Peeking inside the black box: A Commonsense-Aware Generative Framework for Explainable Complaint Detection

Apoorva Singh[1*], **Raghav Jain**[1*], **Prince Jha**[1] and **Sriparna Saha**[1]
[1]Department of Computer Science and Engineering, IIT Patna, India
{apoorva_1921cs19,sriparna}@iitp.ac.in, {raghavjain106, jhapks1999}@gmail.com [*]

Abstract

Complaining is an illocutionary act in which the speaker communicates his/her dissatisfaction with a set of circumstances and holds the hearer (the complainee) answerable, directly or indirectly. Considering breakthroughs in machine learning approaches, the complaint detection task has piqued the interest of the natural language processing (NLP) community. Most of the earlier studies failed to justify their findings, necessitating the adoption of interpretable models that can explain the model's output in real-time. We introduce an explainable complaint dataset, *X-CI*, the first benchmark dataset for explainable complaint detection. Each instance in the *X-CI* dataset is annotated with five labels: complaint label, emotion label, polarity label, complaint severity level, and rationale (explainability), i.e., the causal span explaining the reason for the complaint/non-complaint label. We address the task of explainable complaint detection and propose a commonsense-aware unified generative framework by reframing the multitask problem as a text-to-text generation task. Our framework can predict the complaint cause, severity level, emotion, and polarity of the text in addition to detecting whether it is a complaint or not. We further establish the advantages of our proposed model on various evaluation metrics over the state-of-the-art models and other baselines when applied to the *X-CI* dataset in both full and few-shot settings[1].

1 Introduction

Complaining is an expression of negative emotions communicated due to a discrepancy between reality and expectations (Olshtain and Weinbach, 1985). In pragmatics theory, Trosborg (2011) proposed four primary complaint severity levels: (a) no explicit reproach, (b) disapproval, (c) accusation, and (d) blame. Recent studies on complaint detection have mainly dealt with automatically identifying binary complaints and the associated severity levels from social-media data (Preotiuc-Pietro et al., 2019; Jin and Aletras, 2021). However, these studies primarily focused on improving complaint detection performance with the help of various models without providing any perspective or evaluation of the outcome's explainability. Since the introduction of explainable artificial intelligence (AI) (Gunning et al., 2019), providing interpretation for any AI algorithm's decision has become essential. Therefore, rather than enhancing performance by increasing computational burden, there is a push to construct trustworthy and transparent interpretable models.

As sentient beings, we use our commonsense to establish connections between what is explicitly said and inferred. As a result, we believe that adding external knowledge or commonsense knowledge (Sabour et al., 2021a) to complaint detection systems can help them better grasp the user's circumstances and concerns, resulting in more efficient models. In Figure 1, the user shares information about a registered case with customer service. Based on the given information, we can say the user is waiting for a callback (xNeed) and wants to receive a callback for the registered case (xWant). To the best of our knowledge, all the previous attempts to incorporate commonsense reasoning have only been done for conversational agents and summarization tasks.

Previous research (Saha et al., 2021; Singh et al., 2022) has shown that closely related tasks benefit each other when learned concurrently. However, this strategy entails several problems, such as negative transfer (where multiple tasks, rather than benefiting the learning process, begin to hinder the training process) (Crawshaw, 2020) and optimization scheme (assigning weights to different tasks during training) (Wu, 2020). To address

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

[1]The dataset and code are available at `https://github.com/appy1608/ACL2023`

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 7333–7347
July 9-14, 2023 ©2023 Association for Computational Linguistics

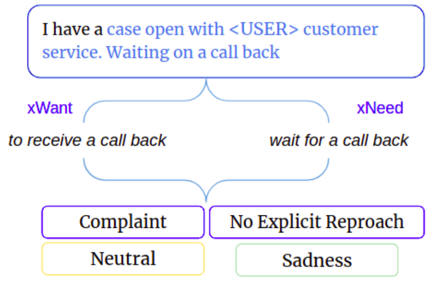


Figure 1: Example from the *X-CI* dataset in which causal span annotation (highlighted in blue) and commonsense knowledge are used to identify complaints and associated tasks.

the aforementioned multitask learning issues, and motivated by generative language models' ability to solve downstream tasks in a generative manner in both full and few-shot (low resource and data constrained) scenarios (Brown et al., 2020a), we propose using text-to-text generation to accomplish the tasks of complaint identification, severity level classification, and cause extraction.

Research Objectives: Following are the research objectives of the current study:

1) We aim to understand the effect of using causal information on the complaint detection, and severity classification tasks in the proposed framework.

2) We intend to do a comparative study of the discriminative approach (multitask learning) and generative approach (text-to-text generation), specifically in the complaint detection domain.

3) Finally, this work aims to study how commonsense knowledge can further boost the performance of the generative approach.

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

1) This is the first study on explainable complaint identification; where we determine the cause for classifying social media data as complaints. We explore two new challenges: (i) explainable complaint identification and (ii) multitask learning as a text-to-text generation problem.

2) We develop *X-CI*, a new benchmark dataset for explainable complaint detection with causal span annotation of expressed complaint/non-complaint.

3) To simultaneously solve five tasks, complaint identification (CI), severity classification (SC), emotion recognition (ER), polarity recognition (PR), and cause extraction (CE), we propose a commonsense-aware unified generative framework, where CI, SC, and CE are the main tasks, and ER, PR are the auxiliary tasks.

4) We established the superiority of our proposed approach on various evaluation metrics over other baselines and state-of-the-art models. The evaluation results further show that the proposed generative model consistently outperforms all other baselines and state-of-the-art models, in full and few-shot settings.

