Exploring Model Behaviors with Google's Learning Interpretability Tool (LIT)

A Project Report for the evaluation of CMPSC 291I

Master of Science

by

Alisetti Sai Vamsi
Perm Number - A1U0061

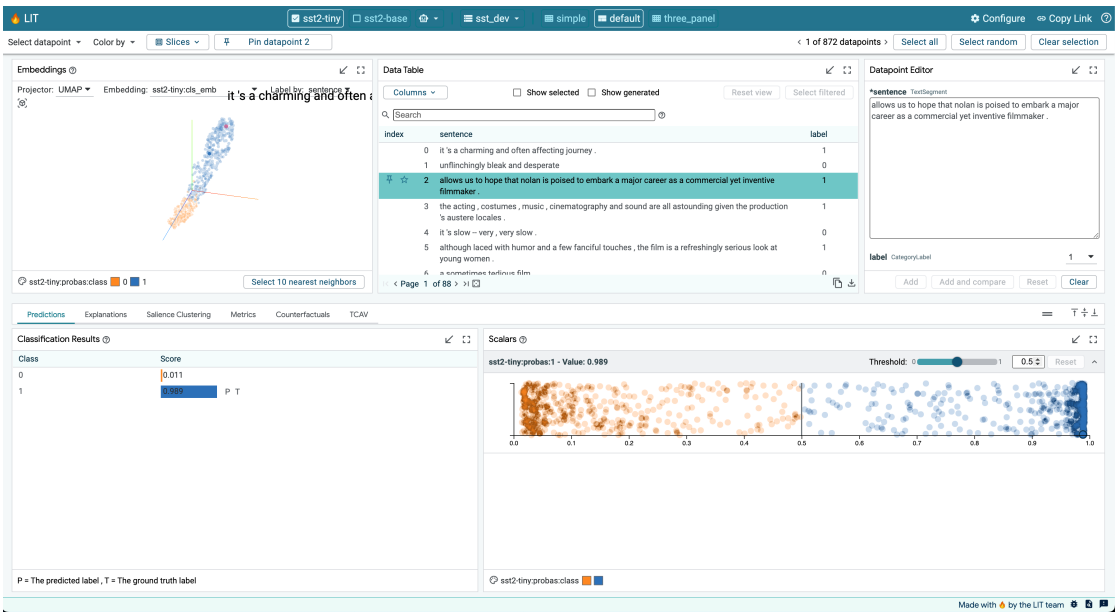
Department of Computer Science
Santa Barbara, CA - 93106
October 2024

Contents

1 Setup

1.1 Language Interpretability Tool

THe Language Interpretability Tool (LIT) is a framework designed to help developers, researchers, and machine learning practitioners understand and analyze natural language processing (NLP) models. The goal of the tool is to provide insights into how language models make decisions, identify model weaknesses, and offer mechanisms to interpret model predictions. I have utilized Google Colab as my environment for interfacing with the LIT tool.



In Google Colab you could simply set it up by installing the python library as such using pip.

```
# Install LIT Library
!pip install lit-nlp
```

Show hidden output

2 Dataset Exploration

The SST-2 (Stanford Sentiment Treebank 2) dataset is a popular benchmark for sentiment analysis. It is a binary classification dataset derived from the Stanford Sentiment Treebank, where each sentence is labeled as either positive or negative.

- Type: Binary classification (positive/negative)
- Source: Movie reviews from the Rotten Tomatoes website

- Size: 67349 (Train), 872 (Validation), 1821 (Test)
- Label: (1: Positive sentiment), (0: Negative sentiment),

Unlike the full SST dataset, SST-2 focuses solely on sentence-level binary classification, omitting neutral sentiments and tree structures from the original treebank. It's commonly used in NLP tasks for evaluating models like BERT and GPT.

2.1 Data Distribution

The validation dataset exhibits a balanced distribution as we can see from Fig. 1, of positive and negative examples, meaning that the number of instances in each class is relatively similar. There is no substantial skew or disproportionate representation of one class over the other, ensuring that both positive and negative examples are fairly represented. This balanced nature reduces the risk of model bias toward one class and provides a more reliable evaluation of the model's performance across both categories.

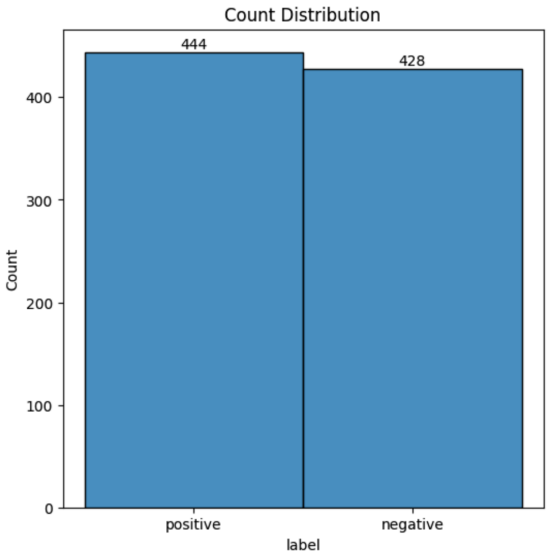


Figure 1: Label Distribution in the validation set of SST2 Dataset.

2.2 Embeddings

When utilizing the SST-Tiny model, a simplified version of BERT, we can analyze the dataset's sentence embeddings. By leveraging the UMAP (Uniform Manifold Approximation and Projection) algorithm for dimensionality reduction, which is built into the LIT (Language Interpretability Tool) Framework, each sentence is transformed into a 3-dimensional space. In this reduced space, a distinct separation between positive and negative sentiment examples becomes apparent. This separation indicates that the model's

embeddings effectively capture the underlying semantic differences between positive and negative sentences, allowing similar sentiments to cluster together. This visualization not only demonstrates the model's ability to encode sentiment-related features but also provides insights into how the sentence embeddings align with the sentiment labels..

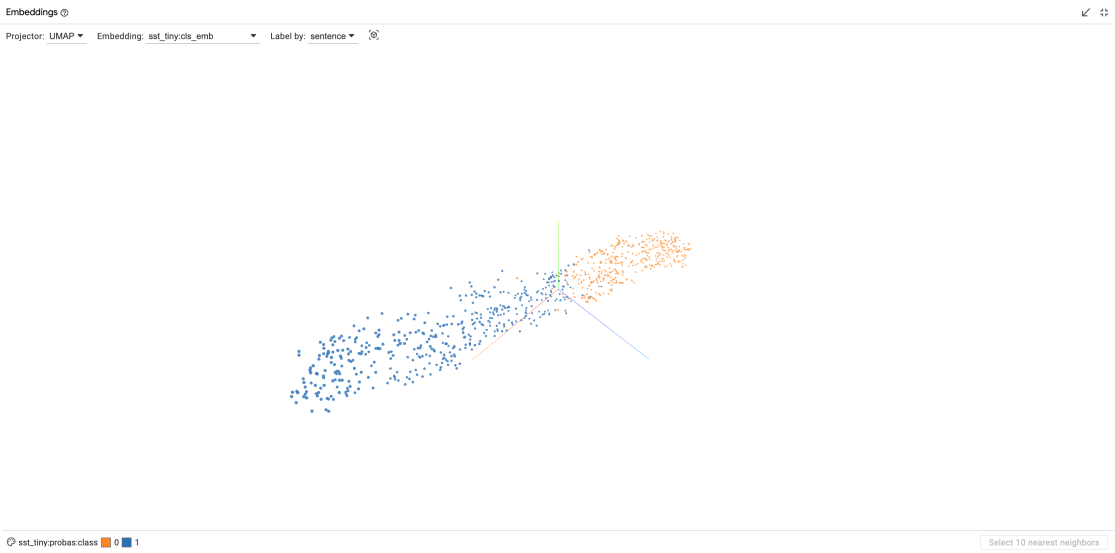


Figure 2: Sentence Embeddings from the SST-Tiny Model, where blue represents positive sentiment and orange represents negative sentiment. We can clearly see two clusters forming indicating the semantic similarity between sentences with same sentiment.

