Finetuning Analysis Report

1. Introduction

This report provides an in-depth analysis of the finetuning job and an dataset analysis. Using the finetuned model and the dataset, we generated text embeddings and applied various techniques to visualize and interpret the data. The analysis includes t-SNE visualization, UMAP projection with KMeans clustering, anomaly detection, and topic identification using KMeans clustering. Additionally, we generated a word cloud to visualize the most frequent words in the dataset. The report aims to provide insights into the underlying structure of the text data and identify potential outliers or anomalies. The following sections present the results of the analysis along with visualizations and interpretations.

2A. Training and Validation Loss Improvements

The table below shows the percentage improvement in training and validation losses over the course of training.

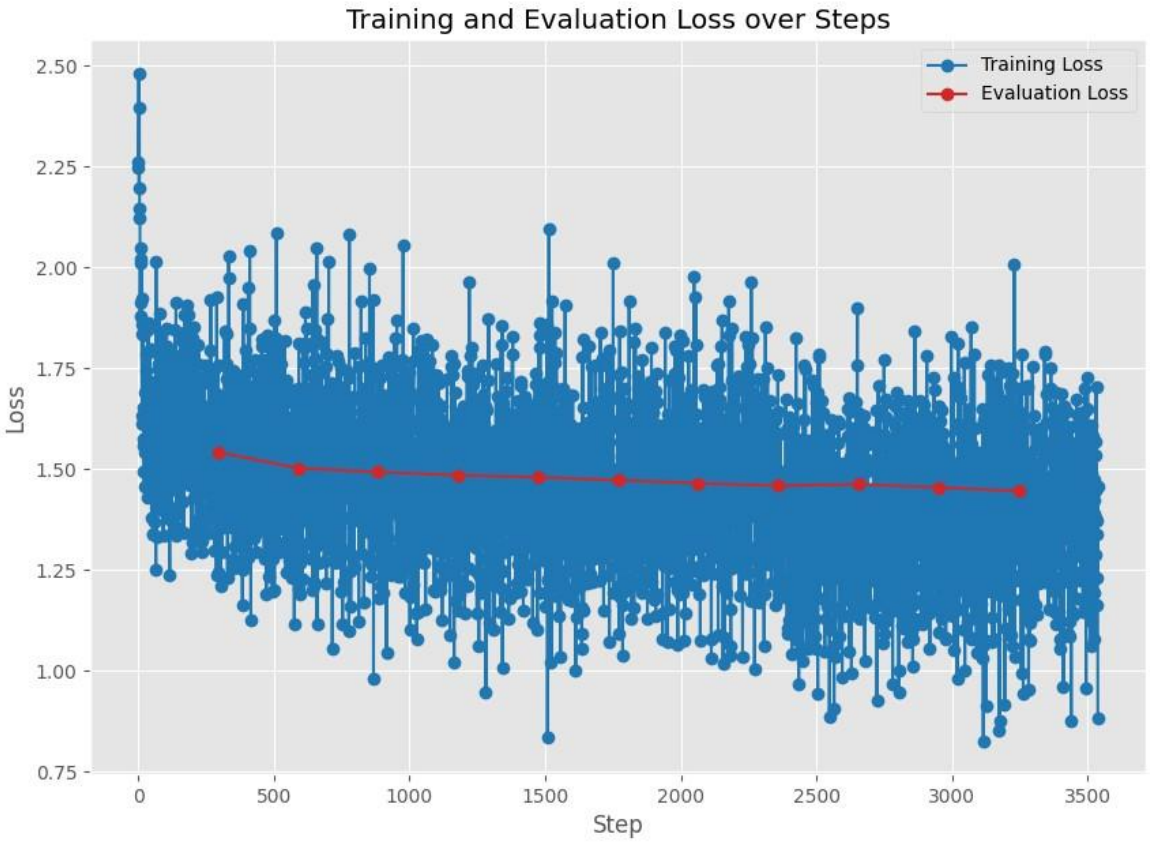
Loss Improvement Metrics

Metric	Value
Train Loss Improvement	60.92%
Validation Loss Improvement	6.25%

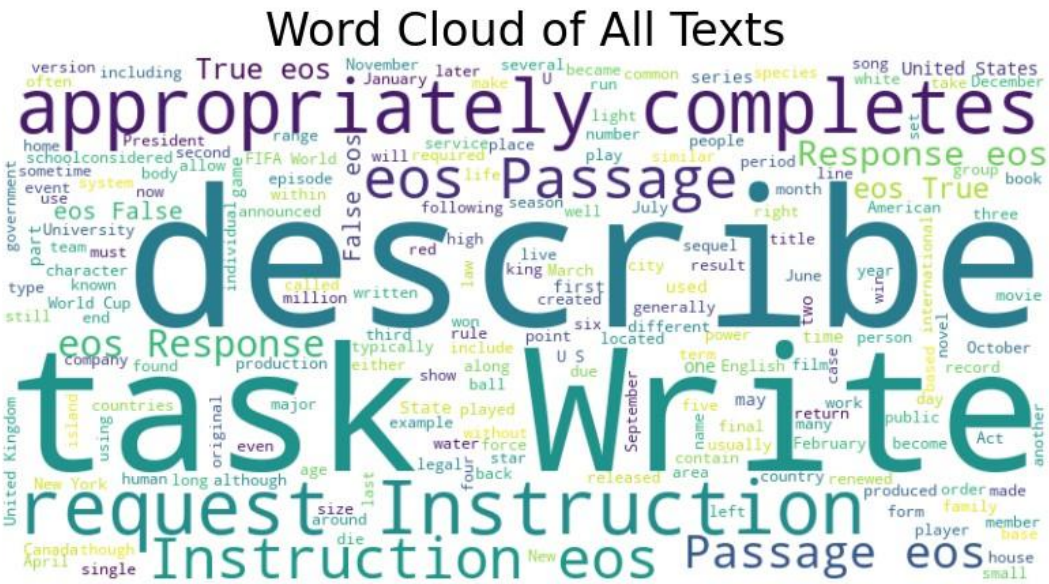
2B. Training Loss Over Steps

The learning curve has not plateaued, indicating that the model was still making progress in learning from the training data. The percentage change in training loss over the last 5 steps is approximately -35.68%, suggesting that further training could potentially lead to improvements.

Overall a percentage difference of 60.92% in the training loss was observed. The training loss plot is crucial for understanding the model's learning progress over time. A downward trend in the loss indicates that the model is learning and improving its predictions. If the last few evaluation losses are relatively constant, it suggests that the model's learning has plateaued, indicating that the model may have reached its learning capacity on the given dataset. However, if the evaluation loss continues to decrease, even subtly, it could mean that the model may benefit from additional training epochs, as there's potential for further convergence. In practice, a balance must be struck to avoid overfitting, where the model learns the training data too well, including its noise and outliers, which can harm generalization to new data.



