Could Augmentation Improve a Conversational Emotion-Cause Pair Extraction Model Toward Low-Resource, Non-Conversational Domains?

Devin Suy
UC Berkeley MIDS
devinsuy@berkeley.edu

Kevin Hoang
UC Berkeley MIDS
khoang97@berkeley.edu

Abstract

Emotion-Cause Pair Extraction (ECPE) in NLP aims to predict emotion labels for a given text along with words or spans that denote the cause for such emotion labels. Recent works on ECPE have primarily focused on conversational context settings involving two or more speakers, often with foreign corpora. In this work, we apply a high-performing, transformer based ECPE model in a conversational domain toward two non-conversational English corpora at 5,000 records in size.

Our experiments investigate whether augmenting training data to 20,000 records via token manipulations and adding data from different domains would increase test F1 scores of both the emotion label and cause span prediction tasks, as compared to a baseline of training on unaugmented data. Our findings partially lean towards the positive: while we identified a slight improvement in classification outcomes, we also recognized a trade-off between improvements in emotion classification and span extraction. Notably, our analysis highlights the inherently subjective nature of both tasks.

1 Introduction

Emotion-Cause Pair Extraction (ECPE) is a task proposed by Xia and Ding (2019), which, given a text input, aims to predict two outputs: first, an emotion label, and second, a sequence of words or spans associated with the cause for that emotion label. This task grew from the simpler Emotion Cause Extraction (ECE) task (Lee et al., 2010), which purely involved extracting cause phrases and bore limitations of requiring that emotion labels be annotated on a piece of text beforehand.

Investigating the ECPE task holds substantial significance due to the variety of practical applications it has. One such application is its utility in product review analysis where it can aid in determining the factors that shape customer satisfaction with a product. For instance, through ECPE analysis of customer reviews, companies can gain insight into how their customers link particular aspects of their products with distinct positive or negative emotions. This stands in contrast to conventional sentiment analysis, as ECPE allows for a focused examination of the specific textual aspects that trigger an emotion, rather than providing a general sentiment associated with an entire text.

Moreover, the study of ECPE also has a key role in advancing conversational AI, particularly seen in the development of chatbots capable of generating empathetic text responses for users. This is notably evident in applications such as mental health support, where understanding the emotional causes of user concerns is vital in generating meaningful and supportive responses.

2 Related Work

Exploring the expression of emotion in text is a key area of investigation in natural language processing. Early approaches focused on recognizing sentiments such as positivity, negativity, or neutrality. Over time the scope of this field has evolved to encompass a more intricate analysis of the subtleties between different emotions on this spectrum. Among these developments emerged the novel Emotion-Cause Pair Extraction (ECPE) task introduced by Xia and Ding (2019), who first approached the task via an LTSM-based model. This task directed focus to the more challenging aspects of both emotion recognition, as well as comprehension of the inherent cause.

Recent endeavors within the ECPE domain have predominantly gravitated toward analyzing conversational texts. In this context, the Recognizing Emotion Cause in CONversations (RECCON) corpus (Poria et al., 2021) has become widely favored for this task. A contemporary contribution to the RECCON-based ECPE investigation is the MuTEC architecture, put forth by Bhat and Modi (2022). Given the evolving state of this field, there exists unexplored potential in adapting existing ECPE architectures to non-conversational documents, such as within self-contained texts. This new area presents a promising opportunity for further exploration and advancement.

3 Methods

3.1 Task

The objective of ECPE is to recognize the expressed emotion in a given text, while also identifying the specific causes or triggers associated with those emotions. The output of this task takes the form of span extraction, where relevant segments of text containing emotion-cause pairs are extracted, alongside the classified emotion.

To accomplish this task, we adopt the MuTEC[1] architecture proposed by Bhat and Modi (2022). MuTEC is a specialized architecture designed to confront the tasks defined by the RECCON (Poria et al., 2021). Our investigation aims to explore the performance of the MuTEC architecture across diverse corpora, specifically in the task of ECPE. When compared to RECCON, the primary distinction of these new datasets lies in the self-contained and non-conversational nature of the text they encompass.

Our proposed datasets for exploring the non-conversational domains in ECPE are relatively low resource in the number of annotated examples. As such, another core aspect of our investigation lies in exploring a variety of data augmentation techniques. Through these methods, we increase the overall size of our datasets and compare their impacts on the performance metrics with respect to the ECPE classification task.