2.3 Confusion Matrix

The confusion matrix reveals that the SST-Tiny model misclassifies 8.5% of the positive examples (74 data points) and 9.4% of the negative examples (82 data points). These misclassifications may be attributed to several factors, one of which is the inherent limitations of the model due to its smaller size and reduced number of hidden layers compared to larger models like BERT or RoBERTa. The compact nature of SST-Tiny, designed to prioritize efficiency, may constrain its ability to capture more complex linguistic patterns or nuanced sentiment expressions within the dataset. As a result, the model may struggle to correctly classify certain ambiguous or contextually rich examples, leading to these errors. Additionally, fewer parameters and hidden layers can limit the model's capacity to learn intricate representations, potentially reducing its overall performance on more subtle or less common sentiment cases.

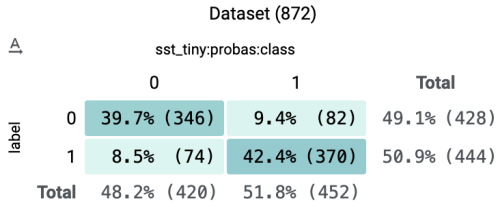


Figure 3: SST-Tiny Confusion Matrix on SST2 Dataset.

3 Model Performance Analysis

3.1 Choice of Models

For this analysis, I selected SST-Tiny (6M parameters), DistilBERT (66M parameters), and RoBERTa (125M parameters) models from the Hugging Face library. The code for loading these models into the LIT framework is provided. To evaluate model performance, I focus on three subsets: positive sentiment, negative sentiment, and length-based subsets. This allows me to assess how well the models handle misclassifications for positive and negative sentiments, as well as their ability to capture long-term dependencies in longer texts. These subsets provide a deeper understanding of the models' strengths and weaknesses across various linguistic challenges.

3.2 Positive Sentiment Subset

3.2.1 Dataset Slice

This subset slice contains all the 444 positive examples as the subset. Analyzing this slice we can understand how well the model is able to perform on positive sentiment, and we can also understand on why its failing on the false positives in the confusion matrix. After filtering through the SST Tiny classifications, we get 74 datapoints that are misclassified as negative. Out of which Fig. 4 shows the top 10 strongly misclassified sentences by SST Tiny.

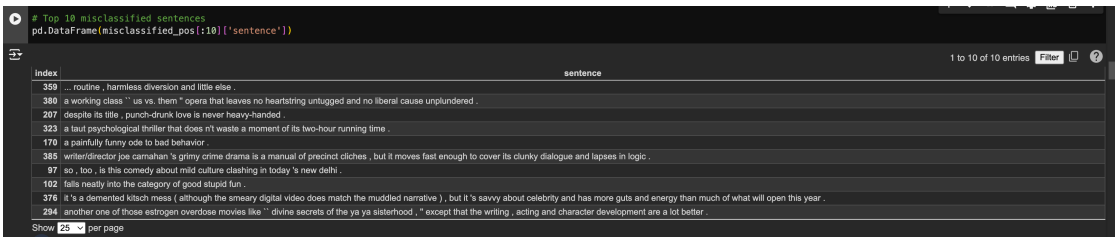
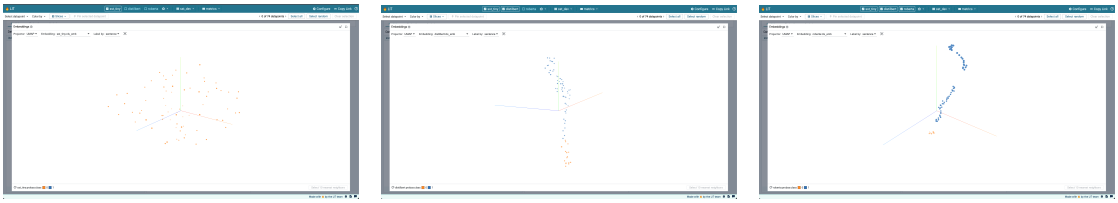


Figure 4: Top 10 sentences strongly misclassified by SST-Tiny.

From the Fig. 5 we can see how the embeddings of 74 misclassifications of positive

sentiment by the SST-Tiny model. We can see how the advanced models like DistilBERT and RoBERTa are clearly clustering these examples as well, showing that they are able to outperform the SST-Tiny model.



SST-Tiny Embeddings DistilBERT Embeddings RoBERTa Embeddings

Figure 5: Embeddings of the 74 misclassified positive sentiment (by SST-Tiny) datapoints

This is further highlighted from the scalar plot shown in Fig. 6 where we can see the datapoints moving closer to the edges in the left and right direction in DistilBERT and RoBERTa models compared to the SST-Tiny.

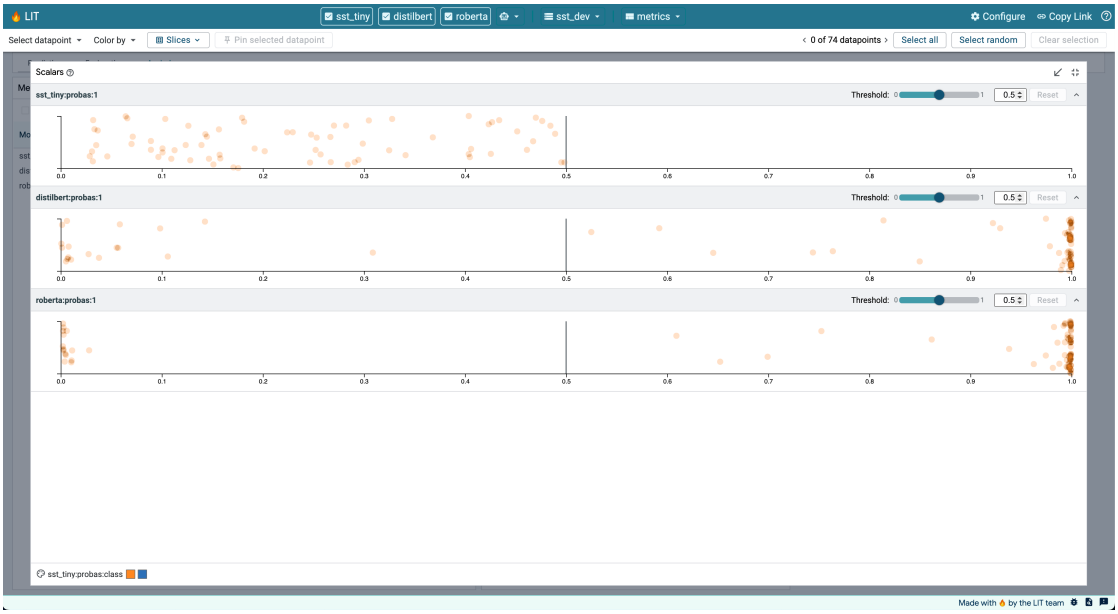
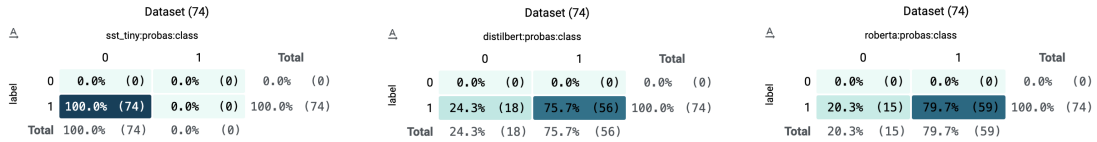


Figure 6: Scalar Analysis Chart.

