Reimagining Complaint Analysis: Adopting Seq2Path for a Generative Text-to-Text Framework

Apoorva Singh[1*], **Raghav Jain**[1*] and **Sriparna Saha**[1]
[1]Department of Computer Science and Engineering, Indian Institute of Technology Patna, India
{apoorva_1921cs19,sriparna}@iitp.ac.in, {raghavjain106}@gmail.com *

Abstract

The escalating volume and frequency of social media complaints necessitate robust automated complaint analysis techniques. Much of the existing body of research in this area has been devoted to two primary aspects: identifying complaint-specific content amidst other non-complaint communications, and predicting the severity of a complaint, which involves classifying complaints into different severity levels based on the anticipated resolution from the complainant's perspective. These automated analysis tools equip companies with the means to effectively manage complaints and generate suitable responses. In our study, we present a unified generative approach for complaint detection, transforming the multi-task learning problem into a text-to-text generation task. As part of our training strategy, we adopt the Seq2Path training paradigm that conceptualizes the outcome as a tree structure as opposed to a traditional sequence. This innovative approach tackles the drawbacks of conventional sequences, such as the lack of order among the outputs, yielding a more coherent and structured output. Our model's effectiveness is assessed against the benchmark *Complaints* dataset, highlighting its superior performance across diverse evaluation metrics when compared with state-of-the-art models and other baselines[1].

1 Introduction

Automated complaint analysis benefits companies by efficiently handling large volumes of data, ensuring objectivity, reducing human biases and errors, and streamlining the process of identifying and classifying complaints. It helps to bridge the gap between customer expectations and reality, serving as a crucial feedback tool to pinpoint areas of dissatisfaction and highlight avenues for improvement. Notably, Trosborg et al. (Trosborg, 2011) proposed a four-tiered framework for understanding the severity of complaints, categorizing them from subtle disapproval to outright blame. With the surge in digital platforms, customer complaints have significantly increased, presenting a daunting task for manual processing. This necessitates a more efficient complaint detection approach, as exhibited in recent studies (Preotiuc-Pietro et al., 2019; Singh et al., 2022a) that have successfully automated the classification of binary complaints and their associated severity levels. Emotion recognition (ER) and sentiment recognition (SR) play vital roles in understanding the affective aspects of customer complaints (Singh and Saha, 2021; Singh et al., 2023a).

In this transition towards automation, inspiration can be drawn from multitask learning (Caruana, 1997)—an approach that mirrors our inherent human ability to learn multiple tasks simultaneously and transfer knowledge across them. This technique, while advantageous, poses its own set of challenges, such as negative transfer (where multiple tasks, rather than benefiting the learning pro-

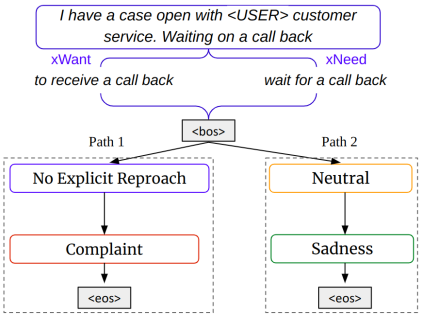

Figure 1: Example of unified generative approach with commonsense knowledge and Seq2Path training used to identify complaints and associated tasks. The two paths represent the related tasks being solved together as a tree-like structure. <bos>: beginning of string, <eos>: end of token.

* The first two authors contributed equally to this work and are jointly the first authors.

[1]The resources are available at https://github.com/Raghav106j/AACL_seq2path_complaint

Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers), pages 241–251
November 1–4, 2023.

cess, begin to hinder the training process) (Crawshaw, 2020) and optimization scheme (assigning weights to different tasks during training) (Wu, 2020). However, a paradigm shift is observed in the natural language processing (NLP) community, where an increasing number of tasks are being formulated as text-to-text generation problems (Du et al., 2021). This approach, harnessing the power of large pre-trained language models and offering unified solutions via a single model, addresses the sheer scale of digital customer complaints. However, it isn't devoid of challenges, such as a) Sequencing, the sequence or order between the outputs does not exist in reality, b) Conditioning, the generation of output should not rely or depend on previously generated outputs.

In this light, the recent success of models transforming text-to-text generation into text-to-tree structures offers an intriguing solution (Mao et al., 2022; Yu et al., 2022; Bao et al., 2022). These models, addressing the challenges of sequencing and conditioning, present outputs as paths of a tree, following a '1-to-n' relationship. In stark contrast to the '1-to-1' relationships depicted by sequence-to-sequence techniques, these models eliminate the need for order or dependence on previously generated outputs. Moreover, our innate capacity to infer implied meanings from explicit expressions—a function of our commonsense—could provide vital insights for automated complaint detection systems. The integration of such external or commonsense knowledge (Sabour et al., 2022) could potentially enrich our understanding of user contexts and situations, contributing to the development of more effective models. The relevance of commonsense reasoning, largely explored in conversational agents and summarization tasks, extends promisingly to complaint detection in this work.

