UIC NLP GRADS at SemEval-2024 Task 3: Two-Step Disjoint Modeling for Emotion-Cause Pair Extraction

Sharad Chandakacherla[*], **Vaibhav Bhargava**[*], and **Natalie Parde**
Department of Computer Science
University of Illinois at Chicago, USA
{schand65, vbharg4, parde}@uic.edu

Abstract

Disentangling underlying factors contributing to the expression of emotion in multimodal data is challenging but may accelerate progress toward many real-world applications. In this paper we describe our approach for solving SemEval-2024 Task #3, Sub-Task #1, focused on identifying utterance-level emotions and their causes using the text available from the multimodal *F.R.I.E.N.D.S.* television series dataset. We propose to disjointly model emotion detection and causal span detection, borrowing a paradigm popular in question answering (QA) to train our model. Through our experiments we find that (a) contextual utterances before and after the target utterance play a crucial role in emotion classification; and (b) once the emotion is established, detecting the causal spans resulting in that emotion using our QA-based technique yields promising results.

1 Introduction

The task of **e**motion **c**ause **a**nalysis in **c**onversations (Wang et al., 2023, ECAC) aims to decipher the expression of human emotion in conversational data, either through unimodal (text-only) or multimodal (e.g., with the addition of video and/or audio) information. On a fundamental level, this is a complex two-part problem: emotion must be identified for a given utterance, and the span of dialogue causing that emotion must subsequently be recognized.[1] SemEval-2024 Task #3 (Wang et al., 2024) was organized around solving this problem, broken into two subtasks varying in the data allowed to build the solution; in Sub-Task #1, identification of emotion cause was limited to the use of only text information. We address Sub-Task #1 in this paper.

We pursued two strategies in our approach toward solving the task. First, we trained a question answering (QA) model (Rajpurkar et al., 2018) to extract causal spans given the reference emotions for non-neutral utterances within the training set. In doing so, we achieved comparable results to those reported by Wang et al. (2023) and Poria et al. (2021), the latter of which is a popular benchmark for this task. Next, we devised a two-step disjoint model that separately learns to classify emotion and extract causal spans during training. During inference we (1) run the emotion classifier, enriching the test set with emotion labels; and (2) run inference on the QA model to extract the causal spans. Our approach achieved third place according to the primary task metric (a weighted-average proportional F_1) and 2nd place on the secondary metric (weighted-average strict F_1; see §A.2 for results on all relevant task metrics). We elaborate on our findings in the remainder of this paper.[2]

2 Background

2.1 Task Description

Given a conversation with a sequence of n emotional utterances $u \in \{u_1, ..., u_n\}$, the twin goals in SemEval Task #3 are to identify (a) the emotion label $e_i \in \{$HAPPINESS, SADNESS, DISGUST, FEAR, SURPRISE, ANGER, NEUTRAL$\}$; and (b) for emotions other than those identified as NEUTRAL, the corresponding span of text c_i that caused u_i to be assigned label e_i.

Input and Output. Each input u_i is a sequence of text. This text is matched with video and audio in the dataset, although for Sub-Task #1 only the text is used for learning and inference. Output for each u_i is a categorical label e_i in the label space defined previously, and a sequence of text c_i signifying the reason why e_i was assigned to u_i.

[*]Authors contributed equally.

[1]*Neutral* utterances have no corresponding causal spans.

[2]Our source code is publicly available at: https://github.com/sharadchandakacherla/MultiModalEmotionCauseAnalysis/tree/main/submission.

Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), pages 1373–1379

Attribute	Frequency
# Conversations	1374
# Utterances	13619
anger	11.85%
disgust	3.03%
fear	2.74%
joy	16.90%
sadness	8.42%
surprise	13.51%
neutral	43.53%

Table 1: Dataset statistics. *# Conversations* and *# Utterances* are raw frequency counts, whereas all emotion categories are percentages of total utterances.