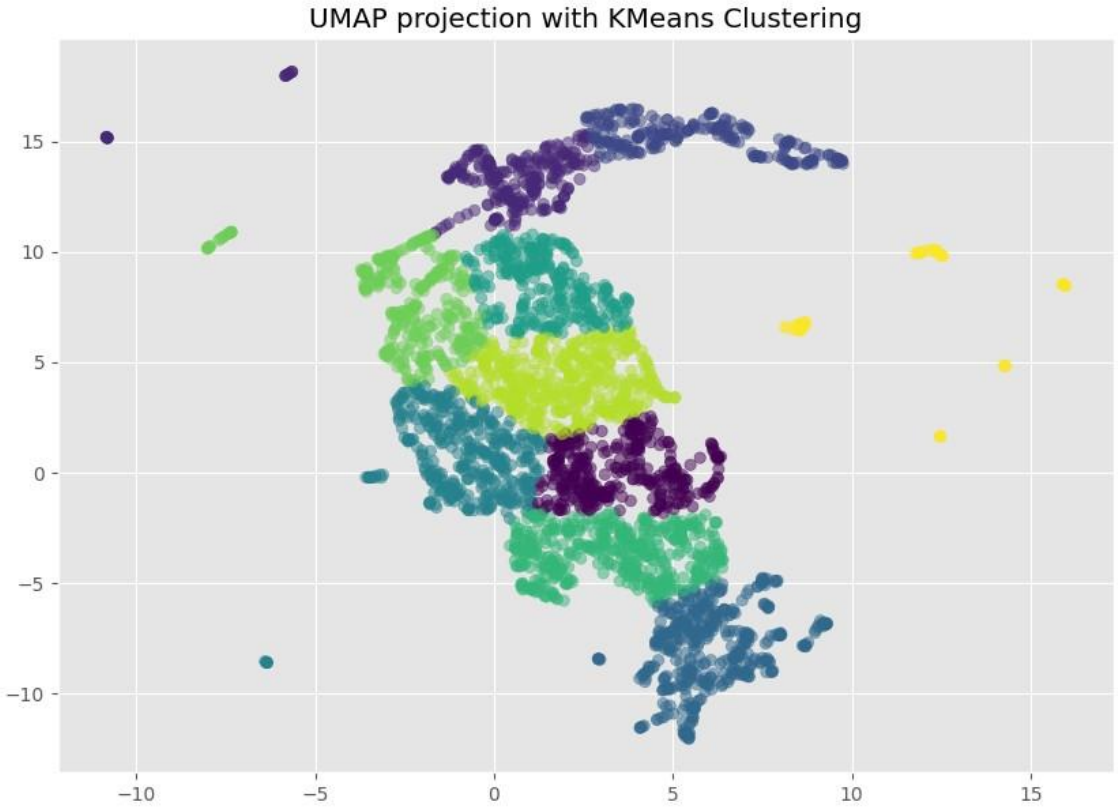
3. Word Cloud of Text Data



The word cloud visualization provides a quick and intuitive way to understand the most prominent terms and themes present in the dataset. The size of each word in the cloud corresponds to its frequency across the text corpus; larger words indicate higher frequency. Such a visualization allows us to gauge the focus and recurring topics in the dataset at a glance.

In reviewing this word cloud, notice which terms are most dominant. These terms often give insight into the overarching subject matter or discourse within the data. It's also useful for identifying any unexpected or unusual words that may warrant further investigation. In some cases, a word cloud can reveal biases, commonalities, or trends that might not be immediately evident from a cursory reading of the texts.

