Multi-Task Learning and Adapted Knowledge Models for Emotion-Cause Extraction

Elsbeth Turcan[1*] **Shuai Wang**[3] **Rishita Anubhai**[3]
Kasturi Bhattacharjee[3] **Yaser Al-Onaizan**[3] **Smaranda Muresan**[2,3]
[1]Department of Computer Science, Columbia University
[2]Data Science Institute, Columbia University
[3]Amazon AI
{eturcan, smara}@cs.columbia.edu
{wshui, ranubhai, kastb, onaizan}@amazon.com

Abstract

Detecting what emotions are expressed in text is a well-studied problem in natural language processing. However, research on finer-grained emotion analysis such as what causes an emotion is still in its infancy. We present solutions that tackle both emotion recognition and emotion cause detection in a joint fashion. Considering that common-sense knowledge plays an important role in understanding implicitly expressed emotions and the reasons for those emotions, we propose novel methods that combine common-sense knowledge via adapted knowledge models with multi-task learning to perform joint emotion classification and emotion cause tagging. We show performance improvement on both tasks when including common-sense reasoning and a multi-task framework. We provide a thorough analysis to gain insights into model performance.

1 Introduction

Utterance and document level emotion recognition has received significant attention from the research community (Mohammad et al., 2018; Poria et al., 2020a). Given the utterance *Sudan protests: Outrage as troops open fire on protestors* an emotion recognition system will be able to detect that *anger* is the main expressed emotion, signaled by the word "outrage". However, the semantic information associated with expressions of emotion, such as the cause (the thing that triggers the emotion) or the target (the thing toward which the emotion is directed), is important to provide a finer-grained understanding of the text that might be needed in real-world applications. In the above utterance, the cause of the anger emotion is the event "troops open fire on protestors", while the target is the entity "troops" (see Figure 1) .

Research on finer-grained emotion analysis such as detecting the cause for an emotion expressed in text is in its infancy. Most work on emotion-cause detection has utilized a Chinese dataset where the cause is always syntactically realized as a clause and thus was modeled as a classification task (Gui et al., 2016). However, recently Bostan et al. (2020) and Oberländer and Klinger (2020) argued that in English, an emotion cause can be expressed syntactically as a clause (*as troops open fire on protestors*), noun phrase (*1,000 non-perishable food donations*) or verb phrase (*jumped into an ice-cold river*), and thus we follow their approach of framing emotion cause detection as a sequence tagging task.

We propose several ways in which to approach the tasks of emotion recognition and emotion cause tagging. First, these two tasks should not be independent; because the cause is the trigger for the emotion, knowledge about what the cause is should narrow down what emotion may be expressed, and vice versa. Therefore, we present a multi-task learning framework to model them jointly. Second, considering that common-sense knowledge plays an important role in understanding implicitly expressed emotions and the reasons for those emotions, we explore the use of common-sense knowledge via adapted knowledge models (COMET, Bosselut et al. (2019)) for both tasks. A key feature of our approach is to combine these adapted knowledge models (i.e., COMET), which are specifically trained to use and express common-sense knowledge, with pre-trained language models such as BERT, (Devlin et al., 2019).

Our primary contributions are three-fold: (i) an under-studied formulation of the emotion cause detection problem as a sequence tagging problem; (ii) a set of models that perform the emotion classification and emotion cause tagging tasks jointly while

[*]Work done during an internship with Amazon AI.

using common-sense knowledge (subsection 4.2) with improved performance (section 6); and (iii) analysis to gain insight into both model performance and the GoodNewsEveryone dataset that we use (Bostan et al., 2020) (section 7).

2 Related Work

Emotion detection is a widely studied subfield of natural language processing (Mohammad et al., 2018; Poria et al., 2020a), and has been applied to a variety of text genres such as fictional stories (Alm et al., 2005), news headlines (Strapparava and Mihalcea, 2010), and social media, especially microblogs such as Twitter (Abdul-Mageed and Ungar, 2017; Kiritchenko et al., 2014; Rosenthal et al., 2019; Mohammad et al., 2018). Earlier work, including some of the above, focused on feature-based machine learning models that could leverage emotion lexicons (Mohammad and Turney, 2013)), while recent work explores deep learning models (e.g., Bi-LSTM and BERT) and multi-task learning (Xu et al., 2018; Demszky et al., 2020).

However, comparatively few researchers have looked at the semantic roles related to emotion such as the cause, the target or the experiencer, with few exceptions for Chinese (Gui et al., 2016; Chen et al., 2018; Xia and Ding, 2019; Xia et al., 2019; Fan et al., 2020; Wei et al., 2020; Ding et al., 2020), English (Mohammad et al., 2014; Ghazi et al., 2015; Kim and Klinger, 2018; Bostan et al., 2020; Oberländer et al., 2020; Oberländer and Klinger, 2020) and Italian (Russo et al., 2011). We highlight some of these works here and draw connection to our work. Most recent work on emotion-cause detection has been carried out on a Chinese dataset compiled by Gui et al. (2016). This dataset characterizes the emotion and cause detection problems as clause-level pair extraction problem – i.e., of all the clauses in the input, one is selected to contain the expression of an emotion, and one or more (usually one) are selected to contain the cause of that emotion. Many publications have used this corpus to develop novel and effective model architectures for the clause-level classification problem (Chen et al., 2018; Xia and Ding, 2019; Xia et al., 2019; Fan et al., 2020; Wei et al., 2020; Ding et al., 2020). The key difference between this work and ours is that we perform cause detection as a sequence-tagging problem: the cause may appear anywhere in the input, and may be expressed as any grammatical construction (a noun phrase, a verb phrase, or a clause). Moreover, we use common sense knowledge for both tasks (emotion and cause tagging), through the use of adapted language models such as COMET.

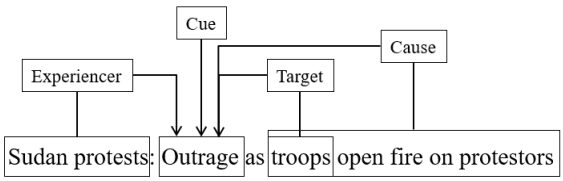


Figure 1: An example of the semantic roles annotated by Bostan et al. (2020)

For English, several datasets have been introduced (Mohammad et al., 2014; Kim and Klinger, 2018; Ghazi et al., 2015; Bostan et al., 2020; Poria et al., 2020b), and emotion cause detection has been tackled either as a classification problem (Mohammad et al., 2014), or as a sequence tagging or span detection problem (Kim and Klinger, 2018; Ghazi et al., 2015; Oberländer and Klinger, 2020; Poria et al., 2020b). We particularly note the work of Oberländer and Klinger (2020), who argue for our problem formulation of cause detection as sequence tagging rather than as a classification task supported by empirical evidence on several datasets including the GoodNewsEveryone dataset (Bostan et al., 2020) we use in this paper. One contribution we bring compared to these models is that we formulate a multi-task learning framework to jointly learn the emotion and the cause span. Another contribution is the use of common-sense knowledge through the use of adapted knowledge models such as COMET (both in the single models and the multi-task models). Ghosal et al. (2020) have very recently shown the usefulness of common-sense reasoning to the task of conversational emotion detection.