The rest of the paper is organized as follows. Section 2 discusses some of the previous studies on this subject. Following that, in Section 3, we explain the dataset development in detail. In Section 4, we summarize our proposed methodology for the unified generative method-based experiments. We analyze the experiments, results, and their outcomes in Section 5. Finally, in Section 6, we conclude our study and identify the scope of future work.

2 Related Studies

In computational linguistics, previous works on complaint detection only pivoted on identifying complaints using feature-based machine learning models (Preotiuc-Pietro et al., 2019; Coussement and Van den Poel, 2008), transformer network-based models (Jin and Aletras, 2021). Recently multitask complaint analysis models have been developed that leveraged polarity and affect information for enhancing the complaint mining task (Singh and Saha, 2021; Singh et al., 2022, 2021, 2023).

In affective computing, identifying the causal span of expressed emotions is crucial for understanding human emotions (Poria et al., 2021). Motivated by previous research, we aim to investigate the reasons behind viewpoints, particularly complaints, expressed on social media

Providing cognitive awareness of the user's circumstances and sentiments is essential to designing effective systems for downstream tasks such as chatbots (Sabour et al., 2021b). Hence, we believe that allowing complaint detection models to use commonsense information and draw conclusions from what the user has openly shared is particularly useful for better understanding the user's circumstances, resulting in more effective and socially aware customer support systems.

Recent breakthroughs in deep learning and pre-trained language models have substantially affected the development in the field of neural text generation (Raffel et al., 2020; Lewis et al., 2019). Models such as GPT-2 (Radford et al., 2019) and GPT-3 (Brown et al., 2020b) are decoder-only transformers that can generate understandable, and consistent text because of being pre-trained on a vast amount of text data. BART (Lewis et al., 2019) and T5 (Raffel et al., 2020) are encoder-decoder transformers that have shown rapid evolution and success in many NLP applications, such as summarization and translation.

After an in-depth literature review, it can be concluded that there is no work on explainable complaint detection. In this paper, we attempt to bridge this research gap.

3 *X-CI* Dataset Development

For this work, we utilize the *Complaints* dataset[2] published in (Preotiuc-Pietro et al., 2019), which includes 3,449 tweet instances in English. We selected this dataset because it is openly available and comprises annotated complaints from the social-networking site Twitter, a popular choice for data analysis. Jin and Aletras (2021) augmented *Complaints* dataset with five severity levels (*no explicit reproach, disapproval, accusation, blame, and non-complaints*). Recently, Singh et al. (2022) enriched the *Complaints* dataset with the sentiment (*negative, neutral, positive*) and emotion (*anger, disgust, fear, happiness, sadness, surprise and other*) classes; the 'other' emotion class depicts tweets that do not fall under the scope of Ekman's six basic emotions (Ekman et al., 1987). We utilize this extended dataset annotated with severity levels, polarity, and emotion classes for our current work.

3.1 Annotations

Three annotators with expertise in creating supervised corpora were assigned the task of annotating causal spans for each sample in the dataset. The chosen annotators come from varied backgrounds and demographics to ensure the elimination of biases.

Annotators were asked to find the causal span, X(I), that best described the foundation of the complaint (C) or non-complaint (NC) label for each occurrence (I) in the *X-CI* dataset. If there is no mention of X(I) for C/NC in I, the sentence was tagged as 'no cause' by the annotators. It is worth noting that 95% of the instances in the dataset have only either complaint or non-complaint cause, while 5% have no cause. Based on previous research on span extraction (Rajpurkar et al., 2016), we use the macro-F1 measure to assess inter-rater agreement and achieve a 0.77 F1 score, suggesting that the annotations are of good quality. The extended dataset *X-CI* now includes the tweet text, complaint label, severity level, polarity label and emotion label, and annotated cause for each instance. Please refer to Section A in the **Appendix** where we provide the details related to the annotation procedure and also provide detailed statistics related to the *X-CI* dataset.

4 Proposed Methodology

This section defines our problem before delving into the details of the proposed framework. The overall framework is shown in Figure 2.

Problem Definition: An explainable complaint detection model should predict the review's cause, severity level, emotion, and polarity class in addition to detecting whether it is complaint or not. Formally, given an input instance $X_i = \{x_0, x_1, .., x_i, .., x_n\}$ where n is the length of input instance, we intend to learn five closely related tasks: (i) complaint identification (c), (ii) polarity classification (p), (iii) emotion recognition (e), (iv) severity level classification (s), (v) cause extraction (ce), where $c \in C$, C is the set of complaint classes, $p \in P$, P is the set of polarity classes, $e \in E$, E is the set of emotion classes, $s \in S$, S is the set of severity levels and $ce(X_i) \in (X_i)$ that is relevant to the complaint label, c.

4.1 Explainable Complaint Detection as Text-to-Text Generation Task

Here, we propose a text-to-text generation paradigm for solving explainable complaint detection and other auxiliary tasks in a single unified manner. To transform this problem into a text generation problem, we first construct a natural language target sequence, Y_i, for input sentence, X_i, for training purposes by concatenating the labels of all the tasks as defined in the following Equation: $Y_i = \{< ce(X_i) > [c][s][e][p]\}$.

We have added special characters after each task's prediction, so that task-specific predictions can be extracted during inference. Now the problem can be reformulated as: given an input sequence X_i, the task is to generate an output sequence, $Y_i^{'}$, containing all the predictions as defined in the above equation using a generative model G; $Y_i^{'} = G(X_i)$.