With a view to mitigate the challenges presented by multitask learning and Seq2Seq models, and inspired by the promising implications of the text-to-tree training paradigm, we introduce a novel approach in this paper. We present a commonsense-aware unified generative framework that employs a sequence-to-path (Seq2Path) training paradigm. This framework aims to streamline the primary tasks of complaint detection and severity level classification. Our proposed approach brings a paradigm shift in this regard, offering a more flexible and context-aware mechanism for output generation, thus enhancing the efficacy of automated complaint analysis. Figure 1 shows a user sharing case details with customer service. The user is awaiting a callback (xNeed) and desires to be contacted regarding the case (xWant). Our Seq2Path training generates multiple outputs in a tree-like structure, solving emotion and sentiment in one path, and severity classification and complaint identification in another, as depicted in Figure 1.

Contributions: Our work's significant contributions are as follows:

1) We propose a unified generative approach for complaint detection to concurrently address four tasks: complaint identification (CI), severity classification (SC), with emotion recognition (ER), and sentiment recognition (SR) as auxiliary tasks.

2) We introduce a Seq2Path training method that views the output as a tree rather than a sequence, enabling each associated task to be treated as an individual tree path

3) Our proposed model, evaluated on a benchmark *Complaints* dataset, demonstrates superior performance across various metrics, surpassing other baselines and state-of-the-art models.

2 Related Works

Automatic complaint detection has received a lot of attention in recent years. Earlier studies concentrated on single-task complaint detection, using feature-based machine learning (Preotiuc-Pietro et al., 2019; Coussement and Van den Poel, 2008) and leveraging transformer networks (Jin and Aletras, 2020, 2021a; Singh et al., 2023b, 2021). Apart from complaint mining, research has concentrated on detecting product hazards and risks (Bhat and Culotta, 2017), as well as the likelihood for escalation of complaints (Yang et al., 2019). Recently, multitask complaint detection models (Singh and Saha, 2021; Singh et al., 2022a) that incorporated sentiment and emotion information to improve the complaint mining process were developed.

Developing effective systems for downstream tasks such as chatbots and customer support systems requires a cognitive understanding of the user's conditions and emotions (Sabour et al., 2021). Consequently, we believe that permitting complaint detection models utilize commonsense information and draw conclusions based on what the customer has openly shared is especially beneficial for explaining the user's circumstances, leading to more efficient and socially conscious customer support systems.

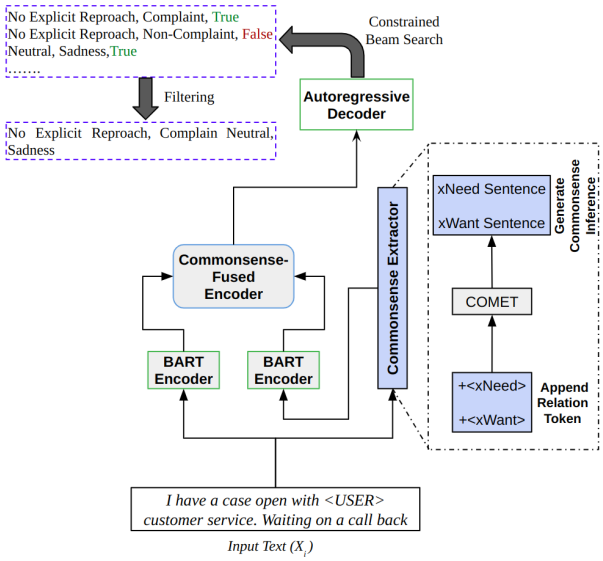

Figure 2: Example of unified generative approach with commonsense knowledge and Seq2Path training used to identify complaints and associated tasks. The two paths represent the related tasks being solved together as a tree-like structure.

Recent advances in deep learning and pre-trained language models have had a significant impact in the area of neural text generation (Raffel et al., 2020a; Lewis et al., 2019). Encoder-decoder Transformers consisting of BART (Lewis et al., 2020) and T5 (Raffel et al., 2020b), have shown massive improvements and success in many NLP tasks such as summarization and translation. Recently, Yan et al. (Yan et al., 2021) solved the task of aspect-based sentiment analysis in a generative manner using the BART model. The BART model is implemented to generate the target sequence in an end-to-end process based on unified task generation.
Past studies on complaint detection leveraged pre-trained language models that were fine-tuned for certain tasks by layering task-specific layers on top of the model. Our current work contrasts in that we redefine the multitask problem as a language generation task, enabling the model to learn to perform the tasks without the requirement for task-specific layers to be trained. A comprehensive literature study has led to the conclusion that Seq2Path, a tree-based generative method, is preferable for complaint identification.

3 Proposed Methodology

We define our problem before getting into the specifics of the proposed model. Figure 2 depicts the overall architecture of our proposed model *CGenPath*.