3.2.2 Classification Performance

From Fig. 7 we can analyze how well DistilBERT is performing compared to SST-Tiny where it is able to classify 56 of the 74 misclassified points correctly, and when it comes to RoBERTa it is able to classify 59 of the 74 misclassified points correctly. Although



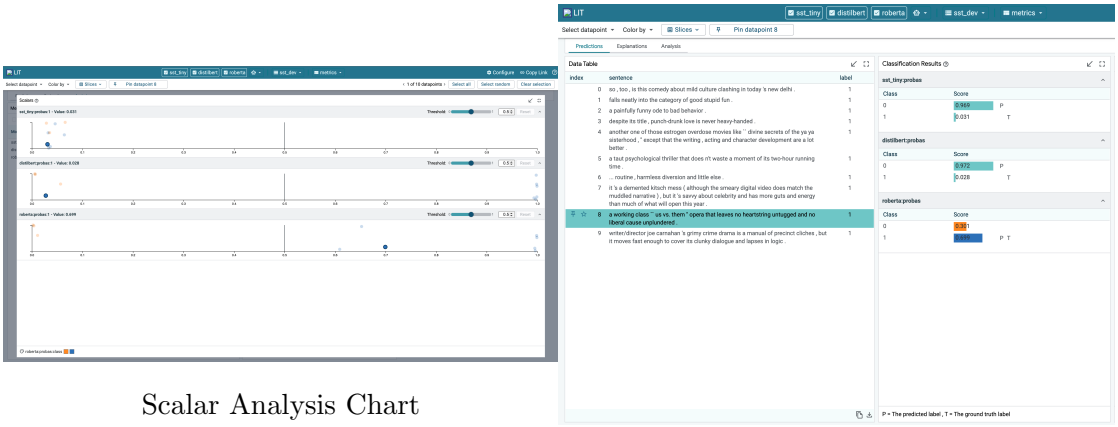
SST-Tiny DistilBERT RoBERTa

Figure 7: Confusion Matrices of the three models

there is a slight difference in the number of correct classifications, RoBERTa is clearly performing better than the other two models.

3.2.3 Analysis

Let us take the top 10 misclassified sentences for our analysis. We can see that there are three datapoints misclassified by all, and one datapoint that RoBERTa managed to predict correctly from Fig. 13.



Scalar Analysis Chart

Data Table

Figure 8: Analysis of Top 10 misclassified datapoints

The sentence ***"a working class " us vs. them " opera that leaves no heartstring untugged and no liberal cause unplundered"*** is correctly predicted by RoBERTa as having positive sentiment. RoBERTa is able to grasp subtle sentiment present in complicated sentences like "no heartstring untugged" and "no liberal cause unplundered" which may require the model to recognize irony or specific idiomatic expressions that are harder for smaller models to grasp.

While on the other hand all three models are misclassifying the sentence ***"writer/director joe carnahan 's grimy crime drama is a manual of precinct cliches , but it moves fast enough to cover its clunky dialogue and lapses in logic ."***

because the sentence contains phrases like “grimy”, “manual of precinct clichés”, “clunky dialogue”, and “lapses in logic”, which are typically associated with negative sentiment. Despite the positive implication of the film ”moving fast enough” the presence of these negative descriptors can heavily influence the models’ assessments. The sentence’s structure presents a mix of sentiments that can confuse the models. The juxtaposition of the film’s flaws and merits in a single sentence requires a more sophisticated understanding of sentiment analysis, which smaller models may lack. RoBERTa, despite being a larger model, might still misinterpret the overall sentiment due to the predominance of negative descriptors.

3.3 Negative Sentiment Subset

3.3.1 Dataset Slice

This subset slice contains all the 428 negative examples as the subset. Analyzing this slice we can understand how well the model is able to perform on negative sentiment, and we can also understand on why its failing on the false negatives in the confusion matrix. After filtering through the SST Tiny classifications, we get 82 datapoints that are misclassified as positive. Out of which Fig. 9 shows the top 10 strongly misclassified sentences by SST Tiny.

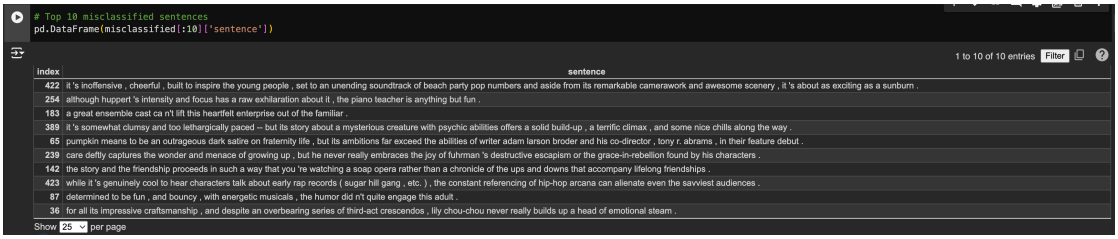
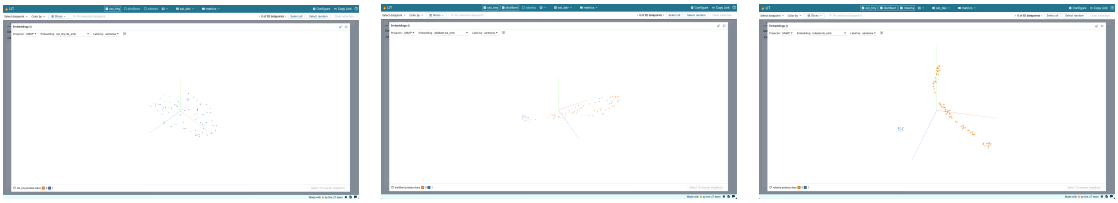


Figure 9: Top 10 sentences strongly misclassified by SST-Tiny

From the Fig. 10 we can see how the embeddings of 82 misclassifications of negative sentiment by the SST-Tiny model. We can see how the advanced models like DistilBERT and RoBERTa are clearly clustering these examples as well, showing that they are able to outperform the SST-Tiny model.



SST-Tiny Embeddings DistilBERT Embeddings RoBERTa Embeddings

Figure 10: Embeddings of the 83 misclassified negative sentiment (by SST-Tiny) datapoints

This is further highlighted from the scalar plot shown in Fig. 11 where we can see the datapoints moving closer to the edges in the left and right direction in DistilBERT and RoBERTa models compared to the SST-Tiny.

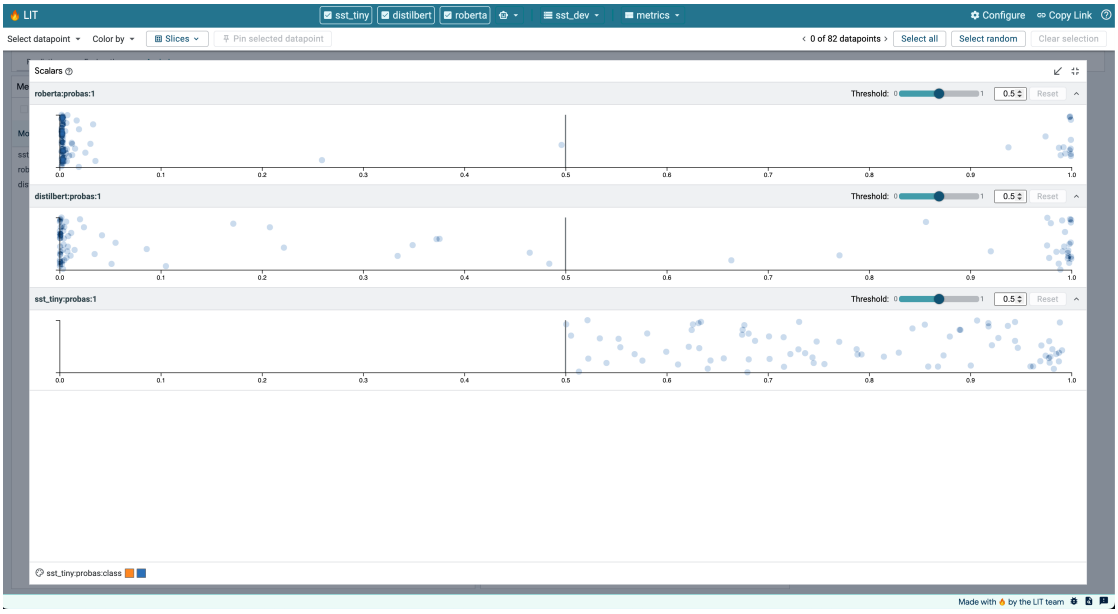
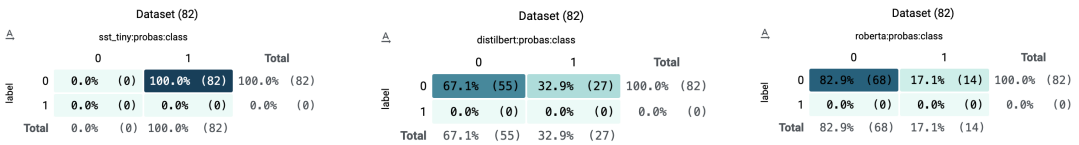