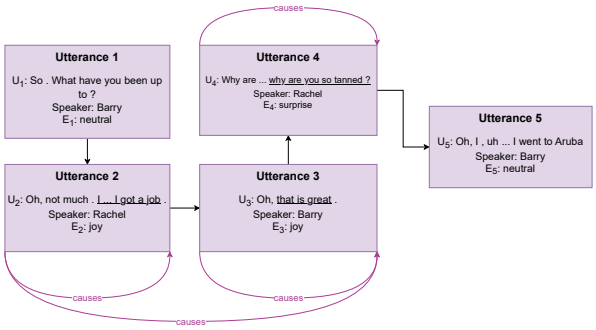

Figure 1: An example conversation from the dataset.

Dataset. All Task #3 entries were trained and evaluated using Wang et al. (2023)'s multi-modal conversational emotion cause dataset (ECF). ECF is an English-language dataset sourced from transcripts, audio, and video clips from the popular television sitcom *F.R.I.E.N.D.S*; the series comprises daily informal conversations involving a cast of six friends living in New York City. Conversations are segmented into individual speaker utterances, referred to as "emotional utterances." Causal spans are linked to emotional utterances in the dataset, and annotators could source them from any utterance in the given conversation. Dataset statistics, including the distribution of emotion labels across utterances in ECF, are presented in Table 1. Sample inputs to the emotion classification model and the causal span extractor are shown in Figure 1.

2.2 Related Work

ECAC has been studied previously to some extent in both disjoint and joint training settings. ECE (Gui et al., 2016) introduced a dataset with emotion causes extracted from a Chinese news article corpus; the language in this dataset is formal and in passive voice. Instances place focus on both clause-level (to capture emotion) and phrase-level (to capture boundaries) annotations. Building on this, ECPE (Xia and Ding, 2019) proposed a joint training model to extract emotion and corresponding causal spans, using word2vec embeddings (Mikolov et al., 2013) pre-trained on a corpus extracted from a Chinese micro-blogging website. They used a two-step process to address the emotion-cause pair extraction task, performing emotion extraction and cause extraction first, followed by emotion-cause pairing and filtering using a hierarchical Bi-LSTM model.

Poria et al. (2019) introduced a novel multimodal, multi-party conversational dataset for emotion recognition in conversations (MELD). Wang et al. (2023) makes use of MELD, and created another corpus of emotional utterances paired with their causes; this corpus also serves as the dataset for our task. Wang et al. (2023) establish baselines for the Multimodal Emotion-Cause Pair Extraction in Conversations (MC-ECPE) task using the same guidelines described in ECPE. The authors of RECCON (Poria et al., 2021) solve the task of recognizing emotion cause in conversations using causal span extraction and causal emotion entailment with transformer-based models. However, they employ their methods on IEMOCAP (Busso et al., 2008) which is a dyadic dataset and DailyDialog (Li et al., 2017) which consists of manually written dialogues focusing on physically-situated topics. Other prior work has used formal conversation datasets or reported speech where emotions are often expressed explicitly (Gui et al., 2016). Performing emotion classification and causal span extraction using a QA-based paradigm for an informal conversational setting is a novel approach to link emotion causes to implicitly expressed emotion.

3 System Overview

We model the task to maximize the probability of finding emotion-cause pairs, (e_i, c_i), for the given conversation context x. We disjointly train models with parameters θ and ϕ to estimate the emotion e_i from x and the causal span c_i from (e_i, x), respectively. We approximate x to the prompts x_e and x_c to provide the appropriate contextual information to our models, as shown in Equation 1.

$$\begin{aligned} P_{\theta,\phi}(e_i, c_i|x) &= P_\theta(e_i|x)P_\phi(c_i|e_i, x) \\ &\approx P_\theta(e_i|x_e)P_\phi(c_i|e_i, x_c) \end{aligned} \quad (1)$$

Our approach is a two-step process, by which we first identify e_i for the given utterance u_i in a

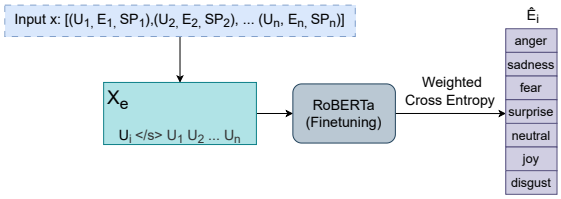

Figure 2: Training the *emotion classifier*.