3.1 Problem Definition

In our study, we aim to explore product reviews through four interconnected tasks: two primary tasks - complaint identification and severity classification; and two auxiliary tasks - sentiment polarity and emotion recognition. Each review, represented as $X_i = \{x_0, x_1, .., x_i, .., x_n\}$, with n denoting the instance length, is classified as follows:
(1) **Complaint Identification** (c): The primary task where a review is assigned a class c from the complaint class set C.
(2) **Severity Classification** (s): Another primary task where the identified complaint is assigned a severity level s from the set S.
(3) **Polarity Classification** (p): An auxiliary task that assigns a sentiment polarity p from the set P to the review.
(4) **Emotion Recognition** (e): An auxiliary task that assigns an emotion e from the emotion class set E to the review.

3.2 Complaint Detection as Seq2Path Task

Here, we conceptualized Complaint detection in multitask setting as a sequence-to-path (Seq2Path) problem (Mao et al., 2022) in which each path can be interpreted as a branch of a tree and can be created separately. At first, we train our commonsense aware seq2seq model (defined in Sec 3.3), where each path is seen as a distinct target, and estimate the average loss. Second, the token generation process is modeled as a tree, and a constrained beam search is used to create paths independently. Finally, given a text as input, the output is a set of all legitimate individual paths with a discriminating token (true or false) attached at the end to pick the correct paths automatically.
Formally, given an input review X_i, our task of obtaining compliant label (c), severity class (s), emotion class (e), and polarity class (p) can be modeled as generating a tree with two paths: (1) One path will generate c and s labels, and (2) second path will generate e and p labels. Finally, the target sequence Y_i is represented as following:

$$Y_i = \{< c >< s >, < e >< p >\} \quad (1)$$

To distinguish valid paths from the others, we added a distinguishing token "true" to the valid paths and "false" to the invalid ones, thereby allowing us to clearly separate them. In order to achieve

better performance on the model, data augmentation is necessary due to the absence of negative samples for discriminative token. To facilitate a good selection of viable paths, each negative sample is augmented with a discriminative token of "false". We generate negative samples by randomly substituting the components of a path. This substitution process helps to boost the model's capacity to recognize valid paths during inference, thus aiding the overall performance of the model.

3.3 Commonsense aware Generative Seq2Path Model for Complaint Detection (*CGenPath*)

We introduce the Commonsense aware Generative Seq2Path Model (CGenPath), a unified generative framework designed to tackle the challenge of complaint detection within a multitask context. Our approach is divided into three main steps for simplicity and clarity: (1) Commonsense Extractor, (2) Commonsense-Fused Encoder, and (3) Seq2Path Training and Inference.

3.3.1 Commonsense Extractor

Our approach begins with the implementation of the Commonsense Extractor, a key component in providing an additional layer of context and commonsense reasoning to typically brief and concise customer reviews. The ATOMIC dataset (Sap et al., 2019) serves as our knowledge base, offering insights into six commonsense relations associated with the entity participating in an event, such as the event's effects (xEffect), the entity's needs (xNeed), and desires (xWant). In the context of complaint detection task, we interpret the review instance as the event and aim to comprehend the customer's needs and desires from their review. Thus, we exclusively focus on two commonsense relations: xNeed and xWant[2]. To generate commonsense reasoning from the customer reviews, we utilize the pre-trained BART (Lewis et al., 2019) based language model, COMET (Hwang et al., 2021), which has been fine-tuned on the ATOMIC dataset. This model is especially effective in providing commonsense reasoning for unseen events (Sabour et al., 2022).

The functioning of the Commonsense Extractor involves appending the commonsense relation tokens (xNeed and xWant) to each review X_i. These concatenated inputs are processed through the pre-trained COMET model to generate two commonsense reasonings $cs^{r_{need}}$ and $cs^{r_{want}}$ for the xNeed and xWant relation tokens, respectively. We generate the final commonsense reasoning CS for each review X_i by concatenating these two reasonings, represented by the following equation: $CS = cs^{r_{need}} \oplus cs^{r_{want}}$.

3.3.2 Commonsense-Fused Encoder

To effectively utilize the commonsense reasoning CS derived from the Commonsense Extractor, we propose a commonsense-aware encoder-decoder framework. This architecture is designed to seamlessly incorporate CS within its sequence-to-sequence learning process, as explained below: In our architecture, the initial stage involves feeding both the review input X_i and the commonsense reasoning CS to a pre-trained BART encoder. This results in two encoded representations, namely U_x and U_{cs}. To integrate the information carried by these two representations, we propose a novel commonsense-fused encoder, an enhancement of the conventional transformer encoder (Vaswani et al., 2017). This involves the generation of two triplets of query, key, and value matrices corresponding to U_x and U_{cs}: (Q_x,K_x,V_x) and (Q_{cs},K_{cs},V_{cs}). Diverging from the standard transformer encoder which projects identical inputs as query, key, and value, our model introduces a cross-attention layer. This layer, consisting of two sublayers of multi-head-cross attention and a normalization layer, exchanges keys and values. It treats (Q_x,K_{cs},V_{cs}) and (Q_{cs},K_x,V_x) as inputs to the cross-attention layer which is computed as defined below, resulting in a cross-infused vector representation.