3.2 Data

We leverage two recent non-conversational English corpora related to ECPE, with both emotion labels and causes annotated:

- EmoCause - Proposed by Kim et al. (2021), a collection of 4,613 situations from the EmpatheticDialogues dataset by Rashkin et al. (2019).

- GoodNewsEveryone (GNE) - Proposed by Bostan et al. (2020), a collection of 5,000 news headlines from sources publicly available as RSS feeds from the Media Bias Chart (Otero, 2023).

3.3 Data Preprocessing

As our task is to extract emotion cause *spans*, we first note that EmoCause annotates emotion cause words for each input text, rather than spans. These emotion words can occur at different positions within the input text, so we define the ground truth cause span for each sample in EmoCause as the span of words beginning with the first annotated cause word and ending with the last annotated cause word. We report on span prediction results for EmoCause in adherence to this definition.

GNE provides contiguous ground truth spans for emotion cause. Particularly, GNE distinguishes "gold" spans which are adjudicated from candidate spans proposed by annotators. However, 201 of the 5,000 records in GNE have missing entries for gold spans, and upon inspection, we noted that the candidate spans for many of these records are not meaningful enough to associate cause to the annotated emotion. Hence, we drop these 201 records and have 4,799 remaining records for GNE.

We generate labels for cause span start and end indices by searching tokenized inputs for tokenized cause spans via the SpanBERT tokenizer, as MuTEC specifically predicts cause spans through their start and end indices within SpanBERT tokenized inputs. When multiple gold spans are given per sample in GNE, we only take the first given.

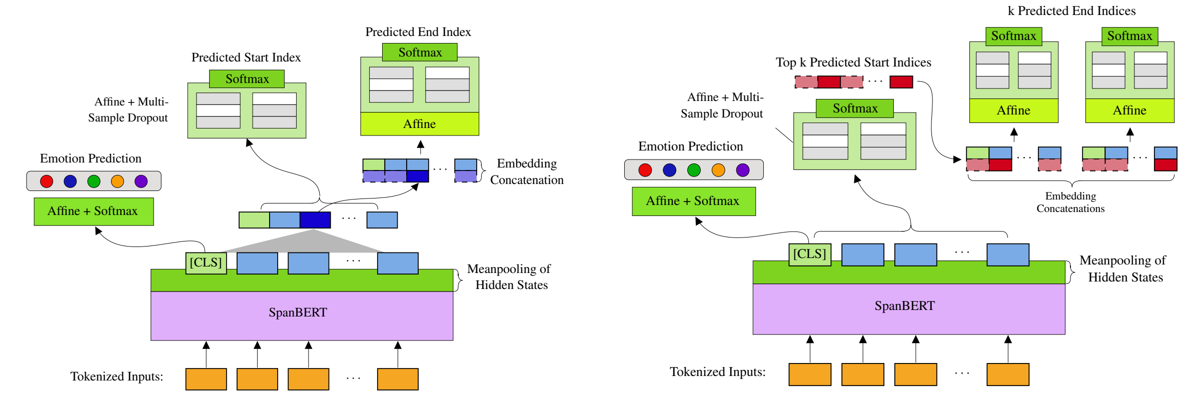
EmoCause provides a test set of 838 out of 4,613 records, while GNE does not have a predetermined train/test split. We thus reserve 960 random records from GNE for testing. We then partition the 3,775 remaining records for EmoCause and the 3,839 records for GNE into train and validation sets via an 80-20 split.

3.4 Model

We summarize the distinctive aspects of MuTEC's architecture for clarity. The inputs to the model are standalone texts[2] tokenized via the SpanBERT tokenizer, with padding of size $p = 124$. The tokenized inputs are passed through the SpanBERT transformer, and we mean pool each token's embeddings throughout all 12 hidden states of the transformer.

[1] Bhat and Modi propose three variants of MuTEC that encapsulate different learning tasks in a conversational setting, but only MuTEC$_{\text{CSE}}$ is relevant toward purely the ECPE task. Each time we refer to MuTEC in our paper, we in actuality refer to MuTEC$_{\text{CSE}}$.

[2] In Bhat and Modi's study, the inputs would be concatenations of dialogue utterances, and in our study, we input EmoCause situations and GNE headlines.



(a) Training architecture: start indices are given (dark blue block). Meanpooled embeddings in the start index are concatenated to those in each token to predict end index.