3 Data

For our experiments, we use the GoodNewsEveryone corpus (Bostan et al., 2020), which contains 5,000 news headlines labeled with emotions and semantic roles such as the target, experiencer, and cause of the emotion, as shown in Figure 1.[1] We focus on the emotion detection and cause tagging tasks in this work. To our knowledge, GoodNewsEveryone is the largest English dataset labeled for

[1]While the dataset labels both the most dominant emotion expressed in text and the reader's emotion, for this paper we only focus on the former.

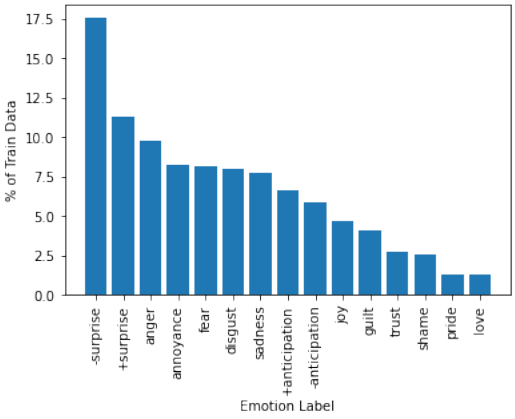


Figure 2: Distribution of adjudicated emotion labels in the GoodNewsEveryone train data, as a percentage of the data points. "Positive" and "Negative" are abbreviated as + and -.

both of these tasks.

In our experiments, we limit ourselves to the data points for which a cause span was annotated (4,798). We also note that this dataset uses a 15-way emotion classification scheme, an extended set including the eight basic Plutchik emotions as well as additional emotions like *shame* and *optimism*. While a more fine-grained label set is useful for capturing subtle nuances of emotion, many external resources focus on a smaller set of emotions. We also note that the label distribution of this dataset heavily favors the more basic emotions, as shown in Figure 2. Therefore, for our work, we choose to limit ourselves to the six Ekman emotions (*anger*, *fear*, *disgust*, *joy*, *surprise*, and *sadness*). We also choose to keep *positive surprise* and *negative surprise* separated, to avoid severely unbalancing the label distribution for our experiments. We randomly split the remaining data (2,503 data points) into 80% train, 10% development, and 10% test.

4 Models

An important feature showcased by the GoodNewsEveryone dataset is that causes of emotions can be expressed through different syntactic constituents such as clauses, verb phrases, or noun-phrases. Thus, we approach the cause detection problem as a sequence tagging problem using the IOB scheme (Ramshaw and Marcus, 1995): $\mathcal{C} = \{\text{I-cause}, \text{O}, \text{B-cause}\}$. Our approach is supported by very recent results by Oberländer and Klinger (2020) and Yuan et al. (2020) who show that modeling emotion cause detection as a sequence tagging problem is better suited than a clause classification problem, although not much current work has yet adopted this formulation. We tackle the emotion detection task as a seven-way classification task with $\mathcal{E} = \{\text{anger}, \text{disgust}, \text{fear}, \text{joy}, \text{sadness}, \text{negative surprise}, \text{positive surprise}\}$.

4.1 Single-Task Models

As a baseline, we train single-task models for each of emotion classification and cause span tagging. We use a pre-trained BERT language model[2] (Devlin et al., 2019), which we fine-tune on our data, as the basis of this model. Our preprocessing strategy for all of our models consists of the pretrained BERT vocabulary and WordPiece tokenizer[3] (Wu et al., 2016) from Huggingface (Wolf et al., 2020). Therefore, for a sequence of n WordPiece tokens, our input to the BERT model is a sequence of $n+2$ tokens, $X = [[\text{CLS}], x_1, x_2, ...x_n, [\text{SEP}]]$, where each x_i is from a finite WordPiece vocabulary and [CLS] and [SEP] are BERT's begin and end tokens. Passing X through BERT yields a sequence of vector hidden states $H = [h_{[CLS]}, h_1, h_2, ..., h_n, h_{[SEP]}]$ with dimension $d_{BERT} = 768$. For emotion classification, we pool these hidden states and allow hyperparameter tuning to select the best type: selecting the [CLS] token ($h_f = h_{[CLS]}$), mean pooling ($h_f = \frac{\sum_{i=1}^{n} h_i}{n}$), max pooling ($h_{f,j} = \max h_{i,j}$), or attention as formulated by Bahdanau et al. (2015):

$$h_f = \sum_{i=1}^{n} \alpha_i h_i \tag{1}$$

where $\alpha_i = \frac{\exp(W_a h_i + b_a)}{\sum_{j=1}^{n} \exp(W_a h_j + b_a)}$ for trainable weights $W_a \in \mathbb{R}^{1 \times d_{BERT}}$ and $b_a \in \mathbb{R}^1$. Then, the final distribution of emotion scores is calculated by a single dense layer and a softmax:

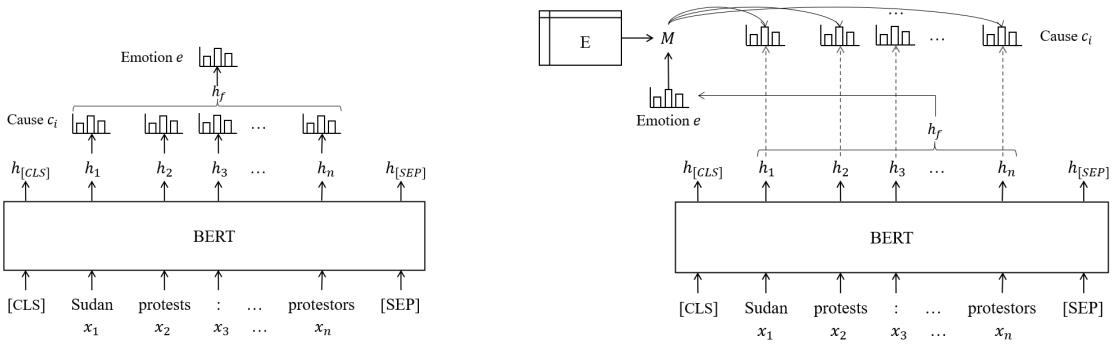
$$e = \text{softmax}(W_e h_f + b_e) \tag{2}$$

with $e \in \mathbb{R}^{|\mathcal{E}|}$ and for trainable parameters $W_e \in \mathbb{R}^{|\mathcal{E}| \times d_{BERT}}$ and $b_e \in \mathbb{R}^{|\mathcal{E}|}$. For cause tagging, a tag probability distribution is calculated directly on each hidden state:

$$c_i = \text{softmax}(W_c h_i + b_c) \tag{3}$$

[2]We use BERT-BASE-UNCASED. We experimented with BERT-BASE-CASED, but it underperformed as the headlines incorporated into GoodNewsEveryone come from different news sources and have different capitalization styles.