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \quad (2)$$

where (Q,K,V) represents the set of the query, key, and value and d_k represents the dimension of the query and key.

The cross-attention layer enables a two-way flow of information between U_x and U_{cs}. As a result, the outputs of the multihead cross-attention layer, namely $U_{x->cs}$ and $U_{cs->x}$, are enriched with information from each other. In the next step, we merge $U_{x->cs}$ and $U_{cs->x}$ into a single output, U_z. This merged output is then processed via a self-attention layer, normalization layers, and fully connected layers with residual connections, thereby producing the final output of our commonsense-fused encoder. At last, we bring

[2]We performed a thorough comparative analysis of all the commonsense relations available in the ATOMIC dataset.

together the original text representation U_x, the commonsense reasoning U_{cs}, and the output from the commonsense-fused encoder, culminating in the final commonsense-fused input representation vector, Z. Once we obtain the commonsense-aware input representation vector, denoted as Z, we proceed to feed Z along with all the output tokens until time step $t-1$, represented as $Y_{<t}$, into the decoder module. This process allows us to obtain the hidden state at time step t, which can be defined as follows: $H^t_{DE} = G_{Decoder}(Z, Y_{<t})$ where $G_{Decoder}$ denotes the decoder computations.

The conditional probability of predicting the output token at the t-th time step, given the input and the previous $t-1$ tokens, is determined by applying the softmax function to the hidden state H^t_{DEC} as follows:

$$P_\theta(Y_t^{'}|R, Y_{<t}) = F_{softmax}(\theta^T H^t_{DE}) \quad (3)$$

where $F_{softmax}$ represents softmax computation and θ denotes weights of our model.

3.3.3 Seq2Path Training and Inference

In this section, we discuss training and inference of our commonsense aware model specific to the Seq2Path paradigm.

Training: For a given review text X_i, we want to generate a set of different paths. Our dataset consists of pairs of (X_i, Y_i) where $Y_i = \{y^1, y^2, ..., y^k\}$. As depicted in Figure 1, Y_i can be expressed as a tree, and each y has a distinct path in the tree. The total number of paths is represented by k. When predicting $Y_i^{'}$ from X_i, the loss can be defined as the average loss of the k paths motivated by Mao et al. (Mao et al., 2022).

$$L(Y^{'}, Y_i|X_i) = \frac{\sum_{y \in Y_i} L_{MLE}(y^{'}, y|X_i)}{k} \quad (4)$$

where L_{MLE} is the standard loss function based on the maximum likelihood estimation (MLE) objective function. In our case, as defined in equation 1, we have 2 paths setting the total number of paths parameter k as 2.

Inference: At the inference stage, we use beam search with constrained decoding motivated by different text-to-tree generation works (Mao et al., 2022; Yu et al., 2022; Bao et al., 2022). Beam Search considers multiple alternative options based on the hyperparameter beamwidth (B) and conditional probability which is more optimal than a simple greedy search technique that only selects a single best token at each time step. Utilizing beam search, we output the top-k paths with diminishing probabilities which signify the possibility of the paths being right. During the decoding process, constrained decoding is employed in order to restrict beam search to look only within specific candidate tokens. As suggested by Cao et al. (De Cao et al., 2020), these tokens can be both derived from the given input text and/or some special tokens intended for the specific task at hand[3]. We also applied a filtering process to filter out the invalid paths. We output the valid paths with a discriminative token "true" and remove the other paths with a discriminative token "false". This step ensured that only the valid paths were retained and any invalid ones were excluded.

4 Experiments and Results

This section describes the dataset used, experiments, results, and analysis of our proposed model. The experiments are intended to address the following research questions:

RQ1: How does the generative paradigm perform in comparison to traditional multi-task models?

RQ2: How does Seq2Path paradigm overcome the drawbacks of Seq2Seq and multi-task models?

RQ3: What is the impact of external knowledge and the Seq2Path training paradigm on the performance of our framework?

RQ4: Is the proposed model able to outperform state-of-the-art models for complaint detection and severity classification tasks?

4.1 Dataset Description

In the current study, we use the *Complaints* dataset provided in (Preotiuc-Pietro et al., 2019), which comprises of 3,449 tweet instances in English. We chose this dataset since it is open source and includes annotated complaints from Twitter, a prominent data-analysis platform. Recently, Jin et al. (Jin and Aletras, 2021a) added five severity levels to the *Complaints* dataset (no explicit reproach, disapproval, accusation, blame, and non-complaints). In the work, Singh et al. (Singh et al., 2022a) added sentiment (negative, neutral, positive) and emotion (anger, disgust, fear, happiness, sadness, surprise, and other) classes to this dataset; the 'other' emotion class represents tweets that are not covered by Ekman's six basic emotions (Ekman et al., 1987). For our current study, we use this extended dataset

[3]We used the special tokens method.