[2] https://github.com/danielpreotiuc/complaints-social-media

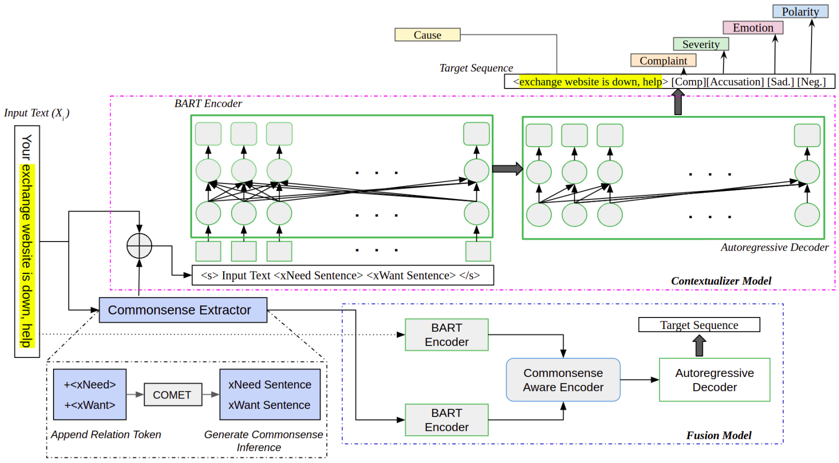


Figure 2: The overall architecture of the proposed model (*C2D*). The two variations of our proposed model, *ConC2D* and *FuseC2D* are depicted by the enclosed red and blue dotted boxes, respectively.

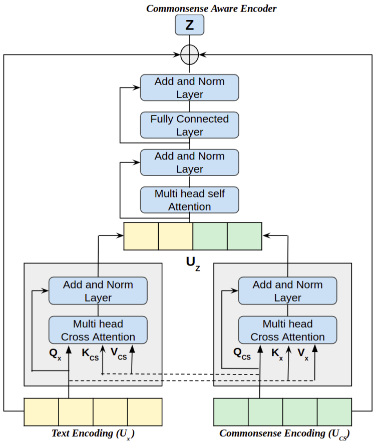


Figure 3: Commonsense-aware encoder internal architecture

4.2 Commonsense-aware Complaint Detection (*C2D*)

We propose Commonsense-aware Complaint Detection, a commonsense-aware unified generative framework to solve the task of explainable complaint detection. We divide our approach into three steps for better understanding: 1) Commonsense Extraction Module, 2) Commonsense-aware Transformer Model, and 3) Reinforcement Learning-based Training.

4.2.1 Commonsense Extraction Module

We use a commonsense extraction module to provide more context in the form of commonsense reasoning to a review, as customers' reviews are usually short and cursory. We use the ATOMIC dataset (Sap et al., 2019) as our knowledge base for the commonsense extraction module. The ATOMIC knowledge base provides commonsense reasoning for six commonsense relations for the entity involved in that event, such as what is the effect of the event on the entity (xEffect), what is their need from the event (xNeed), what they want from the event (xWant), etc. In our problem statement of complaint detection, the event refers to the tweet instance, and we want to understand what is the need and desires of customers from their tweets, so we consider only two commonsense relations[3]: xNeed and xWant. To generate the commonsense reasoning from customer's reviews, we employed a pre-trained BART (Lewis et al., 2019) based language model COMET (Hwang et al., 2021), which is fine-tuned on the above-mentioned ATOMIC dataset as this model is more suitable to provide commonsense reasoning for the unseen events (Sabour et al., 2021a). The Commonsense Extraction Module is outlined as follows: (i) We append two commonsense relation tokens (xNeed and xWant) to the customer's review for each X_i. (ii) We then feed these two commonsense relations concatenated inputs to the pre-trained COMET model to generate two commonsense reasoning $cs^{r_{need}}$ and $cs^{r_{want}}$ for xNeed and xWant relation tokens, respectively. To obtain the final commonsense reasoning CS for each review X_i, we concatenate the two generated commonsense reasoning $CS = cs^{r_{need}} \oplus cs^{r_{want}}$.

4.2.2 Commonsense-aware Transformer

To leverage the commonsense reasoning CS obtained from the Commonsense Extraction Module, we have proposed two variations of commonsense-aware encoder-decoder architecture (*ConC2D* and

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

FuseC2D) that are capable of incorporating CS in their sequence-to-sequence learning process as detailed below:

Contextualizer C2D (*ConC2D*): Given an input review X_i and corresponding commonsense reasoning CS, the task to generate the target sequence $Y_i^{'}$ can be modeled as the following conditional text generation model: $P_\theta(Y_i^{'}|X_i, CS)$, where θ is a set of model parameters. *ConC2D* models this conditional probability by first concatenating the tokens of the input review X_i and the commonsense reasoning CS separated by a unique token <*SEP*> to provide us a final input sequence:$R_i = X_i \oplus CS$. Now, we are given a pair of input sentences and target sequences (R_i, Y_i), the first step is to feed $R_i = \{x_0, x_1, .., x_n, < SEP >, cs_0.., cs_n\}$ to the encoder module to obtain the hidden representation of input as defined next; $H_{EN} = G_{Encoder}(\{x_0, x_1, .., x_n, < SEP >, cs_0.., cs_n\})$ where $G_{Encoder}$ represents encoder computations.

After obtaining the hidden representation, H_{EN}, we will feed H_{EN} and all the output tokens till time step $t-1$ represented as $Y_{<t}$ to the decoder module to obtain the hidden state at time step t, defined as: $H_{DE}^{t} = G_{Decoder}(H_{EN}, Y_{<t})$ where $G_{Decoder}$ denotes the decoder computations.