[3]In the tagging setting, we ignore all tags predicted for subword tokens and use only the tag of the first subword.



(a) The Multi$_{C\to E}$ model.

(b) The Multi$_{E\to C}$ model.

Figure 3: Our multi-task models.

with $c_i \in \mathbb{R}^{|\mathcal{C}|}$ and for trainable parameters $W_c \in \mathbb{R}^{|\mathcal{C}|\times d_{BERT}}$ and $b_c \in \mathbb{R}^{|\mathcal{C}|}$. We refer to both of these single-task models as BERT; if the task is not clear from the context, we will refer to the emotion detection model as BERT$_E$ and the cause tagging model as BERT$_C$. Our training loss for emotion classification as well as emotion cause tagging is the mean negative log-likelihood (NLL) loss per minibatch of size b:

$$\text{NLL}_{\text{emo}} = -\frac{1}{b}\sum_j \sum_k y_{jk} \log e_{jk} \quad (4)$$

$$\text{NLL}_{\text{cause}} = -\frac{1}{b}\sum_i \sum_j \sum_k y_{ijk} \log c_{ijk} \quad (5)$$

where j is the index of the sentence in the minibatch, k is the index of the label being considered (emotion labels for NLL$_{\text{emo}}$ and IOB tags for NLL$_{\text{cause}}$), i is the index of the i^{th} token in the j^{th} sentence in the minibatch, $y_{jk} \in \{0, 1\}$ is the gold probability of the k^{th} emotion label for the j^{th} sentence, $y_{ijk} \in \{0, 1\}$ is the gold probability of the k^{th} cause tag for the i^{th} token in the j^{th} sentence, and e_{jk} and c_{ijk} are the output probabilities of the k^{th} emotion label and of the k^{th} cause label for the i^{th} token, both for the j^{th} sentence.

4.2 Multi-Task Models

Our hypothesis is that the emotion detection and cause tagging tasks are closely related and can inform each other; therefore we propose three multi-task learning models to test this hypothesis. For all multi-task models, we use the same base architecture (BERT) as the single models. Additionally, for these models, we combine the losses of both tasks and weight them with a tunable lambda parameter: $\lambda\text{NLL}_{\text{emo}} + (1-\lambda)\text{NLL}_{\text{cause}}$, using NLL$_{\text{emo}}$ and NLL$_{\text{cause}}$ from Equation 4 and Equation 5.

Multi. The first model, Multi, is the classical multi-task learning framework with hard parameter sharing, where both tasks share the same BERT layers. Two dense layers for emotion classification and cause tagging operate at the same time from the same BERT layers, and we train both of the tasks simultaneously. That is, we simply calculate our emotion scores e and cause tag scores c from the same set of hidden states H.

We further develop two additional multi-task models with the intuition that we can design more explicit and concrete task dependencies than simple parameter sharing in the representation layer.

Multi$_{C\to E}$. We assume that if a certain text span is given as the cause of an emotion, it should be possible to classify that emotion correctly while looking only at the words of the cause span. Therefore, we propose the Multi$_{C\to E}$ model, the architecture of which is illustrated in Figure 3a. This model begins with the single-task cause detection model, $BERT_C$, which produces a probability distribution $P(y_i|x_i)$ over IOB tags for each token x_i, where $P(y_i|x_i) = c_i$ from Equation 3. Then, for each token, we calculate the probability that it is part of the cause as $P(\text{Cause}|x_i) = P(B|x_i) + P(I|x_i) = 1 - P(O|x_i)$. We feed the resulting probabilities through a softmax over the sequence and use them as an attention distribution over the input tokens in order to pool the hidden representations and perform emotion classification: attention is computed as in Equation 1, where $\alpha_i = \frac{\exp P(\text{Cause}|x_i)}{\sum_{j=1}^{n} \exp P(\text{Cause}|x_i)}$, and emotion classification as in Equation 2. For the Multi$_{C\to E}$ model, we apply teacher forcing at training time, and the gold cause spans are used to

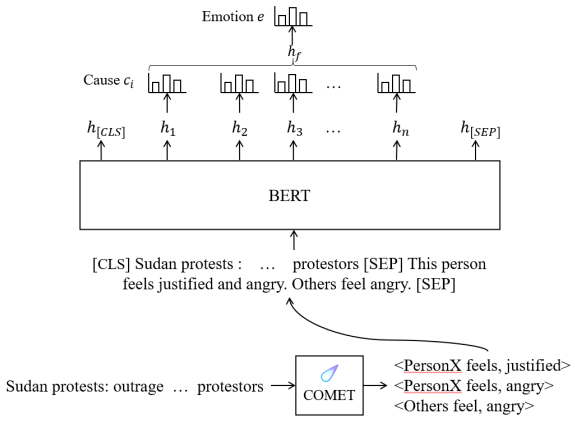


Figure 4: The architecture of our proposed $\text{Multi}_{C\rightarrow E}^{COMET}$ model.

create the attention weights before emotion classification (which means that $P(\text{Cause}|x_i) \in \{0, 1\}$). At inference time, the model uses the predicted cause span instead.