(b) Inference architecture: take the top k predicted start indices (dark red blocks), output k start-end index pairs. Predict the pair with highest joint probability.

Figure 1: MuTEC$_{\text{CSE}}$ architecture.

For the emotion label extraction, the embeddings of the meanpooled CLS tokens are fed through an affine layer with softmax activation to predict the emotion label.

The prediction methods of the start and index positions of cause spans in MuTEC differ in training versus inference/evaluation.

Training: For the start index prediction, the meanpooled hidden states are fed through a multi-sample dropout (MSD) layer[3] of output size 1 to produce a logit for each token. We apply softmax to the logits to predict start index probabilities.

For the end index prediction, we are given start positions in training. For each training sample, we take the meanpooled hidden state of the token in the given start position label and concatenate that state to the meanpooled hidden state of each input token, outputting a $(\text{batch size}) \times p \times 1536$ tensor. This tensor is passed through an affine layer with $\tanh$ activation to output a $(\text{batch size}) \times p \times 768$ tensor, which is passed similarly through an MSD layer of output size 1 and then a softmax layer to output end index probabilities.

Inference: Here, we predict the top k start index probabilities as in training and extract the k respective start indices. We predict k end probabilities via the same method in training, one for each start index within the top k start index probabilities, via concatenating the meanpooled hidden state for the token in that start index. k is referred to by Bhat and Modi as the "beam size," which we set to $k = 4$ in our study. The model will predict the start-end index pair corresponding to the highest multiplied/joint probability.

3.4.1 Loss Functions

The model tracks three categorical cross-entropy loss functions that correspond respectively to the predicted emotion label, the predicted cause span start index, and the predicted cause span end index. The overall loss function is a weighted average of these three cross-entropy functions, with a weight of 0.5 for the emotion prediction, and weights of 0.25 for the start and end index predictions. All parameters are trainable, including those in SpanBERT, to minimize overall loss.

3.5 Experimental Hypothesis and Procedures

Put directly, our experimental hypothesis is whether, for both EmoCause and GNE, training MuTEC on augmented training sets to ~20,000 samples will increase weighted average F1 scores of the emotion classification task and average F1 scores for the span prediction task on each test dataset by at least 3 points, compared to a baseline of training MuTEC on unaugmented training data at size ~3,000.

The weighted F1 score takes its usual definition in classification problems, but we define average F1 scores for the span prediction task as follows, in line with Bhat and Modi's paper on MuTEC:

> Given an input text sample, let M be the number of matching token indices

[3] Multisample dropout is a recently developed technique that copies an affine layer and its weights multiple times, applies distinct dropout masks across the copies, and averages the outputs across the copies back together for improving model generalizability (Inoue, 2020).

between the predicted and ground truth spans. The precision P and recall R are the ratios of M to the total number of tokens in the predicted and ground truth spans, respectively. Then,

$$F1_{sample} = \frac{2 * P * R}{P + R}.$$

The average F1 score is the average of $F1_{sample}$ across all samples.

The augmentations we describe consist of the following operations, the first three of which we take from research by Wei and Zou (2019).

1. **Synonym Replacement (SR):** Randomly selecting a whole number n from 1 to 5 inclusive, and swapping n random tokens in the input text with their closest synonym.

2. **Random Insertion (RI):** Finding a random token in the input text, inserting the closest synonym of that token in a random position in the text, and repeating a random number of times, ranging from 1 to 1/8th the total number of tokens in the text.

3. **Random Deletion (RD):** Randomly removing each token in the input text with a probability $p = 0.2$.

4. **Appending the training sets** of EmoCause to that of GNE and vice-versa, and remapping emotion labels.

All random selections described in SR and RI are done uniformly. The choices of how many tokens to swap/insert along with the deletion probability are consistent with Wei and Zou's recommendation to alter no more than 20% of the input text. We define a token's closest synonym as follows:

> Let SpanBERT(t) be the embedding vector for token t in the last hidden layer of SpanBERT, upon solely inputting t. t's closest synonym t_c is defined such that $t \neq t_c$ and $u := t_c$ maximizes the cosine similarity between SpanBERT(t) and SpanBERT(u) with respect to tokens u in the SpanBERT vocabulary.