The conditional probability for the predicted output token at t^{th} time step, given the input and previous $t-1$ tokens is calculated by applying the softmax function over the hidden state, H_{DEC}^{t}:

$$P_\theta(Y_t^{'}|R, Y_{<t}) = F_{softmax}(\theta^T H_{DE}^{t}) \qquad (1)$$

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

Fused C2D (*FuseC2D*): In this setup, we first feed both input review X_i and commonsense reasoning CS to a pre-trained BART encoder to obtain encoded representations, U_x and U_{cs}, respectively. To fuse the information between these two representations, we have proposed a commonsense-aware encoder (shown in Fig. 3), an extension of the original transformer encoder (Vaswani et al., 2017). We create two triplets of queries, keys and values matrices corresponding to U_x, U_{cs}, respectively: (Q_x,K_x,V_x), (Q_{cs},K_{cs},V_{cs}). Unlike the original transformer encoder where we project the same input as query, key, and value, in *FuseC2D*, we propose a cross-attention layer consisting of two sublayers of multi-head-cross attention and normalization layer that exchanges the key and value by considering (Q_x,K_{cs},V_{cs}) and (Q_{cs},K_x,V_x) as inputs to cross attention layer which computes cross infused vector representation as defined below: $Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V$ where (Q,K,V) represents the set of the query, key, and value and d_k represents the dimension of the query and key.

This cross-attention layer facilitates the exchange of information between U_x and U_{cs}. Now, these multi-head cross attention outputs ($U_{x->cs}$ and $U_{cs->x}$) contain information about each other. Following this, we concatenate $U_{x->cs}$ and $U_{cs->x}$ and pass concatenated output U_z through a self-attention layer, normalization layers, and fully connected layers with residual connections to obtain the output of the commonsense-aware encoder. At last, we concatenate input text representation U_x and commonsense representation U_{cs} to the commonsense-aware encoder's output to obtain the final commonsense-aware input representation vector, Z. Then, we feed Z to an autoregressive decoder similar to the defined Equation 1.

4.2.3 Reinforcement Learning based Training

We initialize our model's weights θ with weights of the pre-trained sequence to sequence generative model (BART-base). We then fine-tune the model with the following two training objective functions: 1) Maximum likelihood estimation (MLE) objective function, which works in a supervised manner to optimize the weights, θ, as defined in Equation 2.

$$\max_{\theta} \prod_{t=0}^{T} P_\theta(Y_t^{'}|X_i, Y_{<t}) \qquad (2)$$

2) On the top of maximum likelihood estimation (MLE) objective function, we also employed a reward-based training objective function. Inspired from (Sancheti et al., 2020), we use a BLEU (Papineni et al., 2002) based reward function as it calculates the overlap between the target sequence, Y_i, and the predicted sequence, $Y_i^{'}$. We define BLEU based Reward R_{BLEU} as: $R_{BLEU} = (BLEU(Y_i^{'}, Y_i) - BLEU(Y_i^{g}, Y_i))$, where $Y_i^{'}$ denotes the output sequence sampled from conditional probability distribution (Eq. 1) at each decoding time stamp and Y_i^{g} denotes the output sequence obtained by greedily maximizing the conditional probability distribution at each time step. To maximize the expected reward, R_{BLEU} of $Y_i^{'}$, we use the policy gradient technique (Sutton et al., 1999) which is defined as:

	Complaint (CI)		Severity (SC)		Cause (CE)		
Model	F1	A	F1	A	JS	HD	ROS
SOTA(Jin and Aletras, 2021)	86.6	87.6	59.4	55.5	-	-	-
ConC2D$_{CI+SC+CE}$	90.7	91.2	73.3	73.3	83.7	75.0	88.7
ConC2D$_{CI+SC+CE+ER}$	90.7	91.1	72.1	73.0	83.4	74.1	88.2
ConC2D$_{CI+SC+CE+PR}$	90.1	90.5	73.1	73.2	83.6	74.5	88.3
ConC2D$_{All}$	**90.8**†	**91.3**†	**73.7**†	**73.4**†	**84.4**†	**75.2**†	88.9†
FuseC2D$_{CI+SC+CE}$	88.9	89.1	73.0	73.2	77.4	71.1	85.8
FuseC2D$_{CI+SC+CE+ER}$	88.1	88.4	72.8	72.1	77.4	70.7	84.2
FuseC2D$_{CI+SC+CE+PR}$	88.8	89.1	72.9	73.0	78.2	72.1	85.1
FuseC2D$_{All}$	88.5	89.5	73.1	**73.4**	78.1	71.3	85.3
Baseline$_1$(Singh et al., 2022)	81.4	82.8	60.3	62.8	76.1	68.2	81.8
BART	88.9	88.9	62.4	62.6	77.1	69.7	84.2
T5	86.7	86.6	68.7	69.3	84.1	74.1	**89.1**
SpanBERT	-	-	-	-	74.8	70.6	83.5
Baseline$_2$	-	-	-	-	79.2	74.1	87.8

Table 1: Results of different baselines and the two proposed frameworks, *ConC2D* and F*useC2D*. 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 %. JS: Jaccard Similarity, HD: Hamming Distance, and ROS: Ratcliff-Obershelp Similarity. The maximum scores attained are represented by bold-faced values. The † denotes statistically significant findings.

$$\nabla_{\theta}J(\theta) = R_{BLEU} \cdot \nabla_{\theta}logP(Y_i^{'}|X_i, CS; \theta)$$

5 Experimental Results and Analysis

This section describes the 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 multitask models and other baseline models?