Multi$_{E\rightarrow C}$. Next, we hypothesize that knowledge of the predicted emotion should help us identify salient cause words. The Multi$_{E\rightarrow C}$ model first performs emotion classification, which results in a probability distribution over predicted emotion labels, as in the BERT$_E$ model and Equation 2. We additionally keep an emotion embedding matrix E, where $E[i]$ is a learnable representation of the i-th emotion label (see Figure 3b) with dimension d_e (in our experiments, we set $d_e = 300$). We use the predicted label probabilities e to calculate a weighted sum of the emotion embeddings, i.e., $M = \sum_i e_i \cdot E[i]$. We then concatenate M to the hidden representation of each token and perform emotion cause tagging with a final dense layer, i.e., $c_i = \text{softmax}(W_{c'}[h_i; M] + b_{c'})$, where ; is the concatenation operator and $W_{c'} \in \mathbb{R}^{|\mathcal{C}|\times(d_{BERT}+d_e)}$ and $b_{c'} \in \mathbb{R}^{|\mathcal{C}|}$ are trainable parameters. In the Multi$_{E\rightarrow C}$ model, we again do teacher forcing and use the gold emotion labels before doing the sequence tagging for cause detection (i.e., e is a one-hot vector where the gold emotion label has probability 1 and all other emotion labels have probability 0). At inference time, the model will use the predicted emotion distribution instead.

4.3 Adapted Knowledge Models

Recent work has shown that fine-tuning pre-trained language models such as GPT-2 on *knowledge graph tuples* such as ConceptNet (Li et al., 2016) or ATOMIC (Sap et al., 2018) allows these models to express their implicit knowledge directly (Bosselut et al., 2019). These adapted *knowledge models* (e.g., COMET (Bosselut et al., 2019)) can produce common-sense knowledge on-demand for any entity, relation or event. Considering that common-sense knowledge plays an important role in understanding implicitly expressed emotions and the reasons for those emotions, we explore the use of common-sense knowledge for our tasks, in particular the use of COMET adaptively pre-trained on the ATOMIC event-centric knowledge base. ATOMIC's event relations include "xReact" and "oReact", which describe the feelings of certain entities after the input event occurs. For example, ATOMIC's authors present the example of <PersonX pays PersonY a compliment, xReact, PersonX will feel good>. xReact refers to the feelings of the primary entity in the event, and oReact refers to the feelings of others (in this instance, oReact yields "PersonY will feel flattered"). For example, using the headline "Sudan protests: Outrage as troops open fire on protestors", COMET-ATOMIC outputs that PersonX feels justified, PersonX feels angry, Others feel angry, and so on (Figure 4). To use this knowledge model in our task, we modify our approach by reframing our single-sequence classification task as a sequence-pair classification task (for which BERT can be used directly). We feed our input headlines into COMET-ATOMIC (using the model weights released by the authors), collect the top two outputs for xReact and oReact using beam search decoding, and then feed them into BERT alongside the input headlines, as a second sequence using the SEP token. That is, our input to BERT is now $X = [[\text{CLS}], x_1, x_2, ..., x_n, [\text{SEP}], z_1, z_2, ..., z_m, [\text{SEP}]]$, where z_i are the m WordPiece tokens of our COMET output and are preprocessed in the same way as x_i. We hypothesize that, since pre-trained BERT is trained with a next sentence prediction objective, expressing the COMET outputs as a grammatical sentence will help BERT make better use of them, so we formulate this second sequence as complete sentences (e.g., "This person feels... Others feel...") (Figure 4).

This approach allows us incorporate information from COMET into all our single- and multi-task BERT-based models; the example shown in Figure 4 is our Multi$_{C\rightarrow E}$ model. We refer to the COMET variants of these mod-

	Emotion Macro F1	Emotion Accuracy	Cause Span F1
BERT	37.25 ± 1.30	38.50 ± 0.84	37.49 ± 1.94
BERT^{COMET}	37.74 ± 0.84	38.50 ± 1.14	39.27 ± 1.85
Multi	36.91 ± 1.48	38.34 ± 1.94	38.35 ± 3.89
Multi$_{C\rightarrow E}$	37.74 ± 2.12	38.74 ± 2.07	39.08 ± 3.73
Multi$_{E\rightarrow C}$	38.26 ± 3.28	39.69 ± 3.41	38.83 ± 1.60
Multi^{COMET}	37.06 ± 2.04	39.05 ± 0.98	**39.50** ± 2.25
Multi$^{COMET}_{C\rightarrow E}$	**39.26*** ± 1.13	**40.79** ± 2.17	38.68 ± 1.36
Multi$^{COMET}_{E\rightarrow C}$	37.44 ± 1.37	38.58 ± 1.44	36.27 ± 1.31

Table 1: The results of our models, averaged over five runs with the same five distinct random seeds. The model with the highest mean performance under each metric is bolded. Results marked with a * are statistically significant above the single-task BERT baseline by the paired t-test ($p < 0.05$).

els as: BERT^{COMET} (single-task models) and Multi^{COMET}, Multi$^{COMET}_{C\rightarrow E}$, Multi$^{COMET}_{E\rightarrow C}$ for the three multi-task models.

5 Experimental Setup

Evaluation Metrics For emotion classification, we report macro-averaged F1 and accuracy. For cause tagging, we report exact span-level F1 (which we refer to as *span F1*), as developed for named entity recognition (e.g., Tjong Kim Sang and De Meulder (2003)), where a span is marked as correct if and only if its type and span boundaries match the gold exactly[4].

Training and Hyperparameter Selection The classification layers are initialized randomly from a uniform distribution over $[-0.07, 0.07]$[5], and all the parameters are trained on our dataset for up to 20 epochs, with early stopping based on the performance on the validation data (macro F1 for emotion, span F1 for cause). All models are trained with the Adam optimizer (Kingma and Ba, 2015). We highlight again that for our Multi$_{C\rightarrow E}$ and Multi$_{E\rightarrow C}$ models, we use teacher forced during training to avoid cascading training error. Because the subset of the data we use is relatively small, we follow current best practices for dealing with neural models on small data and select hyperparameters and models using the average performance of five models with different fixed random seeds on the development set. We then base our models' performance on the average of the results from these five runs (e.g., reported emotion F1 is the average of the emotion F1 scores for each of our five runs). For our joint models, since our novel models revolve around using one task as input for the other, we separately tune two sets of hyperparameters for each model, one based on each of the single-task metrics, yielding, for example, one Multi model optimized for predicting emotion and one optimized for predicting cause. The hyperparameters we tune are dropout in our linear layers, initial learning rate of the optimizer, COMET relation type, lambda weight for our multi-task models, and the type of pooler for emotion classification (enumerated in subsection 4.1).