We apply three augmentation experiments to each training dataset, each augmenting the number of samples to nearly 20,000:

1. **SR only,**

2. **SR, RI, and RD.**

3. **Appending training sets, then SR, RI, and RD.**

The choice to append to 20,000 follows from Wei and Zou's findings that for training sets of 5,000 samples, augmenting to more than 4 augmented inputs per original input results in diminishing returns for classification accuracy.

In our exploration of EmoCause and GNE, we noted minor class imbalance in the former and major class imbalance in the latter. Hence, we oversample minority classes in each augmentation experiment so that each class has the same number of samples in the augmented data (Chawla et al., 2002).

4 Results and Discussion

4.1 Overall Remarks

For both datasets, we found notable boosts (past our 3 point criteria) in label prediction weighted F1 scores via augmentation, though results overall do not appear strong. With the exception of SR + RI + RD on EmoCause, augmentation did not appear to produce even slight gains on span prediction F1 scores. However, MuTEC performs strongly on the span prediction task overall. The increases we observed, though not dramatic, are at least comparable in magnitude to classification accuracy increases on Wei and Zou's original paper.

4.2 Emotion Label Prediction

For the emotion label prediction task, we first note class imbalances and low volume of given data among many classes. In EmoCause, there are 32 classes with a small imbalance: the majority class being "surprise" at 187 samples and the minority class "faithful" at 69 in our training set. In GNE, the labels take 15 classes with a major imbalance: the majority class of “negative surprise” at 871 samples and the minority class of “love including like” at 58 in our training set.

For GNE, our choice to oversample and add noise to inputs via the SR only experiment yielded the best increase in label prediction F1 score, by 9.41 points compared to the baseline. The oversampling allowed MuTEC to better discern minority classes and yield this improvement in emotion classification.

Dataset	Unaugmented		SR Only		SR, RI, RD		Append, SR, RI, RD	
	$\mathrm{F1}_{label}$	$\mathrm{F1}_{span}$	$\mathrm{F1}_{label}$	$\mathrm{F1}_{span}$	$\mathrm{F1}_{label}$	$\mathrm{F1}_{span}$	$\mathrm{F1}_{label}$	$\mathrm{F1}_{span}$
EmoCause	46.09	59.35	47.65	57.88	48.80	**60.50**	**49.76**	58.49
GNE	15.59	**76.16**	**25.00**	75.71	23.81	76.06	20.54	71.67

Table 1: Performance of MuTEC on respective test datasets, with a baseline of training on unaugmented data versus experiments of training on augmented data.

For Emocause, our best label prediction result was via the append operation plus SR, RI, and RD, a 3.67 point boost from baseline. We deduce that adding GNE headlines and adding augmentation diversified the inputs to MuTEC. EmoCause situations mostly use common, general language, so MuTEC's learning on the augmented and diversified text helped its generalizability in predicting emotion labels of the common-sounding test situations in EmoCause.

Even then, F1 scores of less than 50 appear weak. We sought further insight into these low scores via confusion matrices for our best experiments, finding that MuTEC tends to confuse similar negative emotions (anger, annoyance) and positive emotions (joy, positive anticipation/excitement). Further inspection of specific headlines illustrates that in many cases, incorrect predictions are at least reasonable: e.g., in GNE the headline "AFL-CIO president Opposes Green New Deal" has "disgust" as its true label, but MuTEC predicted "anger".

We highlight the relative subjectivity in annotating similar emotions, as well as the volume of data being too little for MuTEC to truly learn distinctions between similar emotions.

Dataset	True	Predicted	#
Emo-Cause	Sentimental	Nostalgic	11
	Angry	Annoyed	10
	Grateful	Joyful	10
	Sad	Devastated	9
	Terrified	Afraid	9
GNE	N-Surprise	N-Anticipation	38
	Joy	P-Surprise	22
	N-Surprise	Fear	17
	N-Surprise	Annoyance	17
	Annoyance	N-Anticipation	17

Table 2: Top 5 Class Confusions in Each Test Set by Quantity. (N - Negative, P - Positive)

4.3 Augmentation on Span Prediction

In the GNE test set, the SR, RI, and RD operations did not yield a boost in span F1 score over our baseline. We observe that GNE headlines have an abundance of proper nouns and particular jargon split by SpanBERT tokenization. Consequently, these operations tend to over-augment the input headline to the point that the augmented headlines lose coherence, which appears to hinder MuTEC's generalizability to test headlines. An example of this limitation follows.