Figure 11: Scalar Analysis Chart

3.3.2 Classification Performance



SST-Tiny DistilBERT RoBERTa

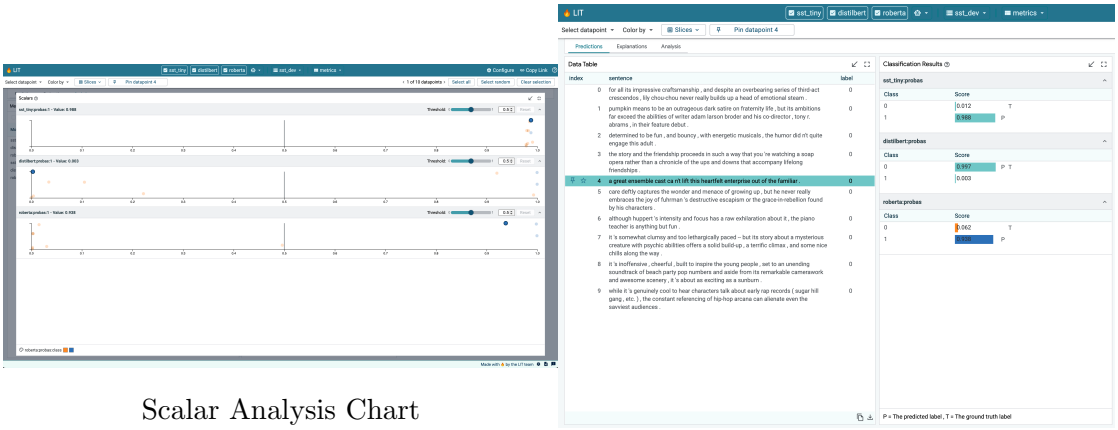
Figure 12: Confusion Matrices of the three models

From Fig. 12 we can analyze how well DistilBERT is performing compared to SST-Tiny where it is able to classify 55 of the 82 misclassified points correctly, and when it comes to RoBERTa it is able to classify 68 of the 74 misclassified points correctly. Although there is a slight difference in the number of correct classifications, RoBERTa is clearly performing better than the other two models in case of negative sentiment as well. We can identify one pattern, that in case of positive sentiment RoBERTa was able to classify 59/74, but in case of the negative sentiment its 68/82, which is higher than that of the positive sentiment indicating that it is biased towards detecting negative

sentiment more than the positive sentiment. But with DistilBERT its the otherway around, where it was able to classify 56/74 for the positive case, and 55/82 for the negative case, indicating a slight bias towards detecting positive sentiment more than the negative one.

3.3.3 Analysis

Let us take the top 10 misclassified sentences for our analysis. We can see that there are three datapoints misclassified by all, and one datapoint that DistilBERT managed to predict correctly which RoBERTa misssed from Fig. 13.



Scalar Analysis Chart

Data Table

Figure 13: Analysis of Top 10 misclassified datapoints

The sentence ***"A great ensemble cast can't lift this heartfelt enterprise out of the familiar."*** poses a challenge due to its nuanced language and context. The phrase ***"a great ensemble cast"*** can initially signal a positive sentiment, suggesting that the actors are commendable. However, the latter part of the sentence ***"can't lift this heartfelt enterprise out of the familiar"*** implies a negative evaluation, indicating that despite the cast's talent, the overall experience is unoriginal or lacks depth. The contrast between the initial positive phrase and the subsequent negative sentiment creates ambiguity that can confuse models. RoBERTa and SST-Tiny may not effectively capture the overall sentiment due to their interpretations of the sentence's structure. They might overly focus on the positive aspect of the cast and fail to fully integrate the negative implications that follow. These models may misjudge the overall sentiment as positive due to the leading phrase, neglecting the significance of the latter part. DistilBERT, while smaller than RoBERTa, appears to have correctly identified the negative sentiment by recognizing that the phrase "can't lift this heartfelt enterprise out of the familiar" serves as a critique, overshadowing the positive note about the cast. RoBERTa's larger architecture is typically more powerful but can also lead to overfitting to certain phrases or structures in its training data. In contrast, DistilBERT's simpler

architecture may have a more straightforward interpretation, allowing it to focus on the critical context of the sentence more effectively.

The sentence ***"Although Huppert's intensity and focus has a raw exhilaration about it, The Piano Teacher is anything but fun."*** is an example of how subtle language and conflicting cues can confuse models. All the three models marked this as positive when infact its a negative sentiment. I feel that the models are prioritizing words like "exhilaration," "intensity," and "focus" and fail to appropriately adjust their sentiment when encountering negations or more subtle signals, such as "anything but fun." This suggests that the models might struggle to handle linguistic subtleties and are more prone to reacting strongly to individual positive words.

3.4 Length Based Subset

About the slice

3.4.1 Dataset Slice

This subset slice divides the dataset into two sub halves based on their length. Analyzing this slice we can understand how well the model is able to perform on text that is long, and we can also understand the model's attention on long-range tokens. The maximum number of words is 48 in the dataset and the minimum number is 3. I calculated the median length as shown in Fig. 14.

Figure 14: Calculating Max, Min, Meadian lengths of sentences in the dataset

After calculating the median length, I divided it into two halves from min-median (lower length), and median-max (upper length), where you get 275 datapoints in the upper half, and 597 datapoints in the lower half. You can see the top 5 longest sentences (Fig. 15) and the top 5 shortest sentences (Fig. 16) below.

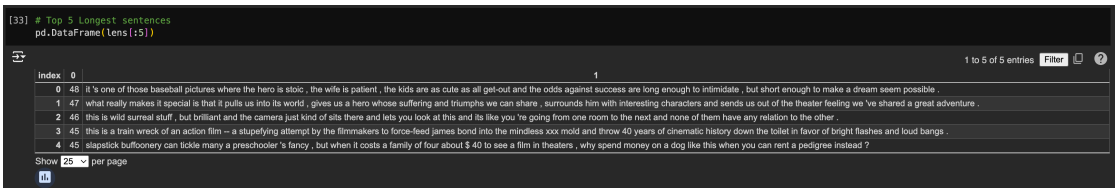


Figure 15: Top 5 Longest Sentences

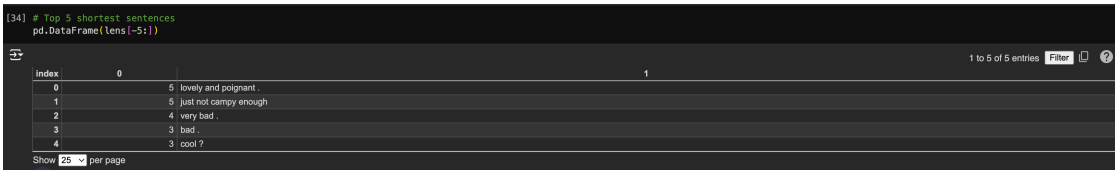
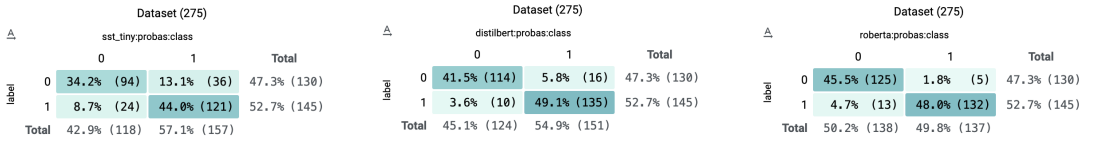


Figure 16: Top 5 Shortest Sentences

3.4.2 Classification Performance

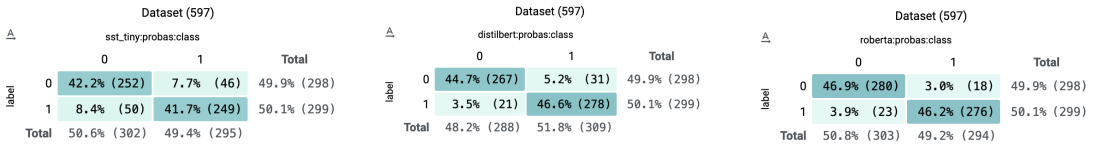
From the upper half confusion matrix we can clearly say that compared to DistilBERT, RoBERTa is quite good at detecting long range negative sentiments, i.e texts having more words with a negative connotation because the number of false negatives (1.8% in RoBERTa compared to 5.8% in DistilBERT and 13.1% in SST-Tiny). Whereas DistilBERT is better at detecting long range positive sentiments based on the percentage of false positives (3.6% in DistilBERT compared to 4.7% in RoBERTa and 13.1% in SST-Tiny).