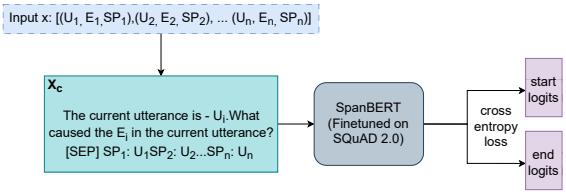

Figure 3: Training the *emotion causal classifier*.

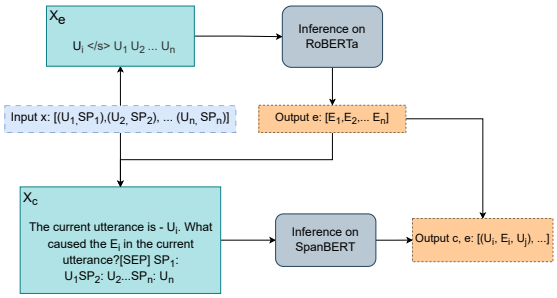

Figure 4: Performing inference at test time.

conversation from x_e, and then identify the causal span c_i for u_i in all cases when $e_i \neq$ NEUTRAL. We fine-tune separate pre-trained language models (PLMs) for these two steps. While fine-tuning for emotion cause spans, we use the emotion labels provided as part of the training set to construct x_c.

Emotion Classifier. We use a RoBERTa base model (Zhuang et al., 2021) as the backbone of our emotion classification model, with a weighted cross entropy loss to penalize emotion label predictions. We use class weights from our training set as weighing terms for the loss function, and fine-tune for 20 epochs using the AdamW optimizer (Loshchilov and Hutter, 2019) with linear rate scheduler with 500 warm-up steps. We use the learning rate 5×10^{-5} with 0.01 weight decay rate, and select the best epoch based on weighted F_1. The input prompt to this model is a space-separated concatenation of u_i, the separator token proposed by Zhuang et al. (2021), and all utterances in the conversation ($U_{\text{all}} = \{u_1, ..., u_n\}$) space-separated in sequential order, as shown in Figure 2.

Emotion Cause Classifier. We frame emotion cause classification as a QA task. To avoid asking our model to answer impossible questions, we skip utterances where the predicted label is NEUTRAL. We use a SpanBERT base model (Joshi et al., 2020) fine-tuned on SQuAD2.0 (Rajpurkar et al., 2018).[3] We then further fine-tune this model on our task. The input prompt to this model is:

> `The current utterance is - [`u_i`]. What caused the` e_i `in the current utterance? <SEP>` U_{all}

This is shown in Figure 3. For fine-tuning SpanBERT, we change the batch size from 32 to 12 and the maximum sequence length from 512 to 400. We set the learning rate to 2×10^{-5} and train the model for five epochs. Figure 4 shows inference using our proposed system. We first perform inference on our *emotion classifier* for the test dataset to augment test x_c with emotion labels e_i, and then perform inference on our *emotion cause classifier*.

4 Experimental Setup

ECF is already split into *train* and *test* sets. We separate a *dev* set from *train* by holding out the last 20% of the training data. We make use of pre-trained RoBERTa (Liu et al., 2019) and SpanBERT (Joshi et al., 2020) models from HuggingFace. Other details regarding our hardware and software libraries can be found in §A.3.