Original: "Harvard study finds trigger warnings do more harm than good"

Augmented: "Harva**fred (SR)** study finds trigger **official (RI)** warnings **(RD)** more harm **(RD)** good"

However, for EmoCause, the SR, RI, and RD experiment produced our best result, a 1.15 point increase in span F1 score compared to the baseline. Our observations about EmoCause situations containing common language and avoiding specific jargon applies to these findings also. The augmentations were sufficient to add positional diversity to cause spans, but not disruptive enough from the coherence of the simpler, common-sounding nature of training texts to reduce span F1 scores.

For each dataset, appending the other dataset and remapping classes yielded decreases in span F1 scores. Here we note semantic and structural differences between input texts/cause spans of the two datasets, that we conclude hinders MuTEC's predictive capabilities toward unseen test spans of a particular dataset under this experiment.

1. GNE spans take up more of the text proportionally compared to EmoCause spans.

2. GNE spans also tend to run to the end of the headline. EmoCause spans have more variation in their location within situations.

3. EmoCause cause spans explain the narrator's emotion while experiencing their situation, while GNE spans explain secondhand readers' emotions in reaction to the headline.

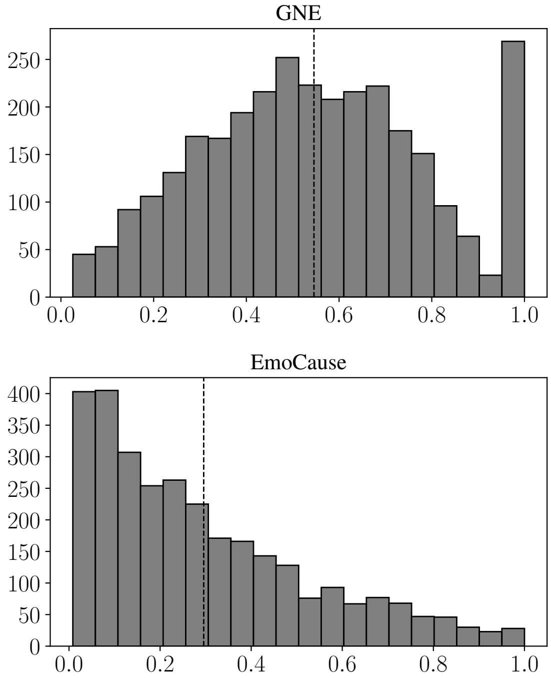


Figure 2: Histograms of Input Counts in Training Data, by % of Input in Cause Span, with Average Lines.

4.4 Span Prediction Examination

We examine the results of span prediction outputs to assess their sensibility and provide further context to the reported F1 scores.

For EmoCause, we observe sensible span predictions even when the ground truth spans differ. The first sample below illustrates the limitation in our defining ground truth spans, given emotion cause *words*, in reasonably evaluating the quality of predicted spans.

	Text
Input	"I had a surprise visit from my mother. She came into town and took us out for dinner."
Pred.	"Surprise visit from my mother"
Actual	"Mother. She came into town"
Input	"When I saw a guy on a bike doing a wheelie."
Pred.	"Guy on a bike doing a wheelie"
Actual	"Guy on a bike doing a wheelie"

Table 3: Sample Test Span Predictions - EmoCause, SR + RI + RD

For GNE, cause spans tend to run to the end of the input headline, and we observe that MuTEC has learned this pattern during training. Though there is variation as to the predicted start indices, the predicted test spans will almost always run to the end of the headline. Still, the spans are quite reasonable.

	Text
Input	"Judi Dench Leads Support for the Mail's Dementia Care Campaign"
Pred.	"the Mail's Dementia Care Campaign"
Actual	"Leads Support"
Input	"Lawyer Alleges Ecuador Spread Lies About WikiLeaks Founder"
Pred.	"Ecuador Spread Lies About WikiLeaks Founder"
Actual	"Spread Lies About WikiLeaks Founder"

Table 4: Sample Test Span Predictions - GNE, SR Only

5 Conclusion and Future Work

In this study, we attempted to train MuTEC, a high-performing, transformer-based ECPE model in a conversational domain, toward two non-conversational English language ECPE corpora that have seen little to no study in previous literature. We attempt to augment training data in these corpora to improve model performance. We find that our augmentation tactics were able to improve MuTEC in both emotion prediction and cause prediction toward EmoCause, a dataset with more general and common-sounding language. We only observed improvements toward emotion prediction toward a dataset with specialized language and terminology in GNE.