SST-Tiny DistilBERT RoBERTa

Figure 17: Confusion Matrices of the three models for upper length dataset

From the lower half confusion matrix we can see the same trend, but the gap is sightly narrowed for RoBERTa and DistilBERT in predicting positive sentiment. It seems like RoBERTa is better at short range positive sentiment detection than long range based on the false positive percentage (3.9% in lower subset, and 4.7% in upper subset). Whereas DistilBERT seems to perform almost the same on short range and long range texts with positive sentiment based on false positive rate (3.6% in upper subset, and 3.5% in lower subset), which is quite surprising given its smaller model size compared to RoBERTa.



SST-Tiny DistilBERT RoBERTa

Figure 18: Confusion Matrices of the three models for lower length dataset

3.4.3 Analysis

To analyze much more closely, we take the longest sentence, and the shortest sentences from the dataset, and see how these models perform on them respectively.

The longest sentence ***"it 's one of those baseball pictures where the hero is stoic , the wife is patient , the kids are as cute as all get-out and the odds against success are long enough to intimidate , but short enough to make a dream seem possible ."*** is of 48 words, and is correctly classified by SST-Tiny, DistilBERT, but not RoBERTa as shown in Fig. 20.

The shortest sentence "cool ? " is 3 words long, including the spaces, and is correctly classified by SST-Tiny, DistilBERT but not RoBERTa as shown in 19

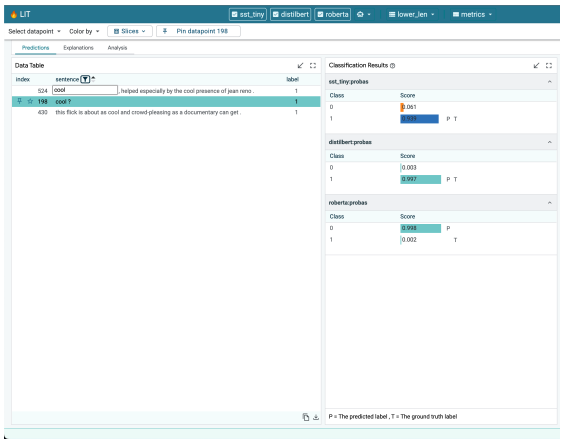


Figure 19: Shortest Sentence in the Dataset

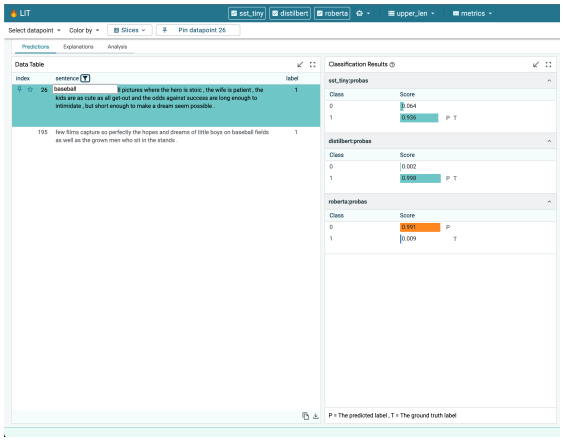


Figure 20: Longest Sentence in the Dataset

This requires saliency and attribution analysis to dig deeper as to why RoBERTa is

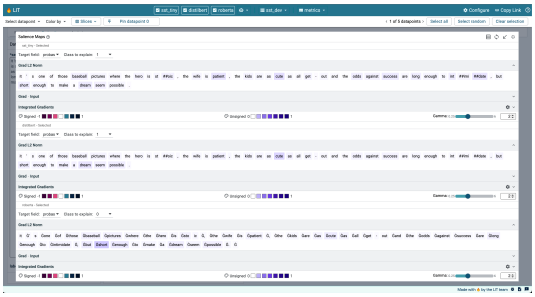
failing on such very long and very short texts.

4 Saliency & Attribution Analysis

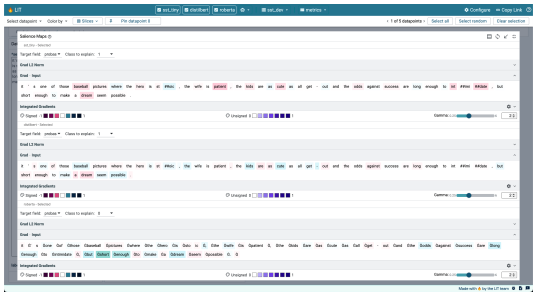
For Saliency & Attribution analysis, let us take the 5 concerning examples from our model analysis section and dig deeper. Below are the 5 concerning sentences that occured during our analysis in the previous section.

4.1 Analyzing Attention in Longer Sentences

Example 1: ***"it 's one of those baseball pictures where the hero is stoic , the wife is patient , the kids are as cute as all get-out and the odds against success are long enough to intimidate , but short enough to make a dream seem possible ."*** (Ground Truth: Positive)



(a) Grad L2 Norm



(b) Grad Input

(c) Integrated Gradients

(d) LIME

Figure 21: Saliency Maps for Example 1

- **Grad L2 Norm**: From the gradient L2 norm analysis, we observe that SST-Tiny places significant attention on tokens such as patient, cute, date, and dream, assigning these words higher weights, which skews the model toward predicting a positive sentiment. Notably, cute receives the highest weight, indicating SST-Tiny's bias toward positive tokens. In contrast, RoBERTa focuses its attention

on the word short, which leads the model to classify the sentiment as negative. DistilBERT's gradient L2 salience is similar to that of SST-Tiny, but it places even more emphasis on cute and comparatively less on other tokens. This suggests that DistilBERT is primarily influenced by stronger sentiment words, making it less sensitive to the context provided by surrounding words.

- **Grad Input**: The gradient input analysis presents a notable contrast to the gradient L2 norm, particularly with how SST-Tiny assigns salience. While the L2 norm highlighted cute as receiving a positive salience, the gradient input reveals that cute is actually being assigned negative salience by SST-Tiny, indicating a potential inconsistency in how the model weighs this word across different analyses. On the other hand, DistilBERT seems to be more consistent in its attention, focusing on contextually relevant words like kids and cute, which aligns better with human intuition for identifying positive sentiment. This suggests that while SST-Tiny may misinterpret certain key words depending on the attribution method used, DistilBERT appears to maintain a more stable and contextually aware focus, which could contribute to its improved accuracy in certain cases. These differences underscore the importance of using multiple interpretability methods to get a more comprehensive understanding of a model's decision-making process.

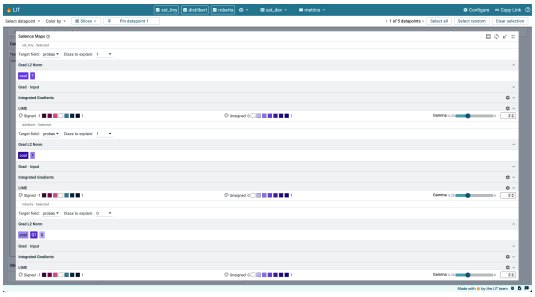
- **Integrated Gradients**: We can see salience for tokens like helping verbs (are, the etc) and stop words (".", "," etc) that are given zero value in Grad Input and Grad L2 Norm are non-zero in integrated gradients showcasing the methods ability to overcome vanishing gradients.

