Analysis of Emotion Cause

Track 3: Personal Project

Nakul Pacheriwala

np2455@nyu.edu

Sayantan Mukhopadhyay

sm9752@nyu.edu

Mentor: Jun Yuan

jy50@nyu.edu

https://github.com/Nakul24-1/Analysis-of-Emotion-Cause

Analysis of Emotion Cause

Nakul Pacheriwala
np2455@nyu.edu

Sayantan Mukhopadhyay
sm9752@nyu.edu

Jun Yuan
Mentor
jy50@nyu.edu

Abstract

We present RoBERTa-emotion-classification for classifying a sequence to a particular emotion and RoBERTa-emotion-extraction to extract spans of texts in a sentence, given an emotion. We preset this as an alternative to black box explainability methods like SHAP and LIME. Our hypothesis is that a pre-trained NLP model like RoBERTa can be fine tuned to make it be able to predict the cause of an emotion better than the explanations provided by SHAP or LIME - since the explainer model is essentially a descendant of the predictor model. We have used the EmoCause dataset, containing human annotations - that we have considered as our baseline. We have conducted a series of experiments comparing the performances of the two model to prove our hypothesis.

1 Introduction

Emotions are deep-rooted in humans; consequently, understanding emotions is a key part of human-like artificial intelligence. One of the many ways to express human emotions is through language. Using Natural Language Processing, we can recognize emotions that have applications in a large number of domains like healthcare and finance.

While there has been substantial progress in the detection and classification of human emotions, understanding and exploring the causes of emotions is a nascent developmental field. It is important to understand the difference between emotion evidence and emotion cause as the dataset we used to train uses annotations explaining emotion cause. Emotion evidence is a part of the text that indicates the presence of emotion in the speaker's emotional state. Emotion cause is a part of the text expressing the reason for the speaker to feel the emotion given by the emotional evidence.

The main goal of this project is to find a method to understand the behavior of Natural Language Processing models by comparing human annotations to explanations by the models. We believe visualizing the levels of intersection between human annotation and model explanations can give us an insight into model behavior which can help us understand where the model fails and how it can be improved.

When working with NLP models the context is very important as the same statement can mean different things in different scenarios, so if we know what words/phrases caused the model to predict a certain outcome, it would be easier to understand if the context was taken into account or not. Due to the intertwined dynamics between the interlocutors, identifying emotional sources at the expressive level in dialogues is a difficult task. If we can see the explanation of the result from the NLP model, a human can decide if they should accept or reconsider the output.

2 Related Works

Poria et. al [4] focus on emotion cause recognition as the task, with the belief that this can aid the interpretability of Contextual Models. For example, some utterances may not contain any emotion-bearing words explicitly, or, the sentences can sound neutral but still carry some emotions that can only be inferred from the context as a whole. They also believe that it would be worth checking if the NLP models assign a high probability to the span of words. Poria et. al defined relevant types of emotional causes namely- No Context, Interpersonal Emotional Influence, Self-contagion, Hybrid Emotion, Unmentioned Latent Cause. They have introduced two sub-tasks: causal span extraction and causal emotion entailment which demand complex reasoning and are thus challenging. They have also set up strong baselines to solve these sub-tasks.

Kim et. al [2] use social cognition to infer causes of emotion from utterances. They have introduced a method based on pragmatics to make dialogue models focus on targeted words. They have annotated the emotion-causing words in emotional situations from the validation and test set of EmpatheticDialogues dataset. We used this as the dataset for our training and testing purposes.

In the paper Shared Interest Boggust et. al [1] quantifies the alignment between these two components by measuring three types of coverage: Ground Truth Coverage (GTC), or the proportion of ground truth features identified by the saliency method; Saliency Coverage (SC), or the proportion of saliency features that are also ground truth features; and IoU Coverage (IoU), the similarity between the saliency and ground truth feature sets. These coverage metrics enable a richer and more structured interactive analysis process by allowing analysts to sort, rank, and aggregate input instances based on model behavior. Leveraging the Shared Interest metrics alongside interactive human annotation enables a question-and-answer process where analysts probe input features and Shared Interest identifies the model's decisions whose saliency feature sets are most aligned.

In RoBERTa: A Robustly Optimized BERT Pretraining Approach, Liu et. all [5] have found that the NLP model BERT was undertrained and could match the performances of models which were made later. They have used overlooked design choices, to provide a model that achieves state of the art results on GLUE, RACE and SQuAD. They call it RoBERTA and we will use RoBERTa as our NLP model.

In SpanBERT: Improving Pre-training by Representing and Predicting Spans, Joshi et. all [3] extend BERT by masking random spans and training the span boundary representations to predict the entire content of the masked span. They present SpanBERT which consistently outperforms BERT and gains on span selection tasks such as question answering.

3 Methods

The following methodology was used for the problem.