For training the *emotion classifier* we make use of weighted F_1 score, choosing the best performing model based on this metric. While training the *emotion cause classifier*, we select models based on metrics defined by Joshi et al. (2020) for span-based learning with PLMs: unweighted exact match, and F_1 score.

5 Results

5.1 Main Quantitative Findings

Our proposed system achieves 3rd place in SemEval Task #3, Sub-Task #1, based on the primary task metric of weighted average proportional F_1 (Wang et al., 2023). We achieve 2nd place overall based on the auxiliary metric of weighted average strict F_1,[4] which accounts for exact span matching. We show

[3] We observe that this additional fine-tuning boosts our model's performance (Appendix A.1).

[4] https://github.com/NUSTM/SemEval-2024_ECAC/tree/main/CodaLab/evaluation

Metric	Score	Ranking
w-avg. Strict F_1	0.1839	2
w-avg. Proportional F_1	**0.2442**	**3**
Strict F_1	0.1851	2
Proportional F_1	0.2397	4

Table 2: Official task scores, shown alongside final leaderboard rankings for Sub-Task #1.

Model	Metric	Score
Our Model	**w-avg. Strict F_1**	**0.2741**
Wang et al.	w-avg. Strict F_1	0.2625

Table 3: Comparing our model's performance on the *dev* set to Wang et al. (2023)'s text-only baseline.

our *test* scores for all official task metrics in Table 2. In Table 3 we compare to Wang et al. (2023)'s baseline for this task, showing that our model improves upon this baseline. Results reported in Table 3 are based on *dev* performance, since *test* was held private by the task organizers.

5.2 Quantitative Analysis

To investigate the performance of our approach, we used the *dev* dataset to perform an ablation study regarding prompt context length and fine-tuning data for the *emotion cause classifier*. We also experimented with varied scaling factors and input context for *emotion classification*.

5.2.1 Emotion Classification

Scaling Factors. We experimented with the use of class size as a scaling factor to improve performance for less-represented classes (e.g., *disgust* or *fear*). In Table 4, models post-fixed with *(w)* are scaled versions of the emotion classification model trained with a weighted cross-entropy loss. We observe large performance differences for under-represented classes under this condition; however, we also observe a slightly reduced F_1 overall. This is a positive observation for our disjoint training regime, as the causal span extractor model isn't trained on neutral cases during training, and it confirms the utility of class weight scaling for this task.

Input Context. We also experimented with varied input context, adjusting the context of the input by passing only `u_i` compared to the longer `u_i <SEP> U_all` used in our final model.

	u_i	u_i (w)	u_i </s> U_{all}	u_i </s> $U_{\text{all}}(w)$
Anger	0.46	0.45	**0.48**	**0.48**
Disgust	0.10	0.23	**0.24**	0.20
Fear	0.17	0.26	0.20	**0.27**
Joy	0.55	0.53	0.59	**0.60**
Sadness	**0.76**	0.74	0.73	0.72
Surprise	0.38	0.35	0.39	**0.40**
Neutral	0.63	0.53	**0.64**	0.58
F_1	0.60	0.58	0.60	0.59

Table 4: Ablation study on different prompts for the *emotion classifier*. </s> is the special token and (w) represents models trained with a weighted cross-entropy loss. F_1 is weighted average strict F_1.

5.2.2 Emotion Cause Classification

We experimented with two QA configurations examining prompt context length and fine-tuning data, shown in Table 5. In the former, we tweaked the model's maximum sequence length for models using the complete set of utterances in a conversation, U_{all}. We compared our results to a model trained only on prior context, i.e., $u \in \{u_1, ..., u_i\}$ where u_i is the current utterance. Interestingly, such models exhibited slightly higher F_1 scores; this is comparable to causal span extraction scores in (Poria et al., 2021). In the latter, we compared versions of our approach using (a) the pretrained SpanBERT directly, and (b) a version that was fine-tuned using SQuAD 2.0 data. We observed that additional training on a model previously trained on the SQuAD 2.0 dataset yields better performance than the pretrained SpanBERT model.