6 Results

We present the results of our models in Table 1[6]. We see that the overall best model for each task is a multi-task adapted knowledge model, with Multi$^{COMET}_{C\rightarrow E}$ performing best for emotion (which is a statistically significant improvement over BERT by the paired t-test, $p < 0.05$) and Multi^{COMET} performing best for cause. These results seem to support our two hypotheses: 1) emotion recognition and emotion cause detection can inform each other and 2) common-sense knowledge is helpful to infer the emotion and the cause for that emotion expressed in text. Specifically, we notice that Multi$_{C\rightarrow E}$ alone does not outperform BERT on either cause or emotion, but Multi$^{COMET}_{C\rightarrow E}$ outperforms both BERT and Multi$_{C\rightarrow E}$ on both tasks. For cause, we also see additional benefits of common-

[4]Our cause tagging task has only one type, "cause", as GoodNewsEveryone is aggregated such that each data point has exactly one emotion-cause pair. We note that this problem formulation leaves open the possibility of multiple emotion-cause pairs.

[5]The default initialization from the `gluon` package: https://mxnet.apache.org/versions/1.7.0/api/python/docs/api/gluon/index.html

[6]Oberländer and Klinger (2020) report an F1 score of 34 in this problem setting on this dataset, but on a larger subset of the data (as they do not limit themselves to the Ekman emotions) and so we cannot directly compare our work to theirs.

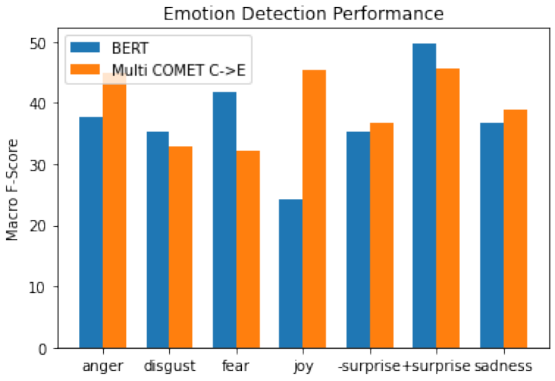


Figure 5: Performance of the BERT and Multi$^{COMET}_{C\rightarrow E}$ models on emotion classification.

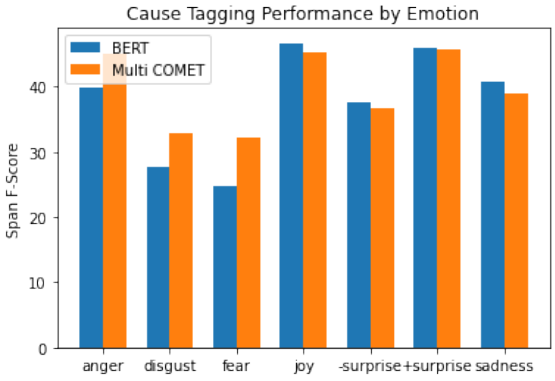


Figure 6: Performance of the BERT and Multi^{COMET} models on cause tagging, broken down by emotion.

sense reasoning alone: BERT^{COMET} outperforms BERT (multi-task modeling alone, Multi, also outperforms BERT for this task) and Multi^{COMET} outperforms Multi. These results speak to the differences between the two tasks, suggesting that common-sense reasoning, which aims to generates implicit emotions, and cause information may be complementary for emotion detection, but that for cause tagging, common-sense reasoning and given emotion information may overlap. The common-sense reasoning we have used in this task (xReact and oReact from ATOMIC) is expressed as possible emotional reactions to an input situation, so this makes intuitive sense.

Finally, we also present per-emotion results for our best model for each task (Multi$^{COMET}_{C\rightarrow E}$ for emotion and Multi^{COMET} for cause) against the single-task BERT baselines in Figure 5 and Figure 6; these per-emotion scores are again the average performance of models trained with each of our five random seeds. We see that each task improves on a different set of emotions: for emotion classification Multi$^{COMET}_{C\rightarrow E}$ consistently improves over BERT by a significant margin on joy and to a lesser extent on anger and sadness. Meanwhile, for cause tagging, Multi^{COMET} improves over BERT on anger, disgust, and fear, while yielding very similar performance on the rest of the emotions.

7 Analysis and Discussion

In order to understand the impact of common-sense reasoning and multi-task modeling for the two tasks, we provide several types of analysis in addition to our results in section 6. First, we include examples of our various models' outputs showcasing the impact of our methods (subsection 7.1). Second, we carry out an analysis of the dataset, focusing on the impact of label variation among multiple annotators on the models' performance (subsection 7.2).

7.1 Example Outputs

We provide some example outputs from our systems for both cause and emotion in Table 2; the various Multi models have been grouped together for readability and because they often produce similar outputs, but the outputs for every model are available in the appendix. In the first example, the addition of COMET to BERT informs the model enough to choose the gold emotion label; in the third and fourth, either COMET or multi-task learning is enough to help the model select key words that should be included in the cause (*return* and *triple shooting*). We also particularly note the second example, in which multi-task learning is needed both for the BERT and BERT^{COMET} models to be able to correctly predict the gold emotion. This suggests that for cause, both common-sense reasoning and emotion classification may carry overlapping useful information for cause tagging, while for emotion, different instances may be helped more by different aspects of our models.

7.2 Label Agreement

Upon inspection of the GoodNewsEveryone data, we discover significant variation in the emotion labels produced by annotators as cautioned by the authors in their original publication[7]. From our inspection of the development data, we see recur-

[7]While the authors selected data according to agreement on the emotion labeling task, they found that in only 75% of cases do at least 3 annotators agree, with diminishing rates of agreement for more annotators.

BERT	Multitask	BERT^{COMET}	Multitask^{COMET}
Mexico reels from shooting attack in El Paso ***fear***			
negative surprise	negative surprise	fear	fear
Insane video shows Viking Sky cruise ship thrown into chaos at sea ***fear***			
negative surprise	fear	negative surprise	fear
Durant could return for Game 3 ***positive surprise***			
for game	could return for game		
Dan Fagan: Triple shooting near New Orleans School yet another sign of city's crime problem ***negative surprise***			
school yet another sign of city's crime	: triple shooting near new orleans school yet another sign of city's crime		

Table 2: Example outputs from our systems. For each example, the gold cause is highlighted in yellow and the gold emotion is given under the text; the first two examples give our models' emotion outputs; the latter two, their causes. Joined cells show that multiple models produced the same output. To make this table easier to read, "Multitask" here may refer to Multi, Multi$_{E\rightarrow C}$, or Multi$_{C\rightarrow E}$ (details on selection and results for each individual model available in appendix; most multi-task models gave similar outputs).

