs obtained from the dense layers of the image and the text modalities, respectively. It is called inter-segment inter-modal attention. We need to divide both the modalities of our dataset into some fixed number of segments beforehand as this can only be applied when both the modalities are divided into the same number of segments. The main objective of this attention mechanism is to learn the dependency of the feature vectors of a segment of the visual modality with the feature vectors of the same segment from the text modality. For each sentence, the feature vectors (i.e., $\in \mathbb{R}^{\mathrm{d}}$) obtained from the two modalities i.e., $\in \mathbb{R}^{2 \times \mathrm{d}}$ are concatenated and then split into into n-segments (i.e.,$\in \mathbb{R}^{2 \times \mathrm{n}}$).

Contextual Inter-modal Attention (A$_{\mathrm{C}}$) Many times, a single review 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. This encourages us to apply attention to the contributing adjacent sentences, as well as multi-modal representations of the same could benefit the system. 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. This attention mechanism is called contextual inter-modal attention and is motivated by the work in (Ghosal et al. 2018).

Attention Fusion We linearly concatenate the computed A$_{\mathrm{S}}$ and A$_{\mathrm{C}}$ vectors with the following X$_{\mathrm{sh}}$, W$_{\mathrm{I}}$ and U$_{\mathrm{T}}$ output vectors and pass the concatenated vector (V) through a fully-connected layer. The following equations can be used to represent the flow of information:

$$W_{\mathrm{I}} = FC_{\mathrm{P}}(Att_{\mathrm{IMG}}(BiLSTM_{\mathrm{P}}) \tag{2}$$

$$X_{\mathrm{sh}} = FC_{\mathrm{S}}(Att_{\mathrm{sh}}(SBiLSTM_2) \tag{3}$$

$$U_{\mathrm{T}} = FC_{\mathrm{S}}(FC_{\mathrm{COM}}(Att_{\mathrm{COM}}(SBiLSTM_1)) \tag{4}$$

$$A_{\mathrm{S}} = [W_{\mathrm{I}}, U_{\mathrm{T}}] \tag{5}$$

$$A_{\mathrm{C}} = [W_{\mathrm{I}}, U_{\mathrm{T}}] \tag{6}$$

$$V = [W_{\mathrm{I}}; U_{\mathrm{T}}; A_{\mathrm{S}}; A_{\mathrm{C}}] \tag{7}$$

Here, ';' indicates the linear concatenation and ',' indicates the inputs of that layer.

Adversarial Loss

The adversarial loss strives to mutually exclude the feature space of shared and task-specific layers. We use a similar methodology to (Liu, Qiu, and Huang 2017), in which a task discriminator (Z) maps the shared feature to its primary task. The adversarial loss is calculated as:

$$L_{adv} = min(max(\sum_{p=1}^{P}\sum_{q=1}^{Q}(z_q^p * log[Z(E(x_q^p))]))) \text{ where,}$$

P denotes the type of tasks, z_q^p signifies the actual label amongst P, and x_q^p is the q^{th} example for task p. The gradient reversal layer (Ganin and Lempitsky 2015) handles the min-max optimization problem.

Output Layer

The final predictions for emotion and sentiment tasks are generated by averaging the adversarial outputs (E_{adv} and S_{adv}) and the shared outputs (E_s and S_s). In the case of the complaint predictions, the multi-modal complaint output (C_m) is also taken into account, i.e., the average of C_{adv}, C_s and C_m.

Calculation of Loss: For the complaint, emotion, and sentiment tasks, we compute the categorical-cross entropy (CE) losses. The integrated loss function (L) of our proposed MCI system is realized as follows: $L = x * L^{\mathrm{C}}{}_{\mathrm{CE}} + y * L^{\mathrm{E}}{}_{\mathrm{CE}} + z * L^{\mathrm{S}}{}_{\mathrm{CE}}$. We aggregate the weighted sum of the losses from the three tasks to compute the overall loss. Here, x, y, and z are constants ranging from 0 to 1 that determine the loss weights that represent the per-task loss-share to the overall loss.