Tweet	Complaint	Severity	Polarity	Emotion
<USER> what is your policy on false advertising? I was refused a sale in westfield due to a company error on pricing.	Complaint	Blame	Negative	Sadness
Thank you <USER> for a better H2 styling I will actually use it now!	Non-Complaint	Non-Complaint	Positive	Happiness
<USER> I have just received an email regarding an order I've placed. I don't recall placing any order with you.plz advise?	Complaint	No explicit reproach	Neutral	Other

Table 1: Different example instances of *Complaints* dataset.

annotated with severity levels, polarity, and emotion labels[4].

Statistics related to *Complaints* dataset: The detailed statistics related to the extended *Complaints* dataset are as follows:

(1) The original work by Preotiuc-Pietro et al. (Preotiuc-Pietro et al., 2019) consists of 1,235 complaints and 2,214 non-compliant tweets in English.

(2) The distribution of tweets across the severity classes (SD task) as mentioned in the work by Jian et al. (Jin and Aletras, 2021a) is as follows: 435 tweets belong to 'No Explicit Reproach', 378 belong to 'Disapproval', 225 belong to 'Accusation', and 197 belong to 'Blame'.

(3) Singh et al.'s (Singh et al., 2022a) study found that 844 tweets were categorized as 'Anger', 7 as 'Disgust', 8 as 'Fear', 473 as 'Joy', 1,479 as 'Other', 626 as 'Sadness', and 12 as 'Surprise' in the emotion classification task.

(4) The sentiment classes for the tweets (SA task) (Singh et al., 2022a) were broken down as follows: 1,041 negative, 1,198 neutral, and 1,210 positive.

(5) In (Preotiuc-Pietro et al., 2019), the inter-rater agreement score for the main task CD is reported as 0.73 (Cohen's Kappa). For the SD task, the inter-rater agreement is 0.64 (Fleiss' Kappa score) according to (Jin and Aletras, 2021a). The auxiliary tasks ED and SA have inter-rater agreement scores of 0.68 and 0.82 (Cohen's Kappa score) respectively, as reported in (Singh et al., 2022a). Table 1 presents example instances from the *Complaints* dataset.

4.2 Baseline Models and Experimental Setup

The primary model in our study, *CGenPath*, uses BART (Lewis et al., 2019) as its foundation, and its configuration and training methodology are outlined in **Appendix** A.1. For comparative baselines, we consider several multitask systems and text-to-text generation models. These include Baseline$_1$, which was inspired by previous work from Singh et al. (Singh et al., 2022b), and MT_{GloVe}, which uses a GloVe pre-trained word embedding and a subsequent BiGRU layer. We also studied BERT-MT, which is a multitask model that's based on BERT, as well as BART and T5 models. Besides these, we developed a variant of *CGenPath* called $CGenPath_{con}$ that simply combines the input text and commonsense reasonings. We also made two versions of our main model that leave out certain features: *CGenPath*$-$Seq2Path, which doesn't use the Seq2Path training mechanism, and *CGenPath*$-$CS, which leaves out commonsense reasoning. Full details about these models' development and training are available in **Appendix** A.2.

4.3 Results and Discussions

It is crucial to emphasize that the primary focus of this study is to improve the performance of the CI (Complaint Identification) and SC (Severity Classification) tasks. Consequently, the results and analysis presented in this research solely consider CI and SC as the primary tasks.

(RQ1) The results presented in Table 2 demonstrate the superior performance of *CGenPath* over all multitask baselines in both the Complaint Identification (CI) and Severity Classification (SC) tasks. *CGenPath* surpasses Baseline1, BERT-MT, and MT$Glove$ by margins of 9.4%, 6.1%, and 7.9%, respectively, in terms of macro-F1 scores for the CI task. A similar trend is observed for the SC task. Notably, other generative methods such as BART, T5, and $CGenPath_{con}$ also exhibit significant performance advantages over the multitask baselines in both tasks. These findings highlight the superiority of pre-trained sequence-to-sequence language models in the context of our experiments.

(RQ2) As can be seen from the table 2 *CGenPath* clearly outperforms other generative baselines and its variant *CGenPath*$_{con}$ across both tasks. *CGenPath* outperforms BART, T5 and *CGenPath*$_{con}$ by a margin of: (1) 1.9%, 4.1%, and 1.1% in CI task on macro-F1 score and (2) 11.3%, 5%, and 3.6% in SC task on macro-F1 score. This performance

[4] https://www.iitp.ac.in/~ai-nlp-ml/resources.html

	Complaint Identification (CI)		Severity Classification (SC)	
Model	F1	A	F1	A
SOTA(Jin and Aletras, 2021b)	86.6	87.6	59.4	55.5
CGenPath	**90.8**†	**91.3**†	**73.7**†	**70.4**†
Baseline$_1$(Singh et al., 2022b)	81.4	82.8	60.3	62.8
BART	88.9	88.9	62.4	62.6
T5	86.7	86.6	68.7	69.3
$CGenPath_{con}$	89.7	89.9	70.1	68.4
MT_{GloVe}	82.9	84.3	52.4	50.1
BERT-MT	84.7	86.1	57.7	53.3

Table 2: Results of different baselines, SOTA and the proposed framework,*CGenPath*. For the CI and SC tasks, the results are in terms of macro-F1 score (F1) and Accuracy (A) values. F1, A metrics are given in %. The maximum scores attained are represented by bold-faced values. The † denotes statistically significant findings.