RQ2: Out of *ConC2D* and *FuseC2D*, which technique performs better?

RQ3: What is the effect of different task combinations in our framework?

RQ4: What is the impact of commonsense knowledge and Reinforcement Learning on the performance of our framework?

RQ5: Is *ConC2D* able to outperform state-of-the-art models for CI and SC tasks on full-shot and few-shot settings?

5.1 Baselines Setup

Multitask systems: Motivated by a recent work in multitask CI framework we develop Baseline$_1$ (Singh et al., 2022) model as one of the multitask baselines. We implement the Baseline$_1$ model for the joint learning of CI, SC, and CE with PR and ER as additional tasks, keeping the experimental setup the same as our current work.

Baselines for Cause Extraction Task: Since cause extraction is a new task in the area of complaint analysis, we drew inspiration from the work of Poria et al. (2021) in the emotion recognition domain and used a pre-trained SpanBERT base model fine-tuned on the SQuAD 2.0 dataset (Rajpurkar et al., 2018). The SpanBERT baseline is used for the CE task only. We also added another baseline for the CE task, Baseline$_2$ where complete review/text is considered as the cause.

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. We fine-tune both these models on the proposed dataset with complaint text as the input sequence and concatenated outputs as the target sequence, and training objective defined in Equation 2.

Ablation Study: The *C2D* model comprises two key components: (1) Commonsense Reasoning (CS) and (2) Reinforcement Learning (RL). In order to establish the necessity of both of these components individually, we conduct an ablation study of the proposed framework.

5.2 Experimental Setup

We have performed all the experiments on the Tyrone machine with Intel's Xeon W-2155 Processor having 196 Gb DDR4 RAM and 11 Gb Nvidia 1080Ti GPU. All the models are executed using a nested 10-fold cross-validation approach similar to that of Jin and Aletras (2021). All the proposed models are trained for 20 epochs with a learning rate of $5{\times}10^{-5}$ and batch size of 16. Adam optimizer is used to train the model with adam epsilon value of $1{\times}10^{-8}$. All the models are implemented using Scikit-Learn[4] and pytorch[5] as a backend. For the CI and SC tasks, accuracy and

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

[5] https://pytorch.org/

	Complaint (CI)		Severity (SC)		Cause (CE)		
Model	F1	A	F1	A	JS	HD	ROS
ConC2D$_{CI+SC+CE}$	**90.7**†	**91.2**†	73.3	**73.3**†	**83.7**†	**75.1**†	**88.7**†
-RL	88.7	88.7	64.5	65.1	81.1	73.2	85.5
-CS	89.8	90.3	**73.4**†	**73.3**†	82.3	74.2	86.1
FuseC2D$_{CI+SC+CE}$	88.9	89.1	73.0	73.2	77.4	71.1	85.8
-RL	88.7	88.8	69.6	69.5	77.2	70.0	84.2
-(RL+CS)	88.9	88.9	62.4	62.6	77.1	69.7	84.2

Table 2: Results of the ablation studies performed on the proposed framework, *C2D*'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.

macro-F1 metrics are used to evaluate predictive performance. For the quantitative assessment of the CE task, we used the Jaccard Similarity (JS), Hamming Distance (HD), and Ratcliff-Obershelp Similarity (ROS) metrics. Dataset split follows the standard 80% training (2671 instances), 10% validation (344 instances), and 10% testing (344 instances).

5.3 Results and Discussions

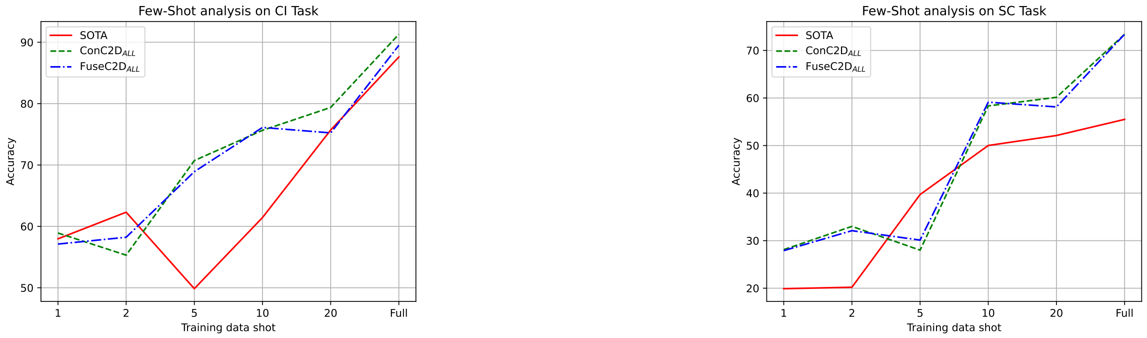
This study aims to enhance the performance of CI, SC, and CE tasks by incorporating two secondary tasks (ER and PR). We present our findings and analyses, focusing solely on CI, SC, and CE as the primary tasks in all combinations.

(RQ1) As can be observed from Table 1, the proposed models, *ConC2D*$_{All}$ and *FuseC2D*$_{All}$ outperform all the other baselines for CI, SC, and CE tasks by a significant margin illustrating the superiority of pre-trained sequence to sequence language models. *Complaints* dataset is a Twitter-based dataset with fixed character constraints, due to which the *ConC2D* model can capture more information in the form of commonsense reasoning compared to all the other baselines. Sample sentences from the dataset, such as *'Thank you', 'I need help'* depict a lack of contextual information. Even in baseline setups, generative baselines (BART and T5) consistently outperform other baselines for all the tasks. These findings validate our idea of re-framing the multitask problem as a text-to-text generation task.