Experiments, Results, and Analysis

Baselines

- **Single-task systems:** We develop a SBERT-based single-task deep learning model for complaint detection with only text (STL$_{\mathrm{T}}$). The BiLSTM output passes through the attention, dense and outer layer (task-specific). For the multi-modal single-task (complaint) model (STL$_{\mathrm{T+I}}$)(Pranesh and Shekhar 2020), to extract the image features Visual Geometry Group Network[7] (VGG19) model has been used and the remaining architecture is similar to STL$_{\mathrm{T}}$.
- **Multi-task systems:** We develop MTL$_{\mathrm{T}}$ and MTL$_{\mathrm{T+I}}$ models for multi-task baselines. The textual embeddings are generated from the pre-trained GloVe[8] (Pennington, Socher, and Manning 2014). The embedding layer's output is forwarded to the word sequence encoder, which analyzes it to extract contextual knowledge from the sentence. For extracting the image features VGG19[9] 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 into the three task-specific dense layers, which are then forwarded to respective output layers. The BERT-Shared Private Model (BSPMF$_{\mathrm{T}}$) (Singh and Saha 2021) is another suitable baseline for multi-task framework.

[7] https://keras.io/api/applications/vgg/

[8] GloVe:urlhttp://nlp.stanford.edu/data/wordvecs/glove.840B.300d.zip

[9] It should be noted that for all our multi-modal experiments, we performed early fusion at the feature level. We also performed experiments with late fusion mechanism, but the results were not satisfactory.

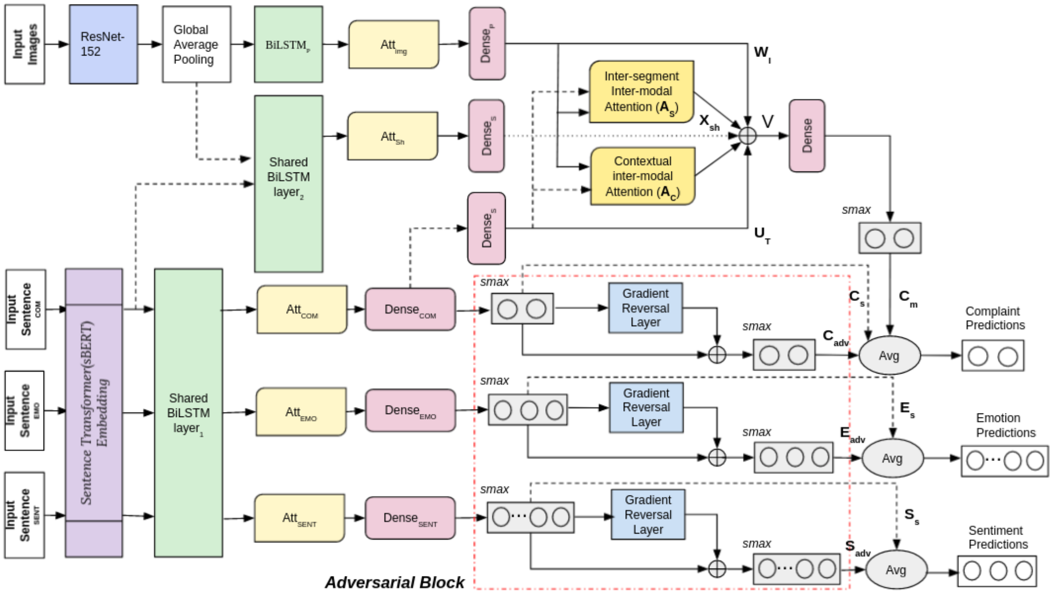


Figure 4: The Multi-modal Complaint Identification (MCI) framework. *smax*: softmax activation function.

