Thesis Topic:

"Evaluating and Enhancing Figurative Language Understanding in Large Language Models: A Study on Multi-Type Figurative Expressions Using the FLUTE Dataset"

Thesis Progress Report

Summary:

This report outlines the progress made so far in the thesis, which focuses on evaluating and improving the ability of large language models, specifically RoBERTa and BERT, to understand figurative language using the FLUTE dataset. Key stages of the project completed so far are summarized below.

1. Installation and Initialization:

The code is being run in a Python environment on Google Colab Pro, utilizing the A100 GPU for optimal performance. Necessary libraries like transformers, datasets, evaluate, torch, and others were installed to support the project. To ensure reproducibility, a seed-setting function was implemented, and compatibility and efficiency were enhanced by enabling fp16 precision and deterministic behaviors in PyTorch.

2. Dataset Preparation and Splitting:

The FLUTE dataset, which contains labeled examples of figurative language, was uploaded, reviewed, and analyzed. Initial exploratory data analysis (EDA) includes reviewing the dataset structure and analyzing key columns such as "premise," "hypothesis," "label," "explanation," and "type" by extracting unique value counts and examining their distributions was performed.

The dataset was then split into training (80%) and validation (20%) subsets to support a robust evaluation process. The split sizes were verified to ensure proper division.

Additionally, the FLUTE test set was uploaded, reviewed, and checked for consistency with the training and validation sets.

3. Model Training and Evaluation:

Initial Model Setup: Two pre-trained models, RoBERTa-base and BERT-base-uncased, were chosen for the classification tasks. The key steps taken include:

- **Tokenization:** The training and validation datasets were tokenized using the respective model tokenizers, with padding and truncation applied to prepare the data for training.
- **Training Process:** Model training configurations were set, including the learning rate, batch size, number of epochs, and early stopping criteria. Both models were trained on the training dataset, with validation performed at the end of each epoch.
- **Evaluation and Performance Analysis:** The performance of both models was evaluated using various metrics to assess their effectiveness in classifying figurative language. Key aspects of the analysis include:
  - **Overall Accuracy:** The overall classification accuracy for each model on the test set.
  - **Class-Wise Accuracy:** The accuracy for specific figurative language types, such as sarcasm and idioms, was calculated to identify model strengths and weaknesses.
  - **Evaluation Metrics:** BLEU, BLEURT, and BERTScore metrics were used to evaluate how closely the model predictions aligned with the ground truths. Also, precision-recall-F1 scores and confusion matrices were analyzed and visualized to provide deeper insights into model performance.
  - **Misclassification Analysis:** Misclassified examples were identified and analyzed to better understand where the models struggled, providing insights into potential areas for improvement.

4. Figurative Language Evaluation Leveraging Holmes' Approach:

Since the "Sarcasm" label was identified as a challenging figurative language type based on performance analysis, the contrastive pairs for sarcasm were created by pairing sarcastic hypotheses with their literal explanations from the FLUTE training set. A likelihood comparison approach was then applied to determine the model's preference for sarcastic versus literal sentences. This evaluation involved:

- Calculating the mean difference in log-likelihood scores between sarcastic and literal sentences.
- Assessing the accuracy of model preferences, based on whether it favored sarcastic or literal pairs.
- Performing statistical tests (t-test and Wilcoxon signed-rank test) to determine the significance of these preferences.

Additionally, the distribution of log-likelihood scores and confidence levels for both sarcastic and literal pairs was analyzed to gain deeper insights into the model's decision-making process.

5. Data Augmentation:

Based on performance analysis, sarcasm was identified as a particularly challenging figurative language type. To address this, data augmentation techniques were applied to enhance the dataset:

- The nlpaug library, specifically the ContextualWordEmbsAug tool, was used to generate paraphrased versions of the hypotheses.
- Each instance was augmented with three additional paraphrased hypotheses, significantly increasing the dataset's diversity.

Post-Augmentation Steps:

- The augmented data was combined with the original training set.
- Data consistency was verified, and any invalid rows were removed.
- The newly expanded dataset was split into updated training and validation sets to ensure a robust evaluation process.

6. Re-Training and Evaluation with Augmented Data:

The augmented dataset was used to re-train the RoBERTa and BERT models, incorporating the new data to assess the impact of augmentation. Key updates include:

- **Re-Tokenization:** The augmented dataset was re-tokenized to ensure compatibility with the models.

- **Training Process:** Both models were re-trained using the expanded dataset to evaluate how augmentation influenced performance.
- **Performance Analysis:** Metrics such as accuracy, BLEU, BLEURT, and BERTScore F1 were recalculated and compared to the results before augmentation.