- **LIME**: I was unable to obtain the LIME representation for RoBERTa due to recurring runtime crashes, likely caused by the larger model size. Analyzing the salience maps of SST-Tiny and DistilBERT reveals contrasting interpretations of certain tokens. For example, SST-Tiny assigns a positive salience to hero and a negative salience to stoic, while DistilBERT does the opposite, giving hero a negative salience and stoic a positive salience. This indicates that DistilBERT is better at understanding token meaning based on context. A notable example is the word enough following long—DistilBERT assigns it a negative salience, while SST-Tiny assigns it a positive one. Additionally, LIME operates differently from other attribution methods as it doesn't use tokenizers, splitting sentences based on whole words instead of tokens. This distinction makes it harder to directly compare LIME results to those of other methods.

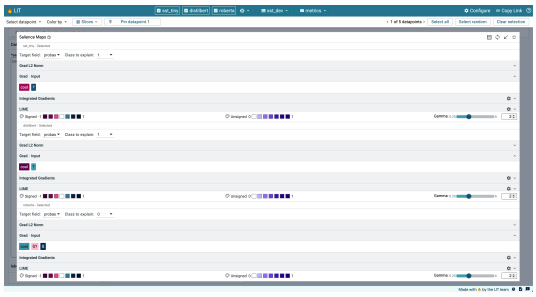
4.2 Analyzing Attention in Shorter Sentences

Example 2: ***"cool ? "*** (Ground Truth: Positive)

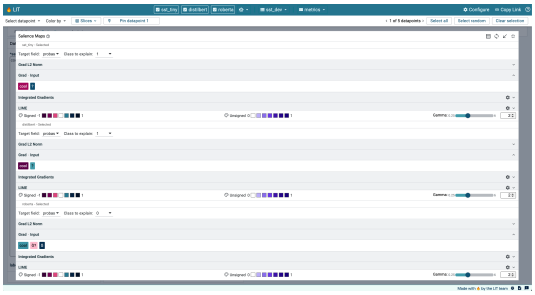
- **Grad L2 Norm**: SST-Tiny appears to assign equal salience to all tokens, showing little differentiation between important words. In contrast, DistilBERT places greater emphasis on the word cool, which is more aligned with the sentence's



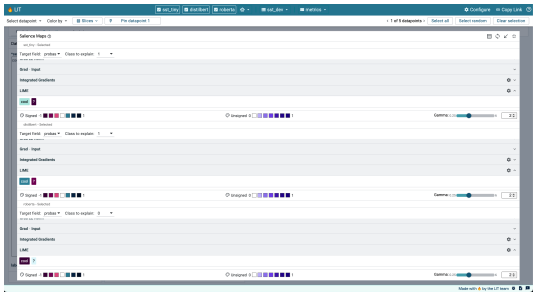
(a) Grad L2 Norm



(b) Grad Input



(c) Integrated Gradients



(d) LIME

Figure 22: Saliency Maps for Example 2

sentiment. Meanwhile, RoBERTa assigns significant salience to the ? token, potentially skewing its interpretation and leading to the misclassification of the sentence as negative. This suggests that RoBERTa's focus on punctuation rather than sentiment-driven words like cool may explain why it flags the sentence with a negative sentiment.

- **Grad Input**: The gradient input analysis presents a contrast to the gradient L2 norm, where cool is assigned negative salience, while the ? symbol is contributing a more positive sentiment effect. However, for RoBERTa, the positive salience of cool actually signals a negative sentiment, as the model is explaining its prediction for label "0" (negative sentiment). In this case, the positive salience of cool outweighs the negative salience of ?, leading to a misclassification, which contradicts the pattern seen in the Grad L2 norm. This suggests that the salience attribution methods provide differing insights into how each token influences the model's decision, highlighting the complexity of the model's interpretation.

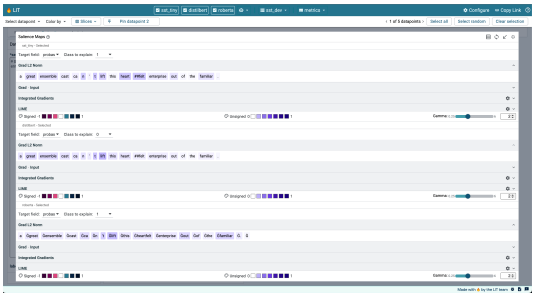
- **Integrated Gradients**: The Integrated Gradients analysis offers a different perspective. For both SST-Tiny and DistilBERT, the word cool is assigned positive salience, indicating that it contributes to the overall positive sentiment of the sentence. However, in RoBERTa, cool is given negative salience, while the ? symbol has positive salience, suggesting that ? plays a role in conveying negative sentiment. This interpretation aligns with human intuition, as a question mark following cool

can imply sarcasm or annoyance, giving the phrase a more negative tone. The contrasting salience attributions across models highlight their varying abilities to capture the subtle nuances of sentiment based on context.

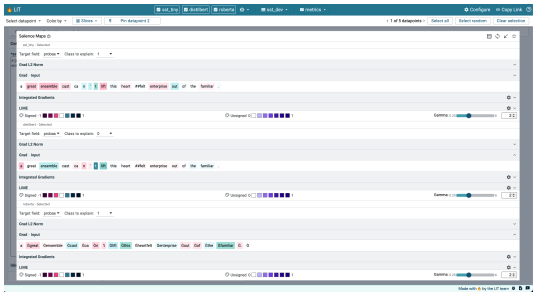
- **LIME**: LIME representation agree with that of integrated gradients only its much more evident with the salience colors being starkly contrasting.

4.3 Large Models aren't always right

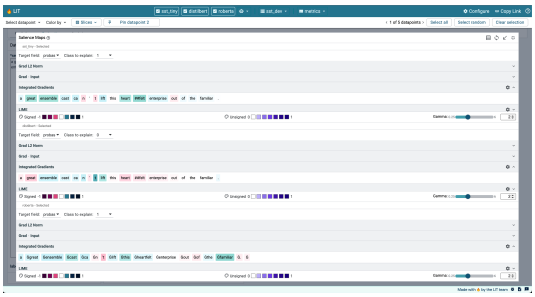
Example 3: ***"a great ensemble cast can't lift this heartfelt enterprise out of the familiar"*** (Ground Truth: Negative)



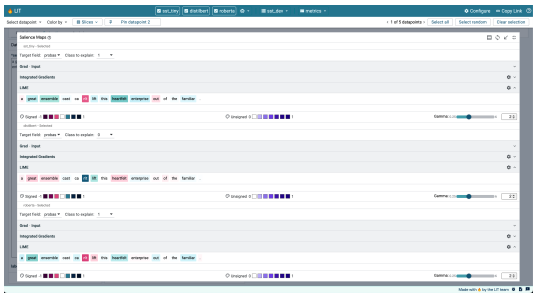
(a) Grad L2 Norm



(b) Grad Input



(c) Integrated Gradients



(d) LIME

Figure 23: Saliency Maps for Example 3

- **Grad L2 Norm**: SST-Tiny assigns higher salience to the tokens heart and felt, which likely biases its decision towards labeling the sentence as positive sentiment. In contrast, DistilBERT and RoBERTa focus more on the token lift, while also giving some attention to great, ensemble, and familiar. Additionally, one issue affecting the models' performance is the tokenization of the word can't, which gets split into sub-tokens, potentially disrupting the models' ability to fully grasp the meaning and sentiment of the sentence. This difference in focus across the models highlights how token salience can influence sentiment classification.
- **Grad Input**: The Grad Input analysis presents a different view from the Grad L2 Norm, particularly for SST-Tiny. Here, out is assigned positive salience, while

heart receives negative salience, and "felt" is given almost zero importance—marking a stark contrast from Grad L2 Norm. However, in the case of DistilBERT, both methods agree, with positive salience assigned to lift and ensemble, supporting the negative sentiment. Similarly, for RoBERTa, both Grad Input and Grad L2 Norm align, as familiar is consistently given positive salience, indicating its contribution to the model's prediction. These differences highlight how attribution methods can vary in their interpretation of token importance.