Metric	BERT	BERT^{COMET}	Multi	Multi$_{E\rightarrow C}$	Multi$_{C\rightarrow E}$	Multi^{COM}	Multi$^{COMET}_{E\rightarrow C}$	Multi$^{COMET}_{C\rightarrow E}$
Acc. (Gold)	38.50	38.50	38.34	39.68	38.74	39.05	38.58	40.79
Acc. (¬Gold)	23.48	23.24	22.37	21.11	22.85	21.26	22.45	20.08

Table 3: Comparison of gold accuracy and non-gold (¬gold) accuracy for our emotion classification models.

ring cases where different annotators give directly opposing labels for the same input, depending on how they interpret the headline and whose emotions they choose to focus on. For example, our development set includes the following example: *Simona Stuns Serena at Wimbledon: Game, Set and "Best Match" for Halep*. The gold adjudicated emotion label for this example is *negative surprise*, but annotators actually included multiple primary and secondary emotion labels including *joy*, *negative surprise*, *positive surprise*, *pride*, and *shame*, which can be understood as various emotions felt by the two entities participant in the event (Simona Halep and Serena Williams). For this input, COMET suggests xReact may be *happy* or *proud* and oReact may be *happy* — these reactions are likely most appropriate for tennis player Simona Halep, but not the only possible emotion that can be inferred from the headline.

Inspired by the variation in the data, we compute also models' accuracy using the human annotations that did not agree with the gold (i.e., a predicted emotion label is correct if it was suggested by a human annotator but was not part of a majority vote to be included in the gold). We denote this ¬Gold, and we compare the performance of our models with respect to Gold and ¬Gold. We present the results of this analysis in Table 3[8]. In this table, a higher ¬Gold accuracy means that the model is more likely to produce emotion labels that were not the gold but were suggested by some annotator. First of all, we note that all models have a relatively high ¬Gold accuracy (about half the magnitude of their gold accuracy); we believe this reflects the wide variety of annotations given by the annotators. We see a trade-off between the Gold and ¬Gold accuracy, and we note that generally the single-task models have higher ¬Gold accuracy and the COMET-enhanced multi-task models have higher Gold accuracy. This suggests that our language models have general knowledge about emotion already, but that applying common-sense knowledge helps pare down the space of plausible outputs to those that are most commonly selected by human annotators. Recall

[8]Note that we perform this analysis on just one of our five runs of the model, so the accuracy numbers do not exactly correspond to those in Table 1.

that this dataset was annotated by taking the most frequent of the annotator-provided emotion labels. Further, since the multi-task models have higher Gold accuracy and lower ¬Gold accuracy than the single-task models, this suggests that also predicting the cause of an emotion causes the model to narrow down the space of possible emotion labels to only those that are most common.

8 Conclusions and Future Work

We present a common-sense knowledge-enhanced multi-task framework for joint emotion detection and emotion cause tagging. Our inclusion of common-sense reasoning through COMET, combined with multi-task learning, yields performance gains on both tasks including significant gains on emotion classification. We highlight the fact that this work frames the cause extraction task as a span tagging task, allowing for the future possibility of including multiple emotion-cause pairs per input or multiple causes per emotion and allowing the cause to take on any grammatical role. Finally, we present an analysis of our dataset and models, showing that labeling emotion and its semantic roles is a hard task with annotator variability, but that common-sense knowledge helps language models focus on the most prominent emotions according to human annotators. In future work, we hope to explore ways to integrate common-sense knowledge more innately into our classifiers and ways to apply these models to other fine-grained emotion tasks such as detecting the experiencer or the target of an emotion.

Acknowledgements

We would like to thank our reviewers as well as the members of Amazon AI team for their constructive and insightful feedback.

Ethical Considerations

Our intended use for this work is as a tool to help understand emotions expressed in text. We propose that it may be useful for things like product reviews (where producers and consumers can rapidly assess reviews for aspects of their products to improve or expand), disaster relief (where those in need of help from any type of disaster can benefit if relief agents can understand what events are causing negative emotions, during and after the initial disaster), and policymaking (where constituents can benefit if policymakers can see real data about what policies are helpful or not and act in their interests). These applications do depend on the intentions of the user, and a malicious actor may certainly misuse the ability to (accurately or inaccurately) detect emotions and their causes. We do not feel it responsible to publicly list the ways in which this may happen in this paper. We also believe that regulators and operators of this technology should be aware that it is still in its nascent stages and does not represent an infallible oracle — the predictions of this and any model should be reviewed by humans in the loop, and we feel that general public awareness of the limitations and mistakes of these models may help mitigate any possible harm. If these models are inaccurate, they will output either the incorrect emotion or the incorrect cause; blindly trusting the model's predictions without examining them may lead to unfair consequences in any of the above applications (e.g., failure to help someone whose text is misclassified as positive surprise during a natural disaster or a worsened product or policy if causes are incorrectly predicted). We additionally note that in its current form, this work is intended to detect the emotions that are expressed in text (headlines), and not those of the reader.

We concede that the data used in this work consists of news headlines and may not be the most adaptable to the use cases we describe above; we caution that models trained on these data will likely require domain adaptation to perform well in other settings. Bostan et al. (2020) report that their data comes from the Media Bias Chart[9], which reports that their news sources contain a mix of political views, rated by annotators who also self-reported a mix of political views. We note that these data are all United States-based and in English. Bostan et al. (2020) do sub-select the news articles according to impact on Twitter and Reddit, which have their own user-base biases[10], typically towards young, white American men; therefore, the data is more likely to be relevant to these demographics. The language used in headlines will likely most resemble Standard American English as well, and therefore our models will be difficult to use directly on other dialects and vernaculars.

[9] https://www.adfontesmedia.com/about-the-interactive-media-bias-chart/

[10] https://www.pewresearch.org/internet/fact-sheet/social-media/

References

Muhammad Abdul-Mageed and Lyle Ungar. 2017. EmoNet: Fine-grained emotion detection with gated recurrent neural networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 718–728, Vancouver, Canada. Association for Computational Linguistics.

Cecilia Ovesdotter Alm, Dan Roth, and Richard Sproat. 2005. Emotions from text: Machine learning for text-based emotion prediction. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, page 579–586, USA. Association for Computational Linguistics.

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

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Çelikyilmaz, and Yejin Choi. 2019. COMET: commonsense transformers for automatic knowledge graph construction. 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 4762–4779. Association for Computational Linguistics.

Laura Ana Maria Bostan, Evgeny Kim, and Roman Klinger. 2020. Goodnewseveryone: A corpus of news headlines annotated with emotions, semantic roles, and reader perception. In *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, pages 1554–1566. European Language Resources Association.