	Complaint (CI)		Severity (SC)	
Model	F1	A	F1	A
CGenPath	**90.8**†	**91.3**†	**73.7**	**70.4**†
-Seq2Path	89.1	88.9	64.5	65.1
-CS	89.8	90.3	71.4	69.3

Table 3: Results of the ablation studies performed on the proposed framework's key components in terms of macro-F1 score (F1) and Accuracy (A) values. The maximum scores attained are represented by bold-faced values. The † denotes statistically significant findings.

gain over BART and T5 can be attributed to the following: (1) Trees can be viewed as a better semantic representation to more effectively capture the structure of different tasks, and (2) Seq2Path training is able to make the model understand the order of output thus making results more consistent. These findings validate the effectiveness of reframing the multitasking problem as a Seq2Path generation task. Furthermore, the performance improvement over $CGenPath_{con}$ highlights the necessity for a well-designed and appropriately formulated fusion technique to effectively integrate diverse sources of information.

(RQ3) *Ablation Study:* To investigate the impact of commonsense reasoning and Seq2Path training on our model's performance, we conducted an ablation study (Table 3). The results clearly indicate that removing the Seq2Path training mechanism leads to a noticeable decline in performance for both the CI (Complaint Identification) and SC (Severity Classification) tasks, as reflected in the evaluation metrics. Furthermore, when the commonsense reasoning (CS) component is removed from our model, we observed a reduction in performance of 1% and 2.3% for the CI and SC tasks, respectively. However, when the CS component and Seq2Path are combined in the CGenPath model, it surpasses all ablated models and baseline models in terms of performance across all evaluation metrics for each subtask. This outcome highlights the significant contribution of each component in improving the model's performance.

(RQ4) *Comparison with State-of-the-art Technique (SOTA)*: Our proposed model (*CGenPath*) is able to outperform the SOTA model (Jin and Aletras, 2021b) on CI and SC tasks. *CGenPath* outperforms the SOTA by a significant margin of 4.2% and 14% on CI and SC tasks in macro-F1 score, respectively. The reasons for these improvements can also be attributed to the facts: 1) Our model, *CGenPath* is leveraging pretrained BART model's knowledge which already has been trained on a huge corpus of data, 2) This model has extra context in the form of commonsense reasoning due to which they are making better predictions, and 3) Seq2Path training can handle complex relations between different tasks and produce more consistent results.

To assess the statistical significance of the obtained results, we conducted a paired T-test. The analysis revealed that the performance improvement achieved by our proposed model compared to the state-of-the-art is statistically significant with a 95% confidence level (i.e., p-value < 0.05).

4.4 Qualitative Analysis

During our qualitative analysis, we found tweets displaying evident complaint markers, such as blame-associated language or accusations, are less prone to misclassification. A comparison of our system's results with those of the state-of-the-art (SOTA) system is presented in Table 4. This comparison clearly shows that incorporating common sense reasoning and the Seq2Path training methodology into the Complaint Identification (CI) task enhances the prediction outcomes. Unlike the SOTA system, which lacks these components, our system provides more precise results. Examining Table 4, it's clear that the SOTA model incorrectly categorizes the complaint

Tweet text	SOTA	Proposed	True Label
It might have been a bad overclock. Also, it normal to have flickering in certain applications while running SLI?	complaint blame	complaint disapproval	complaint disapproval
I am getting server errors when I try to activate a device. Plz help	Non-complaint disapproval	complaint no explicit-reproach	complaint no explicit-reproach

Table 4: Qualitative analysis of the SOTA (Jin and Aletras, 2021b) and the proposed model for CI and SC task predictions.

label as Non-complaint in instance 2, while its predicted severity level is Disapproval, leading to a discrepancy between the two labels. On the contrary, our model successfully predicts both labels accurately, thereby demonstrating that *CGenPath* yields more consistent outcomes.

4.5 Error Analysis

We've conducted a detailed examination of a selection of test set samples, comparing the complaint and severity labels produced by *CGenPath* to those annotated by a human. During this evaluation, we've come across a few instances where the model exhibited errors:
1. Misinterpretation of Implicit Complaints: The model struggles with predictions when the true intention is subtly embedded in the text. In cases where complaints are implied rather than explicitly stated, the model inaccurately identifies them as non-complaints due to a literal interpretation of the text. An example of this is the sentence, <*USER*> *You guys are doing an amazing job ensuring that every week there's a new bug in the software.* This is a complaint, but the model mistakenly labels it as a non-complaint. This issue is likely due to the indirect expression of dissatisfaction or blame by the user.
2. Severity Misclassification Due to Linguistic Overlaps: Misclassification can also take place when instances sharing similar linguistic and structural characteristics exist within adjacent severity levels. Consider the sentence *"I am surprised that a reputable company like yours has such a complex return policy. It's quite disappointing."*. In this case, the model incorrectly identifies the severity level as disapproval instead of the correct label - accusation. This misclassification may be attributed to the presence of words such as "disappointed", which are typically associated with the disapproval class.