(RQ2) Unexpectedly, *FuseC2D* does not improve performance over *ConC2D* as their scores are similar across all multitask variants for all tasks. This is likely because fusion techniques are more effective for fusing different modalities (vision or acoustic) with text data. Some studies also showed that direct concatenation performs similarly to fusion-based methods (Sridhar and Yang, 2022). Another reason for this is the size of our dataset (around 3,449 samples), which may not be enough to train a fusion-based model.

(RQ3) It is also evident from the table that *ConC2D*$_{ALL}$ which includes all the auxiliary tasks, outperforms all the other corresponding model's task variants illustrating that the model can learn the mapping between different tasks during the decoding process. However, unlike multitasking methods, where adding auxiliary tasks results in significant improvement, here the improvement is more subtle. One can also observe from the results that *ConC2D*$_{ALL}$ and generative baseline models outperform other baselines by a significant margin for the other two tasks (SC and CE) illustrating that generative models maintain consistent performance across all the tasks.

(RQ4) Ablation Study: It is evident from Table 2 that when we remove the RL component from *ConC2D*, we notice a significant drop in all the task's performances but especially in the case of the SC task with a drop of 11% in accuracy. This drop-in performance can be attributed to the fact that reward-based learning teaches the model to learn the mapping between different tasks and encourages the model to generate sequences that have a higher overlap with the target sequence, improving the performance of all tasks, especially severity classification. As *ConC2D* - RL only comprises of commonsense component (CS), we can see how commonsense alone is improving the performance of our model over the baseline model, i.e., -(RL+CS) on SC and CE tasks illustrating how providing extra context to model is aiding the predictions of different subtasks. This argument can be further bolstered when the same trend is observed between the *FuseC2D*, *FuseC2D* - RL, and -(RL+CS). Similarly, when we remove the commonsense component (CS) from *ConC2D*, we observe no effect on SC and only a slight drop in the perfor-



(a) Few-Shot analysis for CI Task

(b) Few-Shot analysis for SC Task

Figure 4: Comparative performance of our proposed models, *ConC2D* and *FuseC2D* with the SOTA model on Few-Shot settings for the primary tasks (CI and SC).

mance of CI and CE tasks. However, together with the CS component, the *ConC2D* model can outperform all the ablated models and baseline models on all subtasks' overall evaluation metrics illustrating the contribution of each component.
(RQ5) Comparison with State-of-the-art Technique (SOTA): Both of our proposed models are able to outperform the SOTA model (Jin and Aletras, 2021) on CI and SC tasks. *ConC2D$_{All}$* outperforms the SOTA by a significant margin of 4.2% and 32% on CI and SC tasks, respectively. The reasons for these improvements can also be attributed to the facts: 1) Both *ConC2D* and *FuseC2D* are leveraging the pretrained BART model's knowledge which already has been trained on a huge corpus of data, 2) Both of these models have extra context due to which they are making better predictions, and 3) Adding an auxiliary task of cause extraction that enhances the model's performance for CI and SC tasks. We also compare the performance of our models with SOTA in a few-shot setting where we sampled the training data based on the number of examples per label: [1, 2, 5, 10, 20, Full] (shown in Fig. 4). It can be observed from Fig. 4a both of our models consistently outperform SOTA on all shots (except 2-shot) in the CI task. A similar trend is observed for the SC task (Fig. 4b). These few-shot experiments: 1) illustrate the strength of the generative language model in data-constrained and low-resource settings, and 2) further validate our approach of using the generative language model to solve multitask complaint detection tasks.

All of the results are statistically significant[6] (Welch, 1947).

5.4 Qualitative Analysis

We noticed that tweets with vital complaint signs, such as accusatory expressions or blame-related terms, are less misclassified. The qualitative study of the complaint severity predictions obtained by the SOTA (Jin and Aletras, 2021) and the best performing contextualizer system on a few sample test instances are shown in Table 3. The table shows that the CE task combined with CI and commonsense reasoning led to improved predictions than the SOTA system that lacks these elements. It can also be observed that for both the example instances, both the models correctly predict them as complaints, but the severity level is correctly predicted only by the proposed model.
We also perform a qualitative analysis for the CE task as shown in Table 4. The first row represents the causal span that human annotators picked as relevant for classification. The spans obtained by the proposed models are shown in the following two rows. The rationale for this varies between models, even though the model makes an accurate prediction for the complaint detection and severity classification task.

We also assess the linguistic aspect of the model by involving one expert in the English language, affiliated with the authors, who independently evaluated 100 generated causes from our models (ConC2D and FuseC2D). The evaluator rated the quality of the responses on a scale of 1 (worst), 3 (moderate), and 5 (best) using three predefined metrics:
a) Fluency: The generated cause by the model should be syntactically or grammatically correct.
b) Faithfulness: It measures the factual correctness of the cause with respect to the input tweet.

[6] We used the Student's t-test (p-value < 0.04).

Tweet text	SOTA	Proposed	Actual Label
The <USER> stinks 10mins to take order & stop asking my name like we're friends	complaint blame	complaint disapproval	complaint disapproval
Hey guys, I love this product featured today but don't see a price? Help a girl out?	complaint disapproval	complaint no explicit reproach	complaint no explicit reproach

Table 3: Qualitative study of the CI and SC predictions by the SOTA (Jin and Aletras, 2021) and the proposed model (*ConC2D*). 'Actual Label': true labels for CI and SC tasks, the red colored text indicates the causal span annotation of the sentence.