3.1 Loading the dataset

The EmoCause dataset was downloaded into train and test sections, and loaded onto corresponding dataframes. The three main columns of the dataset are

- original-situation: A line of text containing a situation a person is in
- emotion: The emotion felt by the person at the time of the situation
- labels: The human annotation explaining the cause of the emotion. This is considered to be the baseline for the experiment.

3.2 Fine-tuning RoBERTa-base for Sequence Classification

The pre-trained NLP model RoBERTa-base [?] from the transformers library is fine tuned using the SequenceClassification transformer. It is trained using the text from original-situation as input and the human defined emotions as labels, from the EmoCause train dataset. The model is run for six epochs while recording the loss and accuracy. The version of the model, which predicted the emotion with the least loss while retaining considerable accuracy, is saved. We call this RoBERTa-emotion-classification and have uploaded it to Huggingface for later use. The saved model is tested on the test dataset to measure the performance using metrics like Intersction Over Union (IOU) and Accuracy [ECC]. The model's output performance is also tested using some custom data.

3.3 Fine-tuning RoBERTa-emotion-classification for Question Answering

The fine-tuned model RoBERTa-emotion-classification is further fine tuned using the QuestionAnswering transformer. It is trained using the RoBERTa predicted emotion as input and the human defined annotations from the EmoCause train dataset as label. The model is run for six epochs while recording the loss and accuracy. The version of the model, which predicted the emotion with the least loss while retaining considerable accuracy, is saved. We call this RoBERTa-emotion-extraction and have uploaded it to Huggingface for later use. The saved model is tested on the test dataset to measure the performance using metrics like F-1 score and Accuracy. The model's output performance is also tested using some custom data.

3.4 Using SHAP to generate Explamations

SHAP takes the RoBERTa-emotion-classification model and RoBERTa predicted emotion as input and displays the Shapley values for each word. We take the words with positive Shapley values as SHAP predicted annotation.

3.5 Comparing RoBERTa predicted spans with SHAP generated Explanations

The RoBERTa predicted annotation is compared with SHAP predicted annotation, both against Human Annotations, which we have considered consider as baseline.

3.6 Visualization of the Results

An application which visualizes the spans of text predicted by RoBERTa-emotion-extraction and the SHAP generated explanations is built using DASH. It highlights the texts generated by each method and helps to compare the performance of each method. This application can be used to compare and contrast between SHAP explanations and the explanation given by RoBERTa-emotion-extraction model.

4 Results

The results of the experiment are given in the Table 1.

5 Figures

The highlighted text in the application denotes the spans of text predicted by each explanation model.

6 Discussion

We have performed analysis of NLP model behavior. To understand the working of our model, metrics that quantify instances based on how the transformers model aligns with annotations by humans.Our interactive visualization hosted on a flask based web page allows us to compare the model explanations by SHAP and our Models very clearly as shown in the figure below.Our web page is also able to work with custom inputs of text where it can predict emotion using our fine-tuned RoBERTa classifier which is then utilized by our RoBERTa Question Answering model to explain the cause of the emotion which was predicted. The results above show that our model aligns with human annotations much better than SHAP.