In motivating future research toward low-resource, non-conversational domains for the ECPE task, we emphasize the importance of having corpora with enough volume and clarity to better distinguish between similar emotions such as joy and happiness. We highlight that a span prediction model may learn positional patterns in input texts, propagating to similar patterns in test predictions. Also, we caution readers against haphazardly using data from different domains to train an ECPE model and urge readers to be cognizant of semantic and syntactic features of texts unique to particular domains prior to determining training data.

In a broader perspective, we also observe large potential to use other pretrained models such as T5 (Raffel et al., 2020), to approach the ECPE task via a question answering/multiple choice format, as well as other learning frameworks such as few-shot learning (Beltagy et al., 2022) to build stronger models on data with many classes and limited samples. We hope that the incremental results we observed in this study via augmentation will be an effective platform and early step to explore other methods that can achieve even stronger performance on ECPE.

References

Iz Beltagy, Arman Cohan, Robert Logan IV, Sewon Min, and Sameer Singh. 2022. Zero- and few-shot NLP with pretrained language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 32–37, Dublin, Ireland. Association for Computational Linguistics.

Ashwani Bhat and Ashutosh Modi. 2022. Multi-task learning framework for extracting emotion cause span and entailment in conversations.

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 Twelfth Language Resources and Evaluation Conference*, pages 1554–1566, Marseille, France. European Language Resources Association.

Nitesh Chawla, Kevin Bowyer, Lawrence Hall, and W. Philip Kegelmeyer. 2002. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357.

Hiroshi Inoue. 2020. Multi-sample dropout for accelerated training and better generalization.

Hyunwoo Kim, Byeongchang Kim, and Gunhee Kim. 2021. Perspective-taking and pragmatics for generating empathetic responses focused on emotion causes. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2227–2240, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sophia Yat Mei Lee, Ying Chen, and Chu-Ren Huang. 2010. A text-driven rule-based system for emotion cause detection. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 45–53, Los Angeles, CA. Association for Computational Linguistics.

Vanessa Otero. 2023. The media bias chart. `https://adfontesmedia.com/gallery/`. Accessed: 2023-08-04.

Soujanya Poria, Navonil Majumder, Devamanyu Hazarika, Deepanway Ghosal, Rishabh Bhardwaj, Samson Yu Bai Jian, Pengfei Hong, Romila Ghosh, Abhinaba Roy, Niyati Chhaya, Alexander Gelbukh, and Rada Mihalcea. 2021. Recognizing emotion cause in conversations.

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.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381, Florence, Italy. Association for Computational Linguistics.

Jason Wei and Kai Zou. 2019. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, page 6382–6388, Hong Kong, China. Association for Computational Linguistics.

Rui Xia and Zixiang Ding. 2019. Emotion-cause pair extraction: A new task to emotion analysis in texts. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1003–1012, Florence, Italy. Association for Computational Linguistics.

A Model Implementation Details

In our study, we applied the following hyperparameters to MuTEC not mentioned in the main paper.

- Variable epochs.
- Batch size 24 for EmoCause, 32 for GNE.
- 0.3 dropout ratio on the Multi-Sample Dropout layers.
- 5 multisample dropout samples.
- 0.00005 learning rate.
- Adam optimizer.

During our initial runs of MuTEC on unaugmented EmoCause, MuTEC failed to learn and produce reasonable test predictions after a low number of epochs. We thus end training after observing a decrease in validation F1 scores from one epoch to the next: this occurred after epoch 15 for unaugmented EmoCause, 4 for unaugmented GNE, and 2-5 for all other experiments. In addition, upon encountering memory restrictions for batch size 32 on EmoCause, we needed to drop to 24.

B Naive Baselines

In addition to studying how MuTEC and augmentations would improve model performance, we stepped back and considered whether our experiments would at least beat naive baselines. For label prediction, this would be predicting the majority class, and for span prediction, this would be predicting all tokens in the input text.

Dataset	$F1_{label}$	$F1_{span}$
EmoCause	0.57	61.01
GNE	5.21	62.08

Table 5: Naive Baseline performance on EmoCause and GNE Test Sets.

In GNE, MuTEC plus the experiments yield significant improvements over naive baselines. In EmoCause, we call out the 61.01 span prediction F1 score, higher than in all our experiments. We found that the given test set was not representative of the overall data in terms of span length in comparison to input length. This is another limitation of having to define cause spans from the annotated cause words and explains our unfortunate result.