- **BERT Multi-modal Multi-task Learning (BMML):** This model is similar to the shared private model architecture like $\mathrm{BSPMF_T}$, additionally it also comprises of two attention mechanisms ($\mathrm{A_S}$, $\mathrm{A_C}$) for the complaint task. For the extraction of image features pre-trained ResNet-152 along with the global average pool layer has been used.
- **Ablation models:** To understand the impact of emotion and sentiment prediction individually on complaint classification, we build dual-task variants of our proposed framework ($\mathrm{MCI_{CE}}$, $\mathrm{MCI_{CS}}$). 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 ($\mathrm{A_S}$, $\mathrm{A_C}$) in the MCI framework ($\mathrm{MCI_{A_S}}$, $\mathrm{MCI_{A_C}}$).

Experimental Setup

We utilize Python-based libraries Tensorflow[10] and Scikit-learn[11] (Pedregosa et al. 2011) to implement our proposed framework and all baselines. We report the weighted average F1 score and the accuracy of the models. 70% of the CESAMARD dataset was used as training data, 10% for validation and the rest 20% was used as testing data on all the experimental models. To ensure a fair comparison of the models, a seed value of 42 was chosen, which allowed the models to encounter the same training and testing data. After each of the BiLSTM layers (256 units), we apply a $dropout$ (Srivastava et al. 2014) of 30% each. The output of the task-specific dense layers (128 units) are fed to the softmax layers (smax) for shared outputs. Dense layers of 100 units each are used before sending the text and image modality encodings through the $\mathrm{A_S}$ and $\mathrm{A_C}$ layers. We also employ a $dropout$ of 30% following the fully connected layer to decrease the risk of overfitting. For the inter-segment inter-modal attention, the number of segments (k) is set to 10. In the dense layers, we employ $ReLU$ activation (Glorot, Bordes, and Bengio 2011). $Softmax$ activation with 2, 6, and 3 neurons are used for the output layers for complaint, emotion, and sentiment classification tasks, respectively. Categorical cross-entropy is used as the loss function to train across all the channels. The epoch size is set to 100. $Adam$ (Kingma and Ba 2015) with a learning rate of 0.001 is used as the optimizer. The above values are chosen after thorough hyperparameter tuning using the RandomSearch tuner of the Keras tuner API.

Results and Discussion

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.

Table 2 depicts the classification results from the various experiments. As can be observed, incorporating multi-modal cues in the form of text and images significantly enhances the performance of single modality baselines. 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/review analysis tasks. As seen in Table 2, the proposed approach, which includes all three tasks (CI, ER, and SA), outperforms single-task variants. $\mathrm{MCI_{CE}}$ outperforms $\mathrm{MCI_{CS}}$ in the dual-task variants. This can be driven by the fact that sentiment alone is often insufficient to convey entire information about the customer's mental state.

[10] https://www.tensorflow.org/

[11] https://scikit-learn.org/stable/

Model	**CESAMARD Dataset**			
	Text		**Text+Image**	
	F1	**A**	**F1**	**A**
SOTA	83.18	83.39	-	-
Single-task Baselines				
STL$_{T}$	83.05	84.59	-	-
STL$_{T+I}$	-	-	85.14	85.26
Multi-task Baselines				
MTL$_{T}$	83.09	84.22	-	-
MTL$_{T+I}$	-	-	84.25	85.53
BSPMF$_{T}$	87.04	87.91	-	-
Multi-modal Baselines				
BMML	-	-	88.38	88.50
MCI$_{CE}$	-	-	87.18	86.85
MCI$_{CS}$	-	-	86.20	85.93
MCI$_{A_S}$	-	-	85.13	86.35
MCI$_{A_C}$	-	-	86.25	87.32
Proposed approach				
MCI	-	-	**89.07***	**89.64***

Table 2: Results of all the baselines and the proposed MCI model in terms of weighted average 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.

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.