5 Conclusion

In this study, we aimed to achieve three objectives: firstly, developing a cohesive generative strategy for complaint identification by redefining the multitask learning approach as a text-to-text generation task using the Seq2Path training paradigm; secondly, addressing the shortcomings of previous research that relied on basic multi-task modeling, faced challenges such as negative transfer and optimization strategy; lastly, enhancing model performance by incorporating external knowledge through commonsense reasoning. The Seq2Path technique, which crafts the output as a tree structure, overcomes the inherent drawback of standard sequences, resulting in a more coherent and sequentially ordered output. Experimental tests on the standard *Complaints* dataset demonstrate that our strategy outperforms the current state-of-the-art model and other benchmarks. Future research aims to extend our unified generative strategy to handle multimodal complaints at a granular aspect level and extract valuable insights from product reviews.

Limitations

We attempted to develop a novel unified generative framework for complaint analysis. But the proposed approach is having some limitations as enumerated below:
(1) The proposed methodology has been validated on an English language complaint dataset; further training would be required to scale up to code-mixed language datasets which are prevalent in multilingual countries.
(2) Users often post some images along with text while writing complaints. The current system is unable to handle such multi-modal forms of inputs.

Acknowledgement

Dr. Sriparna Saha would like to acknowledge the support of the Science and Engineering Research Board (SERB)-POWER scheme, India

[SPG/2021/003801-G], a statutory body of the Department of Science & Technology, Government of India) to carry out this research.

References

Xiaoyi Bao, Wang Zhongqing, Xiaotong Jiang, Rong Xiao, and Shoushan Li. 2022. Aspect-based sentiment analysis with opinion tree generation. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 4044–4050. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Shreesh Bhat and Aron Culotta. 2017. Identifying leading indicators of product recalls from online reviews using positive unlabeled learning and domain adaptation. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11, pages 480–483.

Rich Caruana. 1997. Multitask learning. *Machine Learning*, 28.

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.

Michael Crawshaw. 2020. Multi-task learning with deep neural networks: A survey. *arXiv preprint arXiv:2009.09796*.

Nicola De Cao, Gautier Izacard, Sebastian Riedel, and Fabio Petroni. 2020. Autoregressive entity retrieval.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2021. All nlp tasks are generation tasks: A general pretraining framework.

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.

Jena D. Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi. 2021. Comet-atomic 2020: On symbolic and neural commonsense knowledge graphs. In *AAAI*.

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.

Mali Jin and Nikolaos Aletras. 2021a. Modeling the severity of complaints in social media. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 2264–2274. Association for Computational Linguistics.

Mali Jin and Nikolaos Aletras. 2021b. Modeling the severity of complaints in social media. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2264–2274.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. pages 7871–7880.

Yue Mao, Yi Shen, Jingchao Yang, Xiaoying Zhu, and Longjun Cai. 2022. Seq2Path: Generating sentiment tuples as paths of a tree. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2215–2225, Dublin, Ireland. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

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.

Syed Arbaaz Qureshi, Gael Dias, Mohammed Hasanuzzaman, and Sriparna Saha. 2020. Improving depression level estimation by concurrently learning emotion intensity. *IEEE Computational Intelligence Magazine*, 15(3):47–59.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020a. Exploring the limits

of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020b. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Sahand Sabour, Chujie Zheng, and Minlie Huang. 2021. CEM: commonsense-aware empathetic response generation. *CoRR*, abs/2109.05739.

Sahand Sabour, Chujie Zheng, and Minlie Huang. 2022. Cem: Commonsense-aware empathetic response generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11229–11237.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A Smith, and Yejin Choi. 2019. Atomic: An atlas of machine commonsense for if-then reasoning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 3027–3035.

Apoorva Singh, Rohan Bhatia, and Sriparna Saha. 2023a. Complaint and severity identification from online financial content. *IEEE Transactions on Computational Social Systems*.

Apoorva Singh, Siddarth Chandrasekar, Tanmay Sen, and Sriparna Saha. 2023b. Federated multi-task learning for complaint identification using graph attention network. *IEEE Transactions on Artificial Intelligence*.

Apoorva Singh, Arousha Nazir, and Sriparna Saha. 2022a. Adversarial multi-task model for emotion, sentiment, and sarcasm aided complaint detection. In *Advances in Information Retrieval - 44th European Conference on IR Research, ECIR 2022, Stavanger, Norway, April 10-14, 2022, Proceedings, Part I*, volume 13185 of *Lecture Notes in Computer Science*, pages 428–442. Springer.

Apoorva Singh, Arousha Nazir, and Sriparna Saha. 2022b. Adversarial multi-task model for emotion, sentiment, and sarcasm aided complaint detection. In *European Conference on Information Retrieval*, pages 428–442. Springer.

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, Tanmay Sen, Sriparna Saha, and Mohammed Hasanuzzaman. 2021. 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.