For the sake of maintaining the integrity of our study, we keep our test set as is. We note that with a more representative test set, we would expect MuTEC to beat a prediction of the entire span, for the model was able to learn specific start and end index patterns as evidenced by the samples in our main paper. In fact, our EmoCause validation set had a span F1 score of ~39 points for this baseline.

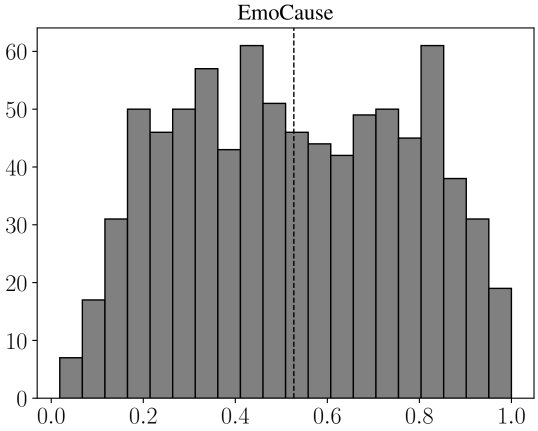


Figure 3: Histogram of Input Counts in EmoCause Test Data, by % of Input in Cause Span, with Average Line.

C Additional Discussion on Classes

Following on our remark about the subjectivity of emotion labels and the tendency of MuTEC to confuse similar emotions, as a supplemental investigation, we ran MuTEC on data with less or combined emotion labels.

For example, we combined test predictions and labels in similar classes on EmoCause, reducing the number of distinct labels to 10: this yielded consistent performance across classes and an emotion prediction weighted F1 score of 69.47.

Label	Precision	Recall	F1	Support
surprised	61.11	84.62	70.97	39
caring	67.81	72.26	69.96	137
angry	71.95	64.13	67.82	92
sad	70.21	71.74	70.97	46
afraid	72.73	52.46	60.95	61
confident	80.95	60.71	69.39	84
joyful	70.66	74.21	72.39	159
anticipating	67.89	74.75	71.15	99
sentimental	81.13	81.13	81.13	53
embarrassed	56.16	60.29	58.16	68
weighted avg.	70.21	69.57	69.47	838

Table 6: Classification Report - EmoCause with Combined Classes, SR + RI + RD

The RECCON dataset (Poria et al., 2021) used by Bhat and Modi (2022) has only 6 emotion classes, though a major imbalance toward the "happiness" label. Our label F1 score on RECCON's test set was at 74.91 after training for just 1 epoch.[4] We observed a span F1 score of 74.53, and both scores are comparable to Bhat and Modi's results.

Label	Precision	Recall	F1	Support
anger	47.11	56.62	51.43	302
disgust	0	0	0	54
fear	0	0	0	42
happiness	93.37	93.85	93.61	1,155
sadness	71.93	41.62	52.73	197
surprise	42.58	75.69	54.50	144
weighted avg.	75.17	76.35	74.91	1,894

Table 7: Classification Report - Unaugmented RECCON

Lastly, we provide the class remappings we used for appending datasets.

GNE Label	EmoCause Label Mapped To
anger	angry
annoyance	annoyed
disgust	disgusted
fear	afraid
guilt	guilty
joy	joyful
love including like	content
neg. anticipation	apprehensive
neg. surprise	disappointed
pos. anticipation	excited
pos. surprise	impressed
pride	proud
sadness	sad
shame	ashamed
trust	trusting

Table 8: Class Mappings - Appending GNE to Emocause

EmoCause Label(s)	GNE Label Mapped To
angry, jealous, furious	anger
annoyed	annoyance
disgusted	disgust
afraid, terrified	fear
guilty	guilt
joyful	joy
sentimental, content, nostalgic, caring, grateful, faithful	love including like
anxious, apprehensive	neg. anticipation
embarrassed, disappointed, devastated	neg. surprise
excited, prepared, anticipating, hopeful	pos. anticipation
surprised, impressed	pos. surprise
proud, confident	pride
sad, lonely	sadness
ashamed	shame
trusting	trust

Table 9: Class Mappings - Appending EmoCause to GNE

[4] We use the 1st fold of RECCON in Bhat and Modi's study, along only positive samples such that the candidate cause utterance in the dialogue contains the cause span. We remove negative samples whose candidate utterances do not contain cause spans.