Significance of Adversarial Multi-task Architecture In terms of all the multi-task baselines (MTL$_{T}$, MTL$_{T+I)}$, BSPMF$_{T}$, BMML) these approaches do not take into account the adversarial loss. Whereas, the proposed model, MCI, incorporates the adversarial loss, which enhances the performance of the multi-task model. We also illustrate the significance of different attention mechanisms for the proposed MCI framework by conducting ablation studies (MCI$_{A_S}$, MCI$_{A_C}$). Moreover, we also report the results by replacing the SBERT embedding model with Glove embeddings (Pennington, Socher, and Manning 2014) (MTL$_{T}$). The results suggest that each of these factors considerably boosted the performance of the proposed MCI framework. *All of the results presented here are statistically significant*[12] *(Welch 1947).*

Comparison with State-of-the-art Technique (SOTA): We also compare our proposed approach with the existing state-of-the-art technique (Jin and Aletras 2020) for single-task CI as we are unaware of any other multi-modal complaint identification framework. SOTA utilizes an array of neural language models boosted by the use of transformer networks. We re-implement it on the CESAMARD dataset and report the results in Table 2. The proposed model outperforms the SOTA technique.

Error Analysis

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

- Skewness of Dataset: The CESAMARD dataset's imbalanced class distribution influences the proposed MCI model's predictions. The complaint class (33%) is underrepresented 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.
- 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 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 CESAMARD dataset are lengthy and heterogeneous in nature, including diverse emotions in a single syllable. 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 really 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.

Conclusion and Future Work

In this work, we propose a dual attention-based multi-modal (text and images), adversarial multi-task framework for simultaneous optimization of complaint classification, emotion, and sentiment analysis. We have created a novel dataset CESAMARD, that contains reviews collected from the website of the retail giant Amazon and annotated with the complaint labels as well as the associated emotion, and sentiment categories. The dual attention mechanism employs both inter-segment and contextual inter-modal attention. Inter-segment inter-modal attention makes use of the relationship between distinct sentence segments across modalities. Whereas, contextual inter-modal attention learns the contextual information for sentence-level complaint prediction across the modalities. Based on experimental results we can conclude that multi-modality and multi-tasking boost the performance of complaint identification in comparison to its uni-modal and single-task alternatives.

In the future, we aim to extend this work with audio, video of the reviews shared on online shopping platforms for complaint identification.

[12] 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).

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

Akhtar, M. S.; Chauhan, D. S.; Ghosal, D.; Poria, S.; Ekbal, A.; and Bhattacharyya, P. 2019. Multi-task Learning for Multi-modal Emotion Recognition and Sentiment Analysis. In *Proceedings of NAACL-HLT*, 370–379.

Akhtar, M. S.; Ekbal, A.; and Cambria, E. 2020. How intense are you? predicting intensities of emotions and sentiments using stacked ensemble. *IEEE Computational Intelligence Magazine*, 15(1): 64–75.

Bahdanau, D.; Cho, K.; and Bengio, Y. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. In Bengio, Y.; and LeCun, Y., eds., *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Chauhan, D. S.; Dhanush, S.; Ekbal, A.; and Bhattacharyya, P. 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*, 4351–4360.

Cohen, A. D.; and Olshtain, E. 1993. The production of speech acts by EFL learners. *Tesol Quarterly*, 27(1): 33–56.

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

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

Devlin, J.; Chang, M.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Burstein, J.; Doran, C.; and Solorio, T., eds., *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.

Do, H. H.; Prasad, P.; Maag, A.; and Alsadoon, A. 2019. Deep learning for aspect-based sentiment analysis: a comparative review. *Expert Systems with Applications*, 118: 272–299.

Ekman, P.; Friesen, W. V.; O'sullivan, M.; Chan, A.; Diacoyanni-Tarlatzis, I.; Heider, K.; Krause, R.; LeCompte, W. A.; Pitcairn, T.; Ricci-Bitti, P. E.; et al. 1987. Universals and cultural differences in the judgments of facial expressions of emotion. *Journal of personality and social psychology*, 53(4): 712.

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

Ganin, Y.; and Lempitsky, V. 2015. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, 1180–1189. PMLR.

Ghosal, D.; Akhtar, M. S.; Chauhan, D.; Poria, S.; Ekbal, A.; and Bhattacharyya, P. 2018. Contextual inter-modal attention for multi-modal sentiment analysis. In *proceedings of the 2018 conference on empirical methods in natural language processing*, 3454–3466.