Sequence Length	EM	F_1
400	**0.4466**	**0.6133**
512	0.4397	0.6095
Model		
SpanBERT & SQuAD 2.0	**0.5147**	**0.6810**
SpanBERT	0.4428	0.6494

Table 5: Ablation study on sequence length (full context) and model for the *emotion causal classifier*. *EM* is exact match, and *F_1* is weighted average strict F_1. The base model of SpanBERT used is `spanbert-base-cased`. We prompt the model with only past and current context.

5.3 Error Analysis

We analyzed mispredicted examples from the *dev* set to identify commonly occurring errors that

might not be readily apparent by the w-avg proportional F1-score, and we observed that some of these conversations had neutral utterances with no corresponding emotion-cause pairs. From the 275 conversations, there were 23 such instances of which 12 were composed of only neutral utterances. In such cases, our span extractor model's output is accurate as it simply skips such utterances by design, and when neutral utterances are predicted correctly, this is the correct action. Conversely, in the cases where there are emotional utterances yet no causal pairs provided, the span model is unpredictable as it is not trained to predict empty causal spans, reinforcing our hypotheses grounded in Equation 1, i.e., that results of span extraction are dependent on the emotion classification model.

We also observed errors where incorrect spans were predicted for correctly identified emotions. In many instances, this involved the prediction of causal spans from *future* utterances. Given the nature of the data (informal conversations), it is possible for future utterances to overlap temporally with the current utterance. In other cases, it might seem to a third-person viewer that the cause of an emotion expressed at a timestep *t* becomes apparent after an utterance from a later timestep. Our model was not able to handle such cases predictably. Following manual review of all 32 predictions made for causal spans appearing in future utterances, we found that only 7 predictions were correct, mostly for the numerous emotion classes. This supports our rationale behind fine-tuning both our models in a full-context setting as explained earlier (§3), but suggests that there is still room for improvement. We suspect that the autoregressive context studied in follow-up analyses (§A.1) may result in better performance by skipping such examples, or perhaps a jointly-trained or multimodal model would help bridge the shortcomings.

Finally, we present a sample of correct predictions and mispredictions in Table 6. We observe that emotion classification for the most representative classes works as hypothesized, i.e., the emotions *joy* and *surprise* are predicted better than *fear*. For the span extraction task, we observe that rows 3 and 4 with non-neutral emotion predictions have "N/A" as their span prediction as, in these instances, the best prediction for an utterance with multiple causes returned identical spans as rows 2 and 5, respectively. One way to avoid such cases could be to pair all utterances u_i and u_j along with u_{all} ($u_i, u_j \in \{u_1, ..., u_n\}$), resulting in a quadratic increase in resource requirements and clipped inputs due to the model's limited token length. However, as this behavior was not consistent across all cases, we opted for the simpler solution described in §3. This also helped with resource constraints.

Utterance	Gold Emo.	Pred. Emo.	Gold Cause	Pred. Cause
Thank you. Oh Joey and look at this crib! It is so cute!	joy	joy	look at this crib! It is so cute!	It is so cute !
I know! I found it on the street.	joy	joy	It is so cute!	look at this crib! It is so cute!
I know! I found it on the street.	joy	joy	I found it on the street.	N/A
Are you serious ... Really ?! It is in such good condition.	surprise	surprise	I found it on the street.	N/A
Are you serious ... Really ?! It is in such good condition.	surprise	surprise	It is in such good condition.	It is in such good condition.
Yeah.	joy	neutral	It is in such good condition.	N/A
Wow! Whoa ... whoa what under the covers?	surprise	surprise	what under the covers?	It is in such good condition.
Ew.	fear	disgust	It is moving.	It is moving.
It is still ... It is got a tail! Get it out of here! Get it out of here!!	fear	fear	It is got a tail!	It is moving .
Ooh! Ah! Okay!	fear	surprise	It is got a tail!	It is moving.