- **Integrated Gradients**: The Integrated Gradients analysis presents a contrast to the Grad Input, while showing some alignment with the Grad L2 Norm. It is noteworthy that DistilBERT effectively captures the human interpretation of sentiment in this instance by assigning negative salience to parts of the sentence that convey irony and sarcasm, alongside positive salience for the initial buildup. In contrast, RoBERTa struggles to recognize this nuance, as it assigns higher salience to the word familiar. This suggests that RoBERTa may prioritize strong, familiar words over the contextual sentiment, potentially leading to misinterpretations of irony and sarcasm in the text.

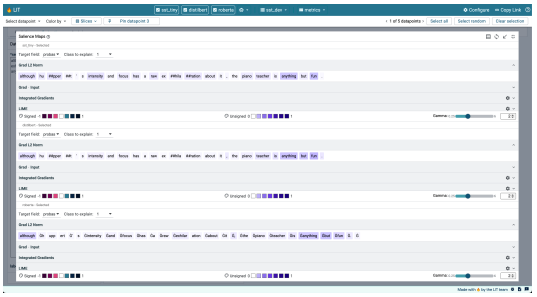
- **LIME**: LIME representation agree with that of integrated gradients only its much more evident with the salience colors being starkly contrasting.

4.4 Long term negations are harder to model

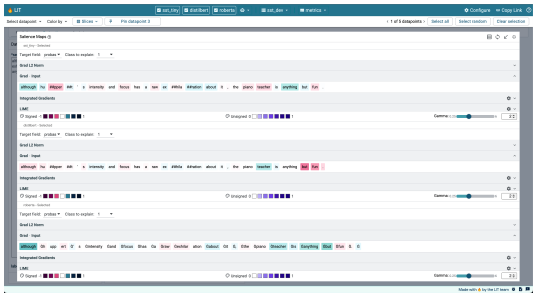
Example 4: ***"although huppert 's intensity and focus has a raw exhilaration about it , the piano teacher is anything but fun ."***

- **Grad L2 Norm**: This example is particularly intriguing because none of the models successfully predict the correct sentiment. The Grad L2 Norm assigns greater salience to words such as anything, fun, and although, which collectively skew the sentiment toward a positive interpretation. This highlights the challenges the models face in accurately capturing the intended sentiment of the sentence.

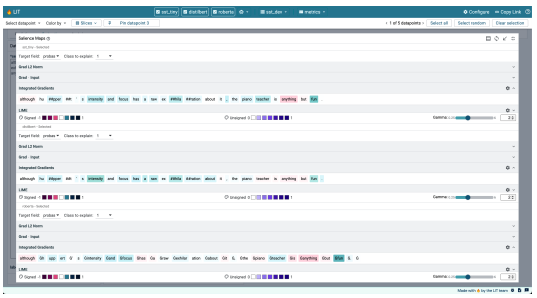
- **Grad Input**: The Grad Input analysis provides a more nuanced examination of the tokens, revealing that words like exhilaration and intensity receive positive salience. Interestingly, the word fun is assigned negative salience across all models, despite its inherently positive connotation. This suggests that the models are attuned to contextual cues, as fun is preceded by the word but, which typically indicates a contrasting sentiment. Additionally, DistilBERT interprets the word although as contributing to a negative sentiment, in contrast to SST-Tiny and RoBERTa, which do not share this perspective. SST-Tiny also uniquely assigns negative salience to the word teacher, while DistilBERT and RoBERTa do not, highlighting the varying interpretations among the models. Overall, the predominance of negative salience in the Grad Input raises questions about how the models arrived at a positive sentiment prediction.



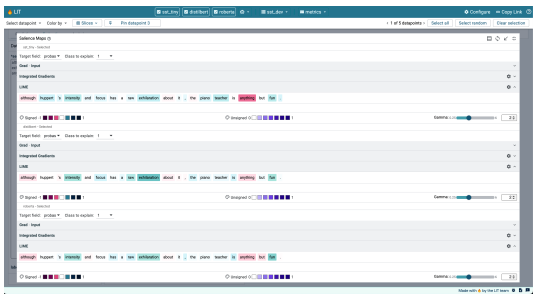
(a) Grad L2 Norm



(b) Grad Input



(c) Integrated Gradients



(d) LIME

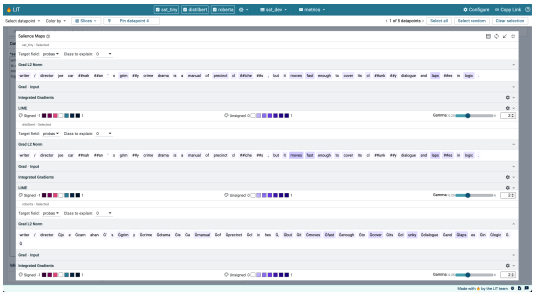
Figure 24: Saliency Maps for Example 4

- **Integrated Gradients**: Integrated Gradients provide a clearer perspective that aligns more closely with human intuition. In this analysis, words with positive connotations receive positive salience, while contrasting terms such as but, although, and anything are assigned negative salience. This alignment suggests that the model is effectively capturing the nuances of sentiment in the text.
- **LIME**: LIME representation agree with that of integrated gradients only its much more evident with the salience colors being starkly contrasting.

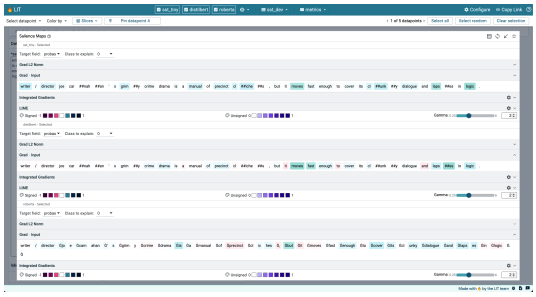
4.5 Capturing long term positive sentiment

Example 5: ***"writer/director joe carnahan 's grimy crime drama is a manual of precinct cliches , but it moves fast enough to cover its clunky dialogue and lapses in logic ."*** (Ground Truth: Positive)

- **Grad L2 Norm**: The saliency scores from the Grad L2 Norm demonstrate a notable consistency across all models, focusing primarily on words such as moves, fast, enough, laps, and covers. This commonality suggests that the models share a similar understanding of the importance of these terms in shaping the overall sentiment of the text.
- **Grad Input**: In this instance, Grad Input aligns with Grad L2 Norm, assigning positive saliency scores to many of the same words that Grad L2 Norm highlights.

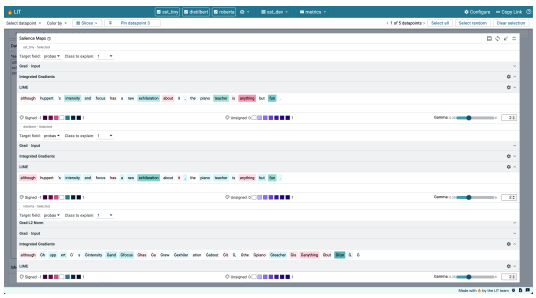


(a) Grad L2 Norm



(b) Grad Input

(c) Integrated Gradients



(d) LIME

Figure 25: Saliency Maps for Example 5

While the sentence predominantly contains words with negative connotations, the overall sentiment is positive. The models struggle to recognize "fast enough" as a positive expression, indicating their difficulty in capturing the nuance of counter-positive statements, particularly over longer contexts.

- **Integrated Gradients**: Integrated Gradients reveal that the RoBERTa model successfully assigns negative salience to the phrase "fast enough," indicating its ability to recognize this positive sentiment. However, the sentence is predominantly filled with words that carry positive salience, suggesting an overall positive sentiment. This discrepancy highlights a limitation of the models: despite the presence of positive expressions, the abundance of negative sentiment words can overshadow the overall sentiment of the sentence.