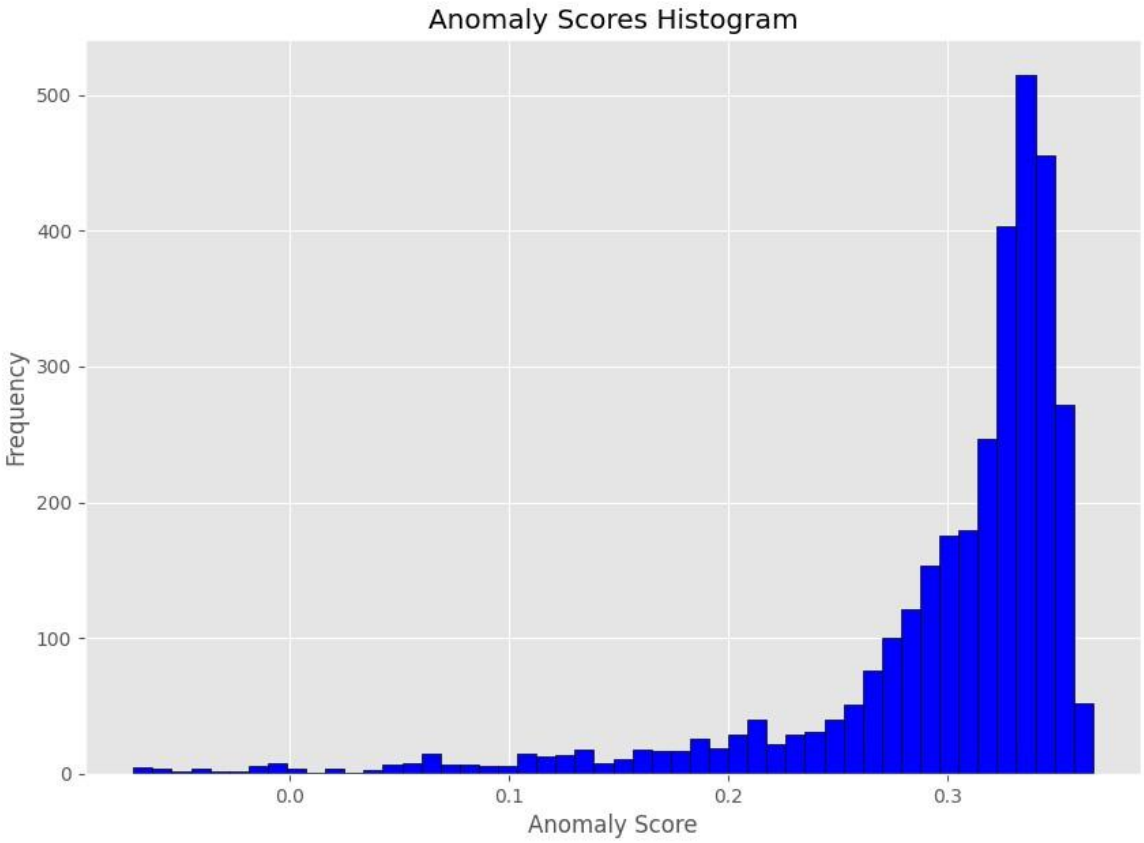
4. UMAP with KMeans Clustering



UMAP, combined with KMeans clustering, identifies distinct topics or themes within the data. Each color in the plot represents a different cluster, suggesting groupings of texts with similar content.

6. Anomaly Scores Distribution

This histogram shows the distribution of anomaly scores given by the Isolation Forest method, this method is used to identify the severity of anomalies in the dataset. Anomalies are data points that significantly differ from the norm, potentially indicating errors or unique patterns. if the distribution is skewed to the right, it indicates a higher number of anomalies. if the distribution is uniform, it indicates a lower number of anomalies.



5. Outliers Detection

Outliers detection is a critical step in data analysis, especially when dealing with large text datasets. Outliers can significantly influence the outcomes of statistical analyses and predictive modeling, leading to skewed results. In this context, outliers are unusual or atypical texts that deviate markedly from the majority of the data. Identifying these outliers helps in understanding the data's underlying structure and possibly uncovering rare but insightful patterns or anomalies that could be of interest. For this analysis we used the Isolation Forest algorithm, a SOTA anomaly detection technique, to compute anomaly scores for each text. These were then applied to systematically identify outliers. This method excels in detecting data points that diverge from the norm, highlighting texts that exhibit unique characteristics compared to the typical patterns observed in the dataset. The following texts are identified as outliers:

1. Below is an instruction that describes a task. Write a response that appropriately completes the request.

###Instruction:<eos>is there a season 9 of pretty little liars<eos>

###Passage:<eos>After an initial order of 10 episodes, ABC Family ordered an additional 12 episodes for season one on Ju...

2. Below is an instruction that describes a task. Write a response that appropriately completes the request. ###Instruction:<eos>will there be a season 4 of game shakers<eos>

###Passage:<eos>When ordered to series in early 2015, it was planned that the first season would consist of 26 episodes. T...

3. Below is an instruction that describes a task. Write a response that appropriately completes the request.

###Instruction:<eos>is love child based on a true story<eos>

###Passage:<eos>The program was created by Sarah Lambert and was first broadcast on the Nine Network on 17 February 2014. The p...