Ying Chen, Wenjun Hou, Xiyao Cheng, and Shoushan Li. 2018. Joint learning for emotion classification and emotion cause detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 646–651, Brussels, Belgium. Association for Computational Linguistics.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. GoEmotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Zixiang Ding, Rui Xia, and Jianfei Yu. 2020. ECPE-2D: Emotion-cause pair extraction based on joint two-dimensional representation, interaction and prediction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3161–3170, Online. Association for Computational Linguistics.

Chuang Fan, Chaofa Yuan, Jiachen Du, Lin Gui, Min Yang, and Ruifeng Xu. 2020. Transition-based directed graph construction for emotion-cause pair extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3707–3717, Online. Association for Computational Linguistics.

D. Ghazi, Diana Inkpen, and S. Szpakowicz. 2015. Detecting emotion stimuli in emotion-bearing sentences. In *CICLing*.

Deepanway Ghosal, Navonil Majumder, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. 2020. COSMIC: COmmonSense knowledge for eMotion identification in conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2470–2481, Online. Association for Computational Linguistics.

Lin Gui, Dongyin Wu, Ruifeng Xu, Qin Lu, and Yu Zhou. 2016. Event-driven emotion cause extraction with corpus construction. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1639–1649, Austin, Texas. Association for Computational Linguistics.

Evgeny Kim and Roman Klinger. 2018. Who feels what and why? annotation of a literature corpus with semantic roles of emotions. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1345–1359, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

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

Svetlana Kiritchenko, Xiaodan Zhu, and Saif M. Mohammad. 2014. Sentiment analysis of short informal texts. *J. Artif. Int. Res.*, 50(1):723–762.

Xiang Li, Aynaz Taheri, Lifu Tu, and Kevin Gimpel. 2016. Commonsense knowledge base completion. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1445–1455, Berlin, Germany. Association for Computational Linguistics.

Saif Mohammad and Peter D. Turney. 2013. Crowdsourcing a word-emotion association lexicon. *CoRR*, abs/1308.6297.

Saif Mohammad, Xiaodan Zhu, and Joel Martin. 2014. Semantic role labeling of emotions in tweets. In *Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 32–41.

Saif M. Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. Semeval-2018 Task 1: Affect in tweets. In *Proceedings of International Workshop on Semantic Evaluation (SemEval-2018)*, New Orleans, LA, USA.

Laura Ana Maria Oberländer and Roman Klinger. 2020. Token sequence labeling vs. clause classification for English emotion stimulus detection. In *Proceedings of the Ninth Joint Conference on Lexical and Computational Semantics*, pages 58–70.

Laura Ana Maria Oberländer, Kevin Reich, and Roman Klinger. 2020. Experiencers, stimuli, or targets: Which semantic roles enable machine learning to infer the emotions? In *Proceedings of the Third Workshop on Computational Modeling of People's Opinions, Personality, and Emotion's in Social Media*, pages 119–128.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, and Rada Mihalcea. 2020a. Beneath the tip of the iceberg: Current challenges and new directions in sentiment analysis research.

Soujanya Poria, Navonil Majumder, Devamanyu Hazarika, Deepanway Ghosal, Rishabh Bhardwaj, Samson Yu Bai Jian, Romila Ghosh, Niyati Chhaya, Alexander Gelbukh, and Rada Mihalcea. 2020b. Recognizing emotion cause in conversations.

Lance A. Ramshaw and Mitchell P. Marcus. 1995. Text chunking using transformation-based learning. *CoRR*, cmp-lg/9505040.

Sara Rosenthal, Saif M. Mohammad, Preslav Nakov, Alan Ritter, Svetlana Kiritchenko, and Veselin Stoyanov. 2019. Semeval-2015 task 10: Sentiment analysis in twitter. *CoRR*, abs/1912.02387.

Irene Russo, Tommaso Caselli, Francesco Rubino, Ester Boldrini, and Patricio Martínez-Barco. 2011. EMOCause: An easy-adaptable approach to extract emotion cause contexts. In *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA 2.011)*, pages 153–160, Portland, Oregon. Association for Computational Linguistics.

Maarten Sap, Ronan LeBras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A. Smith, and Yejin Choi. 2018. ATOMIC: an atlas of machine commonsense for if-then reasoning. *CoRR*, abs/1811.00146.

Carlo Strapparava and Rada Mihalcea. 2010. Annotating and identifying emotions in text. In Giuliano Armano, Marco de Gemmis, Giovanni Semeraro, and Eloisa Vargiu, editors, *Intelligent Information Access*, volume 301, pages 21–38.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Penghui Wei, Jiahao Zhao, and Wenji Mao. 2020. Effective inter-clause modeling for end-to-end emotion-cause pair extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3171–3181, Online. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144.

Rui Xia and Zixiang Ding. 2019. Emotion-cause pair extraction: A new task to emotion analysis in texts. *CoRR*, abs/1906.01267.

Rui Xia, Mengran Zhang, and Zixiang Ding. 2019. RTHN: A rnn-transformer hierarchical network for emotion cause extraction. *CoRR*, abs/1906.01236.

Peng Xu, Andrea Madotto, Chien-Sheng Wu, Ji Ho Park, and Pascale Fung. 2018. Emo2vec: Learning generalized emotion representation by multi-task training. *CoRR*, abs/1809.04505.

Chaofa Yuan, Chuang Fan, Jianzhu Bao, and Ruifeng Xu. 2020. Emotion-cause pair extraction as sequence labeling based on a novel tagging scheme. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3568–3573, Online. Association for Computational Linguistics.

A Appendix

A.1 Hyperparameter Tuning

We include descriptions of our hyperparameter tuning setup and the selected hyperparmeters for each of our models in Table 4; we note that single-task cause models (BERT$_C$ and COMET$_C$) do not tune the `pooler`, all single-task models do not tune the lambda parameter, and all non-common-sense models do not tune `comet_relations`. The parameters selected by all of our models can be seen in Table 5, Table 6, and Table 7. All of our models are trained with minibatches of size $b = 32$.

We used Bayesian optimization as implemented by Amazon SageMaker[11] to tune these parameters, giving the learning rate a logarithmic scale and the dropout and lambda a linear one and allowing 75 iterations of parameter choice before selecting the setting with the best performance on the development set. Each individual instance of each model consisted of five different restarts with five distinct random seeds; one of these instances took approximately five minutes on a single Tesla V100 GPU, for a total of about 6.25 GPU-hours per model and thus 87.5 GPU-hours overall (since each multi-task model was trained twice: once optimized for emotion and once optimized for cause).