Table 6: Correct and incorrect predictions from *dev*.

6 Conclusion

We introduce a disjoint model comprising an emotion classifier and an emotion-cause classifier. Our system addresses emotion cause extraction competitively based on the official leaderboard and on follow-up analyses (§5). We set out with the objectives of developing a disjoint model making use of

the entire conversation to identify emotions in utterances and using a well-known QA paradigm to extract the causes of the emotions, and we achieve this with varying degrees of success. We observe that emotion classification is harder than emotion cause extraction when emotion annotations are present (Tables 4 and 5), and that when the model assigns emotions correctly, it also has a greater chance of extracting causal spans correctly (Table 6). This is more evident when only prior contexts are present, yielding higher scores.

Acknowledgments

We thank the anonymous reviewers for their helpful feedback, and the SemEval-2024 Task 3 organizers for introducing us to this challenging and intriguing research problem.

References

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Ebrahim (Abe) Kazemzadeh, Emily Mower Provost, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. 2008. Iemocap: interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42:335–359.

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.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. SpanBERT: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536, Florence, Italy. 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, Alexander Gelbukh, and Rada Mihalcea. 2021. Recognizing emotion cause in conversations.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

Fanfan Wang, Zixiang Ding, Rui Xia, Zhaoyu Li, and Jianfei Yu. 2023. Multimodal emotion-cause pair extraction in conversations. *IEEE Transactions on Affective Computing*, 14(3):1832–1844.

Fanfan Wang, Heqing Ma, Rui Xia, Jianfei Yu, and Erik Cambria. 2024. Semeval-2024 task 3: Multimodal emotion cause analysis in conversations. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 2022–2033, Mexico City, Mexico. 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.

Liu Zhuang, Lin Wayne, Shi Ya, and Zhao Jun. 2021. A robustly optimized BERT pre-training approach with post-training. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 1218–1227, Huhhot, China. Chinese Information Processing Society of China.

A Appendix

A.1 SQuAD 2.0 Fine-Tuning Affects Emotion Cause Classification

We observe a strong increase in performance of the *emotion cause classifier* if additional fine-tuning is performed using SQuAD 2.0 (Rajpurkar et al., 2018). In this case, the model is prompted with The

current utterance is u_i What caused the e_i in current utterance?. We did not consider utterances $u \in \{u_{i+1}, ..., u_n\}$. The unweighted exact match and F_1 increases, as shown in Table 5.

A.2 Other Metrics for the Model

Metric	Value
Weighted strict precision	0.339
Weighted strict recall	0.235
Weighted strict F-1	0.274
Weighted Proportional precision	0.425
Weighted Proportional recall	0.288
Weighted Proportional F-1	**0.339**
Strict precision	0.348
Strict recall	0.235
Strict F-1	0.280
Proportional precision	0.431
Proportional recall	0.280
Proportional F-1	0.339

Table 7: Additional results for our model on the *dev* set as defined by Wang et al. (2023). Weighted Proportional F_1 was the primary metric used for SemEval Task #3.

We provide additional results for other metrics defined by Wang et al. (2023) in Table 7.

A.3 Hardware and Hyperparameters

We make use of PyTorch 2.2.0,[5] HuggingFace transformers 4.37.2 for the RoBERTa-base implementation,[6] FAIR's[7] spanbert-base-cased, FAIR's SpanBERT fine-tuned on SQuAD2.0 and sklearn 1.3.2[8] to fine-tune our models. We train our models using an Nvidia RTX 3090Ti GPU.

[5] https://pytorch.org/get-started/locally/

[6] https://huggingface.co/docs/transformers/en/installation

[7] https://github.com/facebookresearch/SpanBERT/

[8] https://scikit-learn.org/stable/install.html