Model	Text	Severity
Human Annotator	No email today :(checked spam and everything... I normally get all my Best Buy emails fine.	Disapproval
FuseC2D	No email today :(checked spam and everything ... I normally get all my Best Buy emails fine.	Disapproval
ConC2D	No email today :(checked spam and everything... I normally get all my Best Buy emails fine.	Disapproval

Table 4: Example instances comparing the cause predicted by human annotators and the *FuseC2D* and *ConC2D* models. The span highlighted in red color was selected by the human annotator and the models to be essential for the prediction. The blue-colored text indicates tokens relevant to the model but not to the human annotators.

Model	Fluency	Faithfulness	Redundancy
ConC2D	4.68	4.05	4.33
FuseC2D	4.57	4.11	4.18

Table 5: Average scores obtained by the two proposed models for quality of the responses generated over three quality metrics of Fluency, Faithfulness, and Redundancy.

c) Redundancy: It measures that the generated cause should not contain repeated information.
We report the average scores for these metrics in Table 5. It can be observed that both these models perform reasonably well in Fluency. This can be attributed to the fact the model extracts the content from input only making it more fluent. For faithfulness and redundancy, models perform well but not as well as in Fluency.

5.5 Error Analysis:

We investigate the possible reasons for the proposed model's errors:
Fuzzy Intentions: The model predicts hidden intent sentences inaccurately. When a user voices a complaint without explicitly relaying the actual reason, the model misclassifies it as non-complaint based on the text's literal meaning. For example, *<USER> congratulations. You have reached popular status and the spamming has begun.* The correct class is complaint but the model misclassifies it as non-compliant. The complainant's absence of straightforward disapproval or accusation could be one of the reasons behind this.
Hallucinations: Predicting out of input sentence information as spans: As generative models like BART are designed to generate output based on the complete vocabulary it is trained on, there are some instances where model generates cause which contains some information that is not present in original input review. For example, for review *The number is incomplete*, the model generates the cause of complaint as *The phone number is incomplete*. However the the word *phone* is not present in the input review.

6 Conclusion

In this paper, the explainability factor has been considered while we try to tackle the complaint detection problem. As explainable AI systems increase trustworthiness and confidence when used in real-time, and complaint detection systems benefit different enterprises for robust customer support, generating rationale behind actions performed is a must. The current work makes two contributions: (a) developing the first explainable complaint detection dataset, which includes annotations of the rationale/causal span employed in decision-making (b) a commonsense-aware unified generative framework has been developed to perform five tasks (CD, SC, CE, PR, and ER) simultaneously. In order to take advantage of the knowledge of sizable pre-trained sequence-to-sequence models, this work demonstrates how a multitasking problem could be phrased as a text-to-text generation task. Our proposed model outperforms all the baselines and the SOTA for the three main tasks based on extensive evaluation.
In future, we will focus on expanding explainable complaint identification in the multimodal environment that considers image and text modalities.

Limitations

We attempted to develop a novel framework for explainable complaint identification in a multitask setting. 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.
(3) In some cases, users use an implicit sarcastic tone while writing complaints. In the current setup, sarcasm detection is not considered as a separate task. Thus the proposed system will not be capable of detecting complaints with implicit sarcasm.

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

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020a. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020b. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

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*.

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.

David Gunning, Mark Stefik, Jaesik Choi, Timothy Miller, Simone Stumpf, and Guang-Zhong Yang. 2019. XAI - explainable artificial intelligence. *Sci. Robotics*, 4(37).

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. 2021. Modeling the severity of complaints in social media. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 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*.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2020. Hatexplain: A benchmark dataset for explainable hate speech detection.

E Olshtain and L Weinbach. 1985. Complaints: A study of speech act behavior among native and nonnative speakers of hebrew. the prag-matic perspective.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Soujanya Poria, Navonil Majumder, Devamanyu Hazarika, Deepanway Ghosal, Rishabh Bhardwaj, Samson Yu Bai Jian, Pengfei Hong, Romila Ghosh, Abhinaba Roy, Niyati Chhaya, et al. 2021. Recognizing emotion cause in conversations. *Cognitive Computation*, 13(5):1317–1332.

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.

Daniel Preotiuc-Pietro, Mihaela Gaman, and Nikolaos Aletras. 2019. Automatically identifying complaints in social media. *arXiv preprint arXiv:1906.03890*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

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. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Sahand Sabour, Chujie Zheng, and Minlie Huang. 2021a. Cem: Commonsense-aware empathetic response generation. *arXiv preprint arXiv:2109.05739*.

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

Tulika Saha, Apoorva Upadhyaya, Sriparna Saha, and Pushpak Bhattacharyya. 2021. Towards sentiment and emotion aided multi-modal speech act classification in twitter. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5727–5737.

Abhilasha Sancheti, Kundan Krishna, Balaji Srinivasan, and Anandhavelu Natarajan. 2020. Reinforced rewards framework for text style transfer.

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. 2023. Complaint and severity identification from online financial content. *IEEE Transactions on Computational Social Systems*.

Apoorva Singh, Arousha Nazir, and Sriparna Saha. 2022. 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, Sriparna Saha, Md Hasanuzzaman, and Kuntal Dey. 2021. Multitask learning for complaint identification and sentiment analysis. *Cognitive Computation*, pages 1–16.

Rohit Sridhar and Diyi Yang. 2022. Explaining toxic text via knowledge enhanced text generation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 811–826, Seattle, United States. Association for Computational Linguistics.