Anna Trosborg. 2011. *Interlanguage pragmatics: Requests, complaints, and apologies*, volume 7. Walter de Gruyter.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Sen Wu. 2020. *Automating Knowledge Distillation and Representation from Richly Formatted Data*. Stanford University.

Hang Yan, Junqi Dai, Tuo Ji, Xipeng Qiu, and Zheng Zhang. 2021. A unified generative framework for aspect-based sentiment analysis. *CoRR*, abs/2106.04300.

Wei Yang, Luchen Tan, Chunwei Lu, Anqi Cui, Han Li, Xi Chen, Kun Xiong, Muzi Wang, Ming Li, Jian Pei, and Jimmy Lin. 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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 2 (Industry Papers)*, pages 56–63. Association for Computational Linguistics.

Rongli Yi and Wenxin Hu. 2019. Pre-trained bert-gru model for relation extraction. In *Proceedings of the 2019 8th International Conference on Computing and Pattern Recognition*, pages 453–457.

Chao Yu, Yi Shen, and Yue Mao. 2022. Constrained sequence-to-tree generation for hierarchical text classification. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 1865–1869, New York, NY, USA. Association for Computing Machinery.

A Appendix

A.1 Experimental Setup

In this section, we present the hyperparameters and experimental configurations employed in our study. All experiments were conducted on the Tyrone machine, equipped with Intel's Xeon W-2155 Processor, 196 GB DDR4 RAM, and an Nvidia 1080Ti GPU with 11 GB memory. The dataset was randomly divided, with 80% used for training, 5% for validation, and the remaining 15% for testing. Each model was run ten times, and the average results were reported. For our CGenPath model, we utilized BART (Lewis et al., 2019) as the base model. Training was conducted for a maximum of 60 epochs with a batch size of 16. We employed the Adam optimizer with an epsilon value of

0.00000001. A seed value of 32 was chosen for fair comparisons. The implementation of all models utilized the Scikit-Learn[5], Huggingface library[6], and the PyTorch[7] backend. The predictive performance of our proposed model on all tasks was evaluated using two metrics: accuracy and macro-F1 score. The specifications of the transformer model used are as follows: (1) Number of encoder layers: 6, (2) Number of decoder layers: 6, (3) Dimensionality of layers: 1024, and (4) Embedding size: 1024.

A.2 Baseline Models

Multitask systems: Drawing inspiration from the advancements in the multitask CI framework, we developed Baseline_1 (Singh et al., 2022b) as one of the multitask baselines. This model allows for simultaneous learning of CI (Complaint Identification), SC (Severity Classification), PR (Polarity Classification), and ER (Emotion Recognition), employing the same experimental setup as our current study.

We also developed another baseline model called MT_{GloVe} (Qureshi et al., 2020). This approach leverages pre-trained GloVe word embeddings (Pennington et al., 2014) as its initial step, retrieving embeddings from the GloVe pre-trained word embedding file[8]. The embedding layer's output is then fed into a word sequence encoder, which captures contextual information from the sentence. The MT_{GloVe} model incorporates a fully shared BiGRU layer with 256 units, followed by a shared attention layer. The output of the attention layer is further processed by four task-specific dense layers before being directed to the output layers.

In addition, we developed a Basic Multi-task System (BERT-MT) based on BERT (Yi and Hu, 2019). The architecture of BERT-MT comprises a shared BiGRU layer with 256 units, followed by a shared attention layer. The output of the attention layer is then fed into four task-specific dense layers, each accompanied by its respective output layer.

Text to Text Generation Model: We use BART (Lewis et al., 2019) and T5 (Raffel et al., 2019) as the baseline text-to-text generation models.

a) **BART:** BART is an encoder-decoder-based transformer model which is mainly pre-trained for text generation tasks such as summarization and translation. BART is pre-trained with various denoising pretraining objectives such as token masking, sentence permutation, sentence rotation etc.

b) **T5:** T5 is also an encoder-decoder-based transformer model which aims to solve all the text-to-text generation problems. The main difference between BART and T5 is the pre-training objective. In T5, the transformer is pre-trained with a denoising objective where 15% of the input tokens are randomly masked and the decoder tries to predict all these masked tokens whereas, during pre-training of BART, the decoder generates the complete input sequence.

We fine-tune both these models on the proposed dataset with complaint text as the input sequence and concatenated outputs as the target sequence with Maximum likelihood Estimation as the objective function.

Concatenation based CGenPath: We also proposed a variation of our framework named $CGenPath_{con}$, where we directly concatenate the input review and commonsense reasoning instead of a fusing mechanism.

Ablation Models: The *CGenPath* model comprises two key components: (1) Commonsense Reasoning (CS) and (2) Seq2Path Training. In order to establish the necessity of both of these components individually, we conduct an ablation study of the proposed framework. In this case, we propose two ablated models; (1) *CGenPath*$-$Seq2Path where we replaced the seq2path training with standard seq2seq training, and (2) *CGenPath*$-$CS where we didn't include the commonsense reasoning into our encoder.

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

[6] https://huggingface.co/

[7] https://pytorch.org/

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