4. Below is an instruction that describes a task. Write a response that appropriately completes the request.

###Instruction:<eos>is there a 3rd season of berlin station<eos>

###Passage:<eos>A ten-episode first season premiered on Epix on October 16, 2016. On November 17, 2016, Epix renewed Berlin...

5. Below is an instruction that describes a task. Write a response that appropriately completes the request.

###Instruction:<eos>is there going to be a season 6 of spirit<eos>

###Passage:<eos>Six episodes of the first season premiered on May 5, 2017. The series was renewed for a second season and...

6. Below is an instruction that describes a task. Write a response that appropriately completes the request.

###Instruction:<eos>is it possible to give birth to twins with different fathers<eos>

###Passage:<eos>Superfecundation is the fertilization of two or more ova from the same cycle by sperm ...

7. Below is an instruction that describes a task. Write a response that appropriately completes the request.

###Instruction:<eos>can a person have twins with different fathers<eos>

###Passage:<eos>Superfecundation is the fertilization of two or more ova from the same cycle by sperm from separate ...

8. Below is an instruction that describes a task. Write a response that appropriately completes the request.

###Instruction:<eos>is there going to be a 3rd series of 800 words<eos>

###Passage:<eos>On 19 October 2015, the Seven Network and South Pacific Pictures renewed the show for a second seaso...

9. Below is an instruction that describes a task. Write a response that appropriately completes the request.

###Instruction:<eos>is the tv show taken still on the air<eos>

###Passage:<eos>Taken is a crime drama series based on the film trilogy of the same name. The series acts as a modern-day ori...

10. Below is an instruction that describes a task. Write a response that appropriately completes the request.

###Instruction:<eos>is there a state income tax in louisiana<eos>

###Passage:<eos>The rest of the century balanced new taxes with abolitions: Delaware levied a tax on several classes of in...

11. Below is an instruction that describes a task. Write a response that appropriately completes the request.

###Instruction:<eos>can you get pregnant with twins by two different fathers<eos>

###Passage:<eos>Superfecundation is the fertilization of two or more ova from the same cycle by sperm from...

12. Below is an instruction that describes a task. Write a response that appropriately completes the request.

###Instruction:<eos>will there be another season of the affiar<eos>

###Passage:<eos>A 12-episode second season of The Affair premiered on October 4, 2015. On December 9, 2015, the series w...

13. Below is an instruction that describes a task. Write a response that appropriately completes the request. ###Instruction:<eos>can you have two babies with different fathers<eos>

###Passage:<eos>Superfecundation is the fertilization of two or more ova from the same cycle by sperm from separate ...
14. Below is an instruction that describes a task. Write a response that appropriately completes the request.

###Instruction:<eos>will there be more episodes of spirit riding free<eos>

###Passage:<eos>Six episodes of the first season premiered on May 5, 2017. The series was renewed for a second se...
15. Below is an instruction that describes a task. Write a response that appropriately completes the request.

###Instruction:<eos>did bruce forsyth do the price is right<eos>

###Passage:<eos>It returned to ITV, as Bruce's Price is Right, from 4 September 1995 to 16 December 2001 with Bruce Forsyth...
16. Below is an instruction that describes a task. Write a response that appropriately completes the request.

###Instruction:<eos>is there a 3rd season of 800 words<eos>

###Passage:<eos>On 19 October 2015, the Seven Network and South Pacific Pictures renewed the show for a second season. It premie...
17. Below is an instruction that describes a task. Write a response that appropriately completes the request.

###Instruction:<eos>is the game show the chase still on the air<eos>

###Passage:<eos>After Fox passed up the opportunity to add the series to its lineup, Game Show Network (GSN), in conjun...
18. Below is an instruction that describes a task. Write a response that appropriately completes the request.

###Instruction:<eos>can you have twins from 2 different dads<eos>

###Passage:<eos>Superfecundation is the fertilization of two or more ova from the same cycle by sperm from separate acts o...
19. Below is an instruction that describes a task. Write a response that appropriately completes the request.