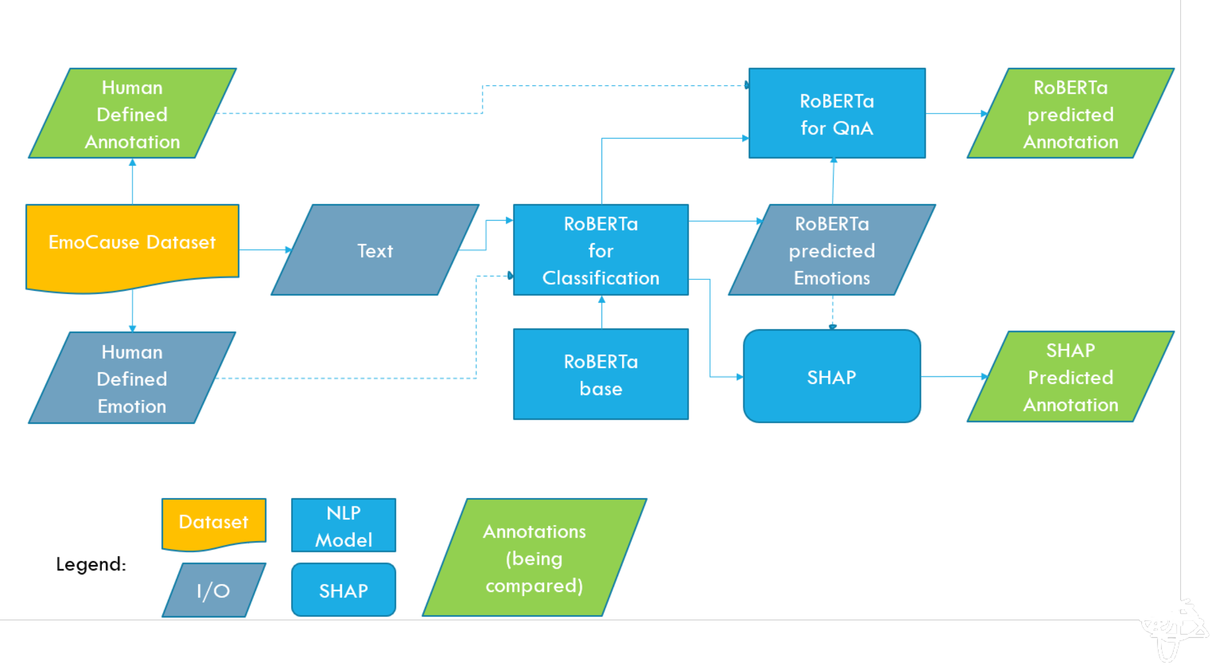
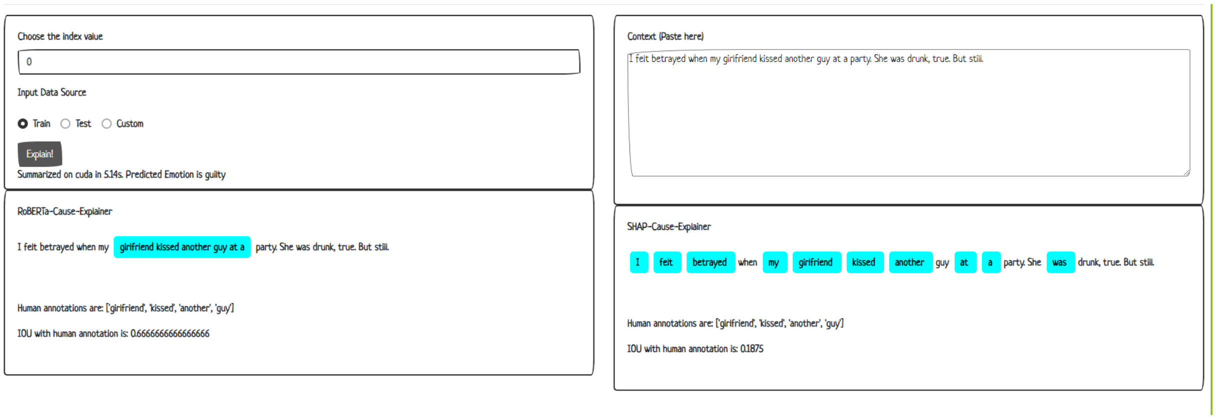

Human Defined Annotation
RoBERTa for QnA
RoBERTa predicted Annotation
EmoCause Dataset
Text
RoBERTa for Classification
RoBERTa predicted Emotions
Human Defined Emotion
RoBERTa base
SHAP
SHAP Predicted Annotation
Legend:
Dataset
NLP Model
Annotations (being compared)
I/O
SHAP

Figure 1. Architecture diagram showing the flow of our project

Table 1. Results

Model	Train Accuracy	Test Accuracy	Train IOU	Test IOU
RoBERTA	0.69	0.45	0.77	0.65
SpanBERT	0.71	0.37	0.82	0.45
SHAP	0.0	0.0	0.12	0.15



Choose the index value

0

Input Data Source

● Train ○ Test ○ Custom

Explain!

Summarized on cuda in 5.14s. Predicted Emotion is guilty

RoBERTa-Cause-Explainer

I felt betrayed when my girlfriend kissed another guy at a party. She was drunk, true. But still.

Human annotations are: ['girlfriend', 'kissed', 'another', 'guy']

IOU with human annotation is: 0.6666666666666666

Context (Paste here)

I felt betrayed when my girlfriend kissed another guy at a party. She was drunk, true. But still.

SHAP-Cause-Explainer

I felt betrayed when my girlfriend kissed another guy at a party. She was drunk, true. But still.

Human annotations are: ['girlfriend', 'kissed', 'another', 'guy']

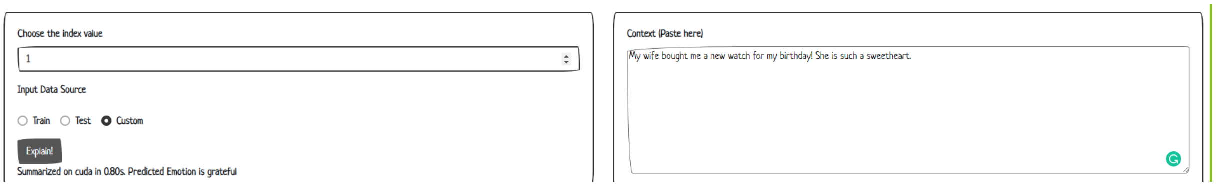
IOU with human annotation is: 0.1875

Figure 2. This Figure shows us how custom input can be utilized to predict emotion and explain it's cause.