Richard S. Sutton, David McAllester, Satinder Singh, and Yishay Mansour. 1999. Policy gradient methods for reinforcement learning with function approximation. In *Proceedings of the 12th International Conference on Neural Information Processing Systems*, NIPS'99, page 1057–1063, Cambridge, MA, USA. MIT Press.

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.

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

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

A Appendix

A.1 Annotation Instructions

Before the annotation procedure began the concept of causal span detection and extraction, specifically for complaint causes was defined to the annotators. *Complaint Cause* is a portion of the text that expresses why the user feels compelled to file a complaint. It is the speech act used by the individual to describe the circumstances in which their expectations have been violated. The annotators were

Tweet	Complaint	Severity	Polarity	Emotion
<USER> stiching on chino's bought 1 month ago in <LOCATION> ripped, no help from shop.	Complaint	Blame	Negative	Sadness
Great phone customer service from <USER> thank you <USER>	Non-Complaint	Non-Complaint	Positive	Happiness
<USER> It started yesterday, but i try again it could work normal. But since last night its just like this.	Complaint	No explicit reproach	Neutral	Other

Table 6: Example instances of annotated *X-Complain* dataset. The red-colored text denotes the rationale span.

Model	IOU F1	Token F1	AUPRC
HateXplain best model (Mathew et al., 2020)	0.222	0.506	0.841
ConC2D	**0.243**	**0.553**	0.851
FuseC2D	0.237	0.551	**0.859**

Table 7: Results of our t wo best performing models, ConC2D and FuseC2D, and the best performing model from the work (Mathew et al., 2020) on the HateXplain dataset. IOU F1: Intersection over Union F1 score, AUPRC: Area Under Precision-Recall Curve. The maximum scores attained are represented by bold-faced values.

instructed to mark causal span based on first encounter with strong expression of complaint reason in the tweet, it could be any length of text. Some example instances were also given before the annotation procedure began.

For example: *'Disappointed with the seller. Product delivery was quite fast, but the display is scratched on the front and left side.'*

Causal span annotated: *'the display is scratched on the front and left side.'* In the given example the first sentence shows weak intensity of complaint cause. But the selected causal span shows strong cause of complaint.

In circumstances where annotators disagree, the final label is determined through majority voting (cause vs. no cause). For ambiguous instances, the authors discuss and clarify them with the annotators during the annotation procedure.

A.2 Statistics related to *X-CI* dataset

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

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

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

(3) The distribution of tweets across the emotion classes (ER task) as mentioned in the work of (Singh et al., 2022) is as follows: 844 tweets belong to 'Anger', 7 tweets belong to 'Disgust', 8 tweets belong to 'Fear', 473 tweets belong to 'Joy', 1,479 tweets belong to 'Other', 626 tweets belong to 'Sadness', and 12 tweets belong to 'Surprise'.

(4) The distribution of tweets across the sentiment classes (Singh et al., 2022) is as follows: 1,041 tweets belong to 'Negative', 1,198 tweets belong to 'Neutral', and 1,210 tweets belong to 'Positive'. For the CE task, we employed the macro-F1 metric to assess inter-rater agreement based on earlier studies on span extraction (Rajpurkar et al., 2016) and obtained a 0.77 F1 score, indicating that the annotations are of decent quality (mentioned in Section 3.2, line no. 206 of the manuscript). The other three tasks, SC, ER, and PR, have been introduced in other works. The inter-rater agreement for the main task SC is 0.64 (Fleiss' Kappa score), as reported in work by (Jin and Aletras, 2021). The inter-rater agreement scores for the auxiliary tasks ER and PR are 0.68 and 0.82 (Cohen-Kappa score), as reported in work by (Singh et al., 2022). Table 6 shows few example instances of causal span annotation from the *X-CI* dataset.

Model	Total Parameters
T5	222903552
ConC2D	204199168
FuseC2D	139420416

Table 8: The trainable parameters for the proposed models (ConC2D, FuseC2D) and the SOTA model (T5)

A.3 Performance of Proposed Models on Different Dataset

To evaluate the performance of our model to detect cause spans from longer texts, we tested our approach on a benchmark dataset ('HateXplain dataset') released in the work Mathew et al. (2020) that contains posts from the online social networking site GAB. It is also annotated with explanations for hate speech labels.

We compared our two best models (ConC2D and FuseC2D) with the best model mentioned in Hatexplain paper (Mathew et al., 2020), and the results are shown in Table 7. We can observe from the results that both of these models (ConC2D and FuseC2D) are able to outperform their best model on IOUF1, TokenF1, and AUPRC metrics.

A.4 Parameter Comparison Study

We report and compare the number of trainable parameters for our models (ConC2D, FuseC2D) with the best SOTA model (T5) in Table 8. It can be observed from Table 8 that ConC2D has the least number of parameters and is able to outperform both models on most of the metrics.

ACL 2023 Responsible NLP Checklist

A For every submission:

□ A1. Did you describe the limitations of your work?
Left blank.

□ A2. Did you discuss any potential risks of your work?
Left blank.

□ A3. Do the abstract and introduction summarize the paper's main claims?
Left blank.

□ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B □ Did you use or create scientific artifacts?

Left blank.

□ B1. Did you cite the creators of artifacts you used?
Left blank.

□ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Left blank.

□ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Left blank.

□ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Left blank.

□ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Left blank.

□ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Left blank.

C □ Did you run computational experiments?

Left blank.

□ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Left blank.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

□ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Left blank.

□ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Left blank.

□ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Left blank.

D □ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

□ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Left blank.

□ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Left blank.

□ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Left blank.

□ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Left blank.

□ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Left blank.