Ghosh, S.; Ekbal, A.; and Bhattacharyya, P. 2021. A Multitask Framework to Detect Depression, Sentiment and Multi-label Emotion from Suicide Notes. *Cognitive Computation*, 1–20.

Glorot, X.; Bordes, A.; and Bengio, Y. 2011. Deep sparse rectifier neural networks. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, 315–323.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

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

Jin, M.; and Aletras, N. 2020. Complaint Identification in Social Media with Transformer Networks. In Scott, D.; Bel, N.; and Zong, C., eds., *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, 1765–1771. International Committee on Computational Linguistics.

Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. In Bengio, Y.; and LeCun, Y., eds., *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Kumar, A.; Ekbal, A.; Kawahra, D.; and Kurohashi, S. 2019. Emotion helps Sentiment: A Multi-task Model for Sentiment and Emotion Analysis. In *2019 International Joint Conference on Neural Networks (IJCNN)*, 1–8. IEEE.

Lailiyah, M.; Sumpeno, S.; and Purnama, I. E. 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)*, 307–312. IEEE.

Lewis, M.; Haviland-Jones, J. M.; and Barrett, L. F. 2010. *Handbook of emotions*. Guilford Press.

Liu, P.; Qiu, X.; and Huang, X.-J. 2017. Adversarial Multi-task Learning for Text Classification. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1–10.

Majumder, N.; Poria, S.; Hazarika, D.; Mihalcea, R.; Gelbukh, A.; and Cambria, E. 2019. Dialoguernn: An attentive rnn for emotion detection in conversations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 6818–6825.

Pawar, S.; Ramrakhiyani, N.; Palshikar, G. K.; and Hingmire, S. 2015. Deciphering review comments: Identifying suggestions, appreciations and complaints. In *International Conference on Applications of Natural Language to Information Systems*, 204–211. Springer.

Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. 2011. Scikit-learn: Machine learning in Python. *Journal of machine learning research*, 12(Oct): 2825–2830.

Pennington, J.; Socher, R.; and Manning, C. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 1532–1543.

Poria, S.; Hazarika, D.; Majumder, N.; Naik, G.; Cambria, E.; and Mihalcea, R. 2018. Meld: A multimodal multi-party dataset for emotion recognition in conversations. *arXiv preprint arXiv:1810.02508*.

Pranesh, R. R.; and Shekhar, A. 2020. MemeSem: A Multi-modal Framework for Sentimental Analysis of Meme via Transfer Learning.

Preotiuc-Pietro, D.; Gaman, M.; and Aletras, N. 2019. Automatically Identifying Complaints in Social Media. In Korhonen, A.; Traum, D. R.; and Màrquez, L., eds., *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.

Qureshi, S. A.; Saha, S.; Hasanuzzaman, M.; and Dias, G. 2019. Multitask Representation Learning for Multimodal Estimation of Depression Level. *IEEE Intelligent Systems*, 34(5): 45–52.

Reimers, N.; and Gurevych, I. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *CoRR*, abs/1908.10084.

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 Jangra, A. 2021. Identifying complaints based on semi-supervised mincuts. *Expert Systems with Applications*, 115668.

Singh, R. P.; Haque, R.; Hasanuzzaman, M.; and Way, A. 2020. Identifying Complaints from Product Reviews: A Case Study on Hindi. In Longo, L.; Rizzo, L.; Hunter, E.; and Pakrashi, A., eds., *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*, 217–228. CEUR-WS.org.

Soleymani, M.; Garcia, D.; Jou, B.; Schuller, B.; Chang, S.-F.; and Pantic, M. 2017. A survey of multimodal sentiment analysis. *Image and Vision Computing*, 65: 3–14.

Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; and Salakhutdinov, R. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1): 1929–1958.

Tjandra, S.; Warsito, A. A. P.; and Sugiono, J. P. 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)*, 1–4. IEEE.

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

Yang, W.; Tan, L.; Lu, C.; Cui, A.; Li, H.; Chen, X.; Xiong, K.; Wang, M.; Li, M.; Pei, J.; et al. 2019. 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)*, 56–63.