6.1 Limitations

Unlike SHAP we need human annotations for our model to work, that could be a serious issue specially for domains where humans are not able to annotate the explanations. In our current model we cannot take into account the history of conversation and how it is affecting the current emotion, it considers the given text as whole and cannot differentiate between multiple entities.

Another major flaw of our RoBERTa-emotion-extraction model is that it only predicts one continuous span that contains majority of important words, however there are many



Choose the index value

1

Input Data Source

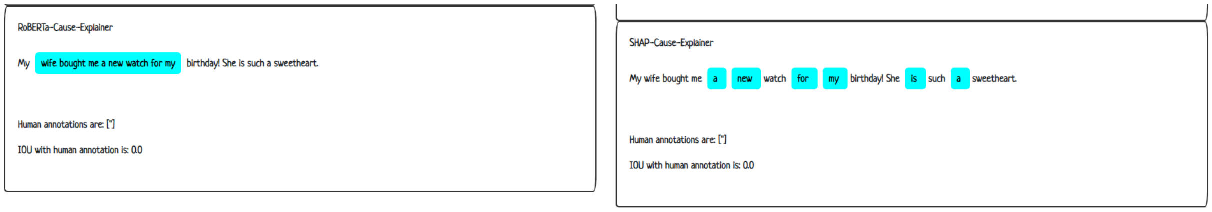
○ Train ○ Test ● Custom

Summarized on cuda in 0.80s. Predicted Emotion is grateful

Context (Paste here)

My wife bought me a new watch for my birthday! She is such a sweetheart.

Figure 3. This is our Flask based UI, where we have visualized the output of the various models we had trained. We can see the predicted emotion and explanations of both our model and SHAP.



RoBERTa-Cause-Explainer

My wife bought me a new watch for my birthday! She is such a sweetheart.

Human annotations are: ['']

IOU with human annotation is: 0.0

SHAP-Cause-Explainer

My wife bought me a new watch for my birthday! She is such a sweetheart.

Human annotations are: ['']

IOU with human annotation is: 0.0

Figure 4. Here we can observe that IoU values are 0, this is because we do not have human annotations for custom text to compare with.

cases where the words causing an emotion are separated by words which add no meaning or are not relevant. Our model would also include those words as it cannot predict multiple spans.

6.2 Future Work

We used only one statement from a conversation in our current experiment. We can extend it further to analyze a full conversation between multiple entities and analyze how the people conversing affect each other's emotion and its cause. Also, We instead of using a Question and Answering model to explain, Bert Summary option to find the most relevant words in the passage can be used as an alternative. We can also compare with other explanation methods like Lime and ELI5.

7 Conclusion

We presented a novel idea which was training a neural network model to explain another neural network.After extensive analysis using metrics like IoU and simple accuracy we have observed that our model consistently outperforms SHAP. RoBERTa-Emotion-Extraction model can be used to understand how the RoBERTa-Emotion-Prediction model works and can be instrumental in further enhancement of the prediction model.

References

[1] Arvind Satyanarayan Hendrik Strobelt Angie Boggust, Benjamin Hoover. March 2022. Shared Interest: Measuring Human-AI Alignment to Identify Recurring Patterns in Model Behavior. *arXiv* (March 2022). https://doi.org/10.48550/arXiv.2107.09234

[2] Gunhee Kim Hyunwoo Kim, Byeongchang Kim. September 2021. Perspective-taking and Pragmatics for Generating Empathetic Responses Focused on Emotion Causes. *arXiv* (September 2021). https://doi.org/10.48550/arXiv.2109.08828

[3] Yinhan Liu Daniel S. Weld Luke Zettlemoyer Omer Levy Mandar Joshi, Danqi Chen. January 2020. SpanBERT: Improving Pre-training by Representing and Predicting Spans. *arXiv* (January 2020). https://doi.org/10.48550/arXiv.1907.10529

[4] Devamanyu Hazarika Deepanway Ghosal Rishabh Bhardwaj Samson Yu Bai Jian Pengfei Hong Romila Ghosh Abhinaba Roy Niyati Chhaya Alexander Gelbukh Soujanya Poria, Navonil Majumder and Rada Mihalcea. September 2021. Recognizing Emotion Cause in Conversations. *Springer* (September 2021). https://doi.org/10.1007/s12559-021-09925-7

[5] Naman Goyal Jingfei Du Mandar Joshi Danqi Chen Omer Levy Mike Lewis Luke Zettlemoyer Veselin Stoyanov Yinhan Liu, Myle Ott. July 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv* (July 2019). https://doi.org/10.48550/arXiv.1907.11692