- **LIME**: LIME representation agree with that of integrated gradients only its much more evident with the salience colors being starkly contrasting. It also shows that even DistilBERT is able to capture this long term positive sentiment but is only capped by the amount of negative words in the sentence.

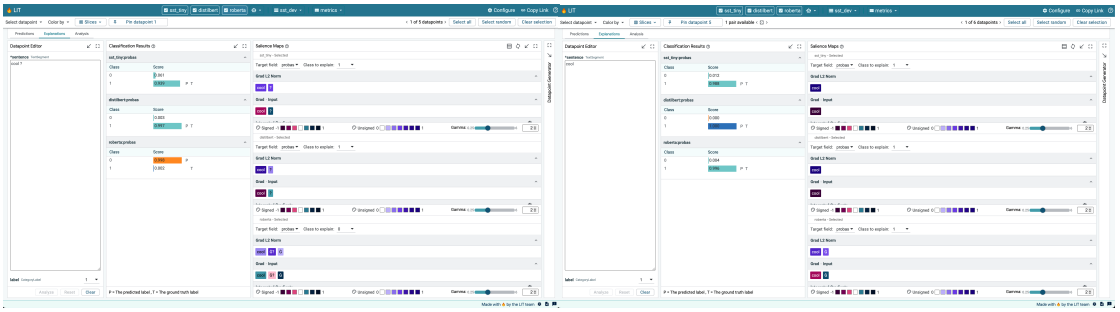
The analysis of model attention reveals a fascinating interplay between the attention mechanisms of different models and human intuition regarding important words in sentences. For instance, when evaluating sentiment, humans often focus on specific contextual clues that signal positivity or negativity. Models like DistilBERT have shown

an ability to align more closely with human sentiment interpretation by assigning appropriate salience to contextually relevant words, such as recognizing sarcasm or irony. In contrast, other models, such as SST-Tiny and RoBERTa, may misinterpret sentiment due to an overreliance on strong lexical cues without adequately considering contextual nuance. For example, the presence of words like "but" or "although" may lead these models to assign negative sentiment to phrases that a human reader would interpret as positive in context. Overall, while some models demonstrate a degree of alignment with human intuition by focusing on the right words, others reveal limitations in capturing the subtleties of language, emphasizing the need for continued refinement in sentiment analysis approaches.

5 Counterfactual Analysis

We can analyze the counterfactuals of the problematic sentences we discussed above for proper context.

5.1 Effect of punctuations



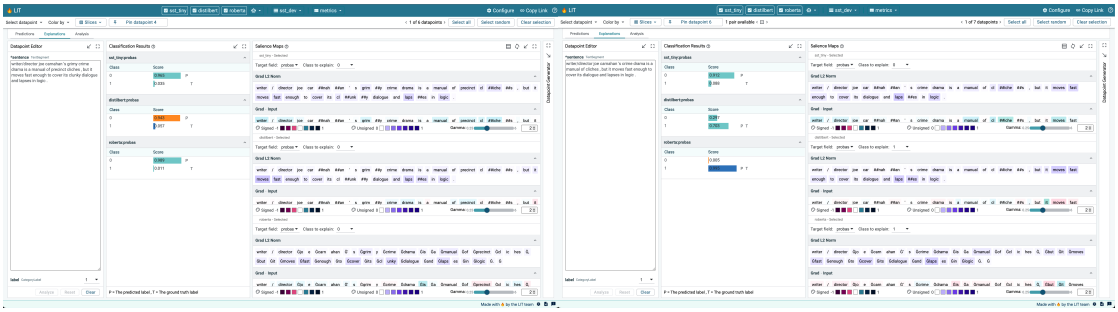
Counterfactual on Example 2 Result

Figure 26: Datapoint Editing in LIT to create counterfactual

From Fig. **??**, the datapoint edit (***"cool ?"*** → ***"cool "***), we can see that the RoBERTa model is heavily influenced by punctuations like "?". After removing the "?" from the shortest sentence, the prediction moved to a positive sentiment as expected.

5.2 Effect of abundance of negative words

From Fig. 27, the datapoint edit (***"writer/director joe carnahan 's grimy crime drama is a manual of precinct cliches , but it moves fast enough to cover its clunky dialogue and lapses in logic ."*** → ***"writer/director joe carnahan 's crime drama is a manual of cliches , but it moves fast enough to cover its dialogue and lapses in logic ."***), we can see that the models DistilBERT and RoBERTa were actually able to capture the long term positive context, but were influenced by the abundance of negative sentiment words in the sentence. After removing



Counterfactual on Example 5 Result

Figure 27: Datapoint Editing in LIT to create counterfactual

words like "grimy", "clunky", "precinct", the models were able to capture the positive sentiment.

6 Discussion on LIT

6.1 Insights Gained About the Model's Behavior and Decision-Making Process

Through the analysis of the models' behaviors, we gained valuable insights into their decision-making processes, particularly in sentiment analysis. For instance, DistilBERT often demonstrated a stronger alignment with human intuition, effectively capturing the nuances of contextual cues such as sarcasm or irony. This model's ability to focus on words that convey sentiment—rather than merely relying on strong lexical indicators—highlights its strength in understanding context. Conversely, models like SST-Tiny and RoBERTa exhibited limitations by overemphasizing certain positive or negative words without adequately considering the overall sentiment conveyed by the context. These discrepancies suggest that while some models can grasp the subtleties of language, others might misclassify sentiments due to their reliance on specific tokens, emphasizing the importance of enhancing contextual understanding in future model training.

6.2 Strengths and Limitations of Using LIT for Model Interpretability

The LIT (Language Interpretability Tool) framework offers significant strengths in enhancing model interpretability. It provides a user-friendly interface for visualizing model behavior, allowing researchers and practitioners to explore various aspects of model performance, including attention weights and token salience. This exploration facilitates a deeper understanding of how models derive their predictions, making it easier to identify areas for improvement.

However, LIT also has limitations. Its reliance on specific attribution methods may lead to a narrow interpretation of model behavior. For example, while Grad L2 Norm and Integrated Gradients offer insights into salience, they might not fully capture the

model's reasoning, especially in complex sentences with nuanced meanings. Additionally, LIT's performance can be constrained by the underlying model's architecture and size, leading to challenges in interpretation when dealing with larger models.

6.3 Strengths and Limitations of Attribution Methods and Counterfactual Generators for Explainability

Attribution methods like Integrated Gradients and Grad Input are powerful tools for explaining model predictions. They provide insights into which tokens contribute positively or negatively to a model's decision, allowing users to trace back the rationale behind specific outputs. This level of granularity aids in diagnosing model errors and understanding how the model interprets language. Additionally LIME offers a better insight into the saliency maps because it is independent of the output gradients, making it easier to integrate newer models.

However, these methods also have limitations. For instance, they may struggle with sentences that contain contradictions or sarcasm, leading to misinterpretations of sentiment. Additionally, the choice of attribution method can significantly impact the explanations provided as some methods may focus too heavily on individual tokens without considering their contextual relationships. Counterfactual generators can enhance explainability by demonstrating how small changes in input can affect output, but they also rely on the model's ability to generalize effectively, which can vary across different architectures. Another good enhancement in this direction could be to extract semantically meaningful subsets from using counterfactual generators.

6.4 Improvements

Semantically Charged Counterfactual Generation: We could utilize the counterfactual generators to alter the To improve model performance in sentiment analysis, training on a more diverse dataset that includes nuanced examples—such as sarcastic or ironic statements—could enhance the model's ability to understand context better. Additionally, incorporating attention mechanisms that prioritize contextual cues over isolated token salience could improve prediction accuracy.

Hybrid Attribution Techniques: Developing hybrid attribution methods that combine the strengths of various approaches (e.g., integrating gradient-based methods with contextual embeddings) could provide more comprehensive insights into model behavior. These refined techniques should aim to capture the interplay between individual token contributions and overall sentence context, leading to more accurate and human-aligned interpretations of model predictions.