A.2 Model Sizes

Our models' sizes are dominated by BERT-base, which has 110 million trainable parameters (Devlin et al., 2019). We note that our trainable dense layers that interface with BERT have sizes 768×7 for emotion classification, 768×3 for cause tagging, and 1068×7 for our Multi$_{E \to C}$ models, while our emotion embedding matrix E has 300×7 trainable parameters. Our fine-tuning process does continue to tune all of BERT's parameters.

A.3 Extended Examples

We include the output of all models for our four selected examples in subsection 7.1 in Table 8, Table 9, Table 10, and Table 11.

[11] https://aws.amazon.com/sagemaker/

Parameter Name	Type	Range or Values
`pooler`	Categorical	`[cls, mean, max, attention]`
`learning rate`	Continuous	$[10^{-6}, 10^{-4}]$
`dropout`	Continuous	$[0, 0.9]$
`lambda`	Continuous	$[0.1, 0.9]$
`comet_relations`	Categorical	`[xReact, oReact, both]`

Table 4: Our hyperparameter search ranges.

Model	**Target Task**	**Parameter Name**	**Parameter Value**
BERT_E	Emotion	pooler	`cls`
		dropout	0.8999992513311351
		lr	$2.0872134970009262 \times 10^{-5}$
BERT_C	Cause	dropout	0.04011659404129298
		lr	$9.609926650689472 \times 10^{-5}$
BERT_E^{COMET}	Emotion	pooler	`cls`
		dropout	0.6467089448672897
		lr	$3.548213539029209 \times 10^{-5}$
		comet_relations	`both`
BERT_C^{COMET}	Cause	dropout	0.8806119007595122
		lr	$9.913585728926367 \times 10^{-5}$
		comet_relations	`xReact`

Table 5: The selected hyperparameter values for our single-task models.

Model	**Target Task**	**Parameter Name**	**Parameter Value**
Multi	Emotion	pooler	`mean`
		dropout	0.1438975482079587
		lr	$2.170218150294524 \times 10^{-5}$
		lambda	0.3736515054477897
	Cause	pooler	`cls`
		dropout	0.8929935089177194
		lr	$9.929740332732521 \times 10^{-5}$
		lambda	0.6103686494768474
$\text{Multi}_{E \to C}$	Emotion	pooler	`max`
		dropout	0.2511612834815036
		lr	$3.179072019077849 \times 10^{-5}$
		lambda	0.4938386162506444
	Cause	pooler	`max`
		dropout	0.763419047616446
		lr	$8.680439371509037 \times 10^{-5}$
		lambda	0.1341940851689314
$\text{Multi}_{C \to E}$	Emotion	pooler	`max`
		dropout	0.8138762283528274
		lr	$4.2586257586160994 \times 10^{-5}$
		lambda	0.8531247637209994
	Cause	pooler	`mean`
		dropout	0.6992099059226856
		lr	$9.859155309987275 \times 10^{-5}$
		lambda	0.4855821360212248

Table 6: The selected hyperparameter values for our multi-task BERT models.

Model	Target Task	Parameter Name	Parameter Value
Multi^{COMET}	Emotion	pooler	max
		dropout	0.22350077887111716
		lr	$3.137385699389837 \times 10^{-5}$
		lambda	0.7676911585403968
		comet_relations	both
	Cause	pooler	mean
		dropout	0.8891347000216091
		lr	$8.123006047625093 \times 10^{-5}$
		lambda	0.1
		comet_relations	both
Multi$^{COMET}_{E\rightarrow C}$	Emotion	pooler	mean
		dropout	0.1372637910712323
		lr	$3.0408118480380588 \times 10^{-5}$
		lambda	0.8968243966922735
		comet_relations	both
	Cause	pooler	max
		dropout	0.5319636087561394
		lr	$7.581334242472624 \times 10^{-5}$
		lambda	0.10896064677810494
		comet_relations	both
Multi$^{COMET}_{C\rightarrow E}$	Emotion	pooler	cls
		dropout	0.7359624181177503
		lr	$1.9853909769532754 \times 10^{-5}$
		lambda	0.7947522633173147
		comet_relations	both
	Cause	pooler	max
		dropout	0.01896406469706125
		lr	$8.360862387915605 \times 10^{-5}$
		lambda	0.14588492191321054
		comet_relations	oReact

Table 7: The selected hyperparameter values for our multi-task COMET models.

Mexico reels from shooting attack in El Paso

fear

Model	Output
BERT	negative surprise
BERT^{COMET}	fear
Multi	negative surprise
Multi$_{C\rightarrow E}$	negative surprise
Multi$_{E\rightarrow C}$	negative surprise
Multi^{COMET}	fear
Multi$^{COMET}_{C\rightarrow E}$	fear
Multi$^{COMET}_{E\rightarrow C}$	fear

Table 8: Full model outputs for our first provided example.

Insane video shows Viking Sky cruise ship thrown into chaos at sea

fear

Model	Output
BERT	negative surprise
BERT^{COMET}	negative surprise
Multi	fear
Multi$_{C\rightarrow E}$	negative surprise
Multi$_{E\rightarrow C}$	fear
Multi^{COMET}	fear
Multi$^{COMET}_{C\rightarrow E}$	fear
Multi$^{COMET}_{E\rightarrow C}$	negative surprise

Table 9: Full model outputs for our second provided example.

Durant could return for Game 3

positive surprise

Model	Output
BERT	for game
BERT^{COMET}	could return for game
Multi	could return for game
Multi$_{C\rightarrow E}$	could return for game
Multi$_{E\rightarrow C}$	could return for game
Multi^{COMET}	could return for game
Multi$^{COMET}_{C\rightarrow E}$	could return for game
Multi$^{COMET}_{E\rightarrow C}$	could return for game

Table 10: Full model outputs for our third provided example.

Dan Fagan: Triple shooting near New Orleans School yet another sign of city's crime problem

negative surprise

Model	Output
BERT	school yet another sign of city's crime
BERT^{COMET}	: triple shooting near new orleans school yet another sign of city's crime
Multi	shooting near new orleans school yet another sign of city's crime
Multi$_{C\rightarrow E}$	: triple shooting near new orleans school yet another sign of city's crime
Multi$_{E\rightarrow C}$	: triple shooting near new orleans school yet another sign of city's crime
Multi^{COMET}	: triple shooting near new orleans school yet another sign of city's crime
Multi$^{COMET}_{C\rightarrow E}$	: triple shooting near new orleans school yet another sign of city's crime
Multi$^{COMET}_{E\rightarrow C}$	: triple shooting near new orleans school yet another sign of city's crime

Table 11: Full model outputs for our fourth provided example.