###Instruction:<eos>is it possible to be pregnant with twins by two different fathers<eos>

###Passage:<eos>Superfecundation is the fertilization of two or more ova from the same cycle by s...

20. Below is an instruction that describes a task. Write a response that appropriately completes the request.

###Instruction:<eos>is there a river in ellicott city md<eos>

###Passage:<eos>The town is prone to flooding from the Patapsco River and its tributary the Tiber River. These floods have had...

21. Below is an instruction that describes a task. Write a response that appropriately completes the request.

###Instruction:<eos>is season 6 of voltron the last one<eos>

###Passage:<eos>The first season premiered on Netflix on June 10, 2016, and consisted of 13 episodes. The series has a 78-episo...

22. Below is an instruction that describes a task. Write a response that appropriately completes the request.

###Instruction:<eos>is penn and teller fool us still on<eos>

###Passage:<eos>On 11 August 2015 the series was renewed for a third season by The CW. The third season premiered on 13 July 20...

23. Below is an instruction that describes a task. Write a response that appropriately completes the request.

###Instruction:<eos>is roman j israel movie based on a true story<eos>

###Passage:<eos>On August 25, 2016, it was revealed that Dan Gilroy's next directorial project was Inner City, a lega...

24. Below is an instruction that describes a task. Write a response that appropriately completes the request. ###Instruction:<eos>are they making more episodes of bob's burgers<eos>

###Passage:<eos>On October 7, 2015, Fox renewed the series for the seventh and eighth production cycles. The eighth ...

25. Below is an instruction that describes a task. Write a response that appropriately completes the request.

###Instruction:<eos>does colorado school of mines have a football team<eos>

###Passage:<eos>The Colorado Mines Orediggers football team represents the Colorado School of Mines in the sport...
26. Below is an instruction that describes a task. Write a response that appropriately completes the request.

###Instruction:<eos>is there a season two of the messengers<eos>

###Passage:<eos>The Messengers is an American television series that aired on The CW during the 2014--15 season. The series...
27. Below is an instruction that describes a task. Write a response that appropriately completes the request.

###Instruction:<eos>is there a season 5 of the tudors<eos>

###Passage:<eos>Showtime announced 13 April 2009, that it had renewed the show for a fourth and final season. The network ordered...
28. Below is an instruction that describes a task. Write a response that appropriately completes the request.

###Instruction:<eos>is there a fourth season of doctor doctor<eos>

###Passage:<eos>On 28 September 2016, Nine renewed the program for a second season after just two episodes having been ai...
29. Below is an instruction that describes a task. Write a response that appropriately completes the request.

###Instruction:<eos>can you have twin with two different father<eos>

###Passage:<eos>Superfecundation is the fertilization of two or more ova from the same cycle by sperm from separate act...
30. Below is an instruction that describes a task. Write a response that appropriately completes the request.

###Instruction:<eos>will there be a 6th season of hell on wheels<eos>

###Passage:<eos>Season one (2011--12) began in 1865, shortly after the assassination of Abraham Lincoln, season two (2...
31. Below is an instruction that describes a task. Write a response that appropriately completes the request.

###Instruction:<eos>can i have dual citizenship in us and czech republic<eos>

###Passage:<eos>The citizenship law of the Czech Republic is based on the principles of jus sanguinis or ``rig...
32. Below is an instruction that describes a task. Write a response that appropriately completes the request.

###Instruction:<eos>will there be new episodes of black mirror<eos>

###Passage:<eos>The show premiered for two series on the British television channel Channel 4 on December 2011 and Febru...
33. Below is an instruction that describes a task. Write a response that appropriately completes the request.

###Instruction:<eos>is there season 3 witches of east end<eos>

###Passage:<eos>On November 22, 2013, Lifetime renewed Witches of East End for a second season to consist of 13 episodes, whi...