7. Results and Observations of the first training of the models:

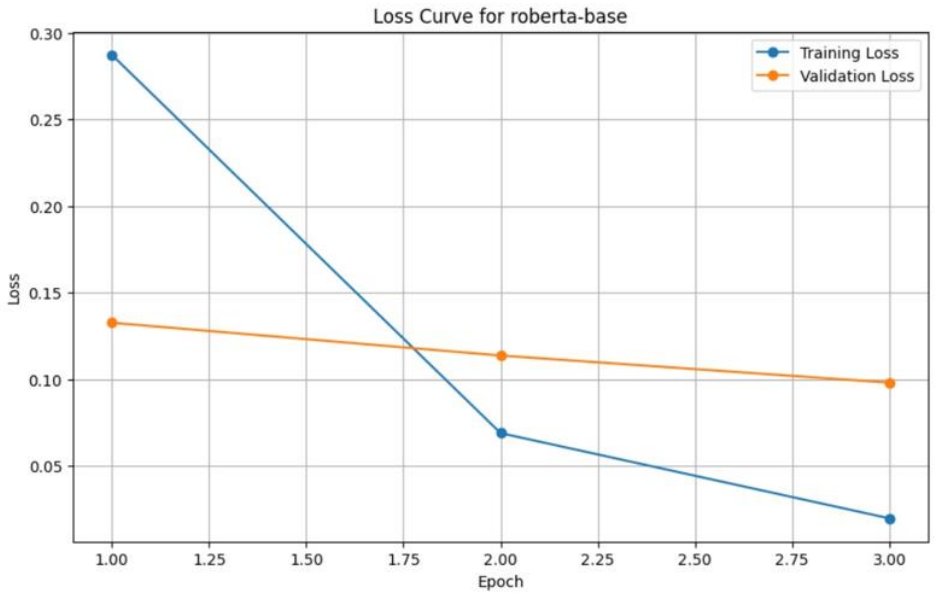
Both RoBERTa-base and BERT-base-uncased, trained and tested for multi-class classification of language figurative expressions on the FLUTE dataset. Below is a detailed analysis of the performance metrics for each model.

Training and Validation Loss: Both models show a steady decrease in training and validation loss across epochs, indicating proper convergence.

RoBERTa-base:

- Training Loss: Starts at 0.2872 and decreases to 0.0197.
- Validation Loss: Starts at 0.1325 and decreases to 0.0870.

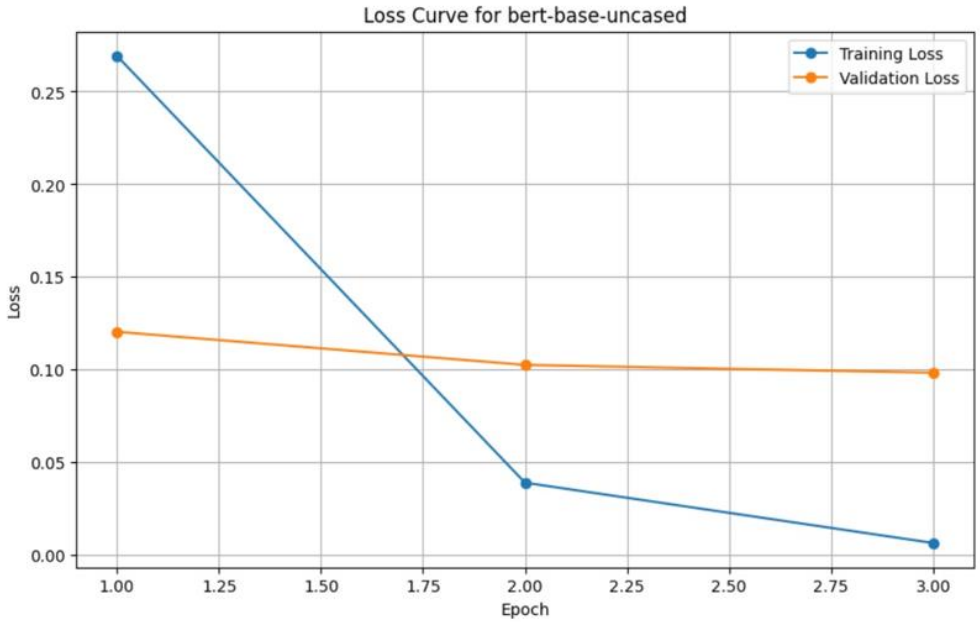
Epoch	Training Loss	Validation Loss
1	No log	0.132586
2	0.287200	0.113678
3	0.068900	0.098025
4	0.019700	0.104894
5	0.019700	0.087026



BERT-base-uncased:

- Training Loss: Starts at 0.2692 and decreases to 0.0066.
- Validation Loss: Starts at 0.1204 and decreases to 0.0869.

Epoch	Training Loss	Validation Loss
1	No log	0.120450
2	0.269200	0.102601
3	0.039100	0.098403
4	0.006600	0.088269
5	0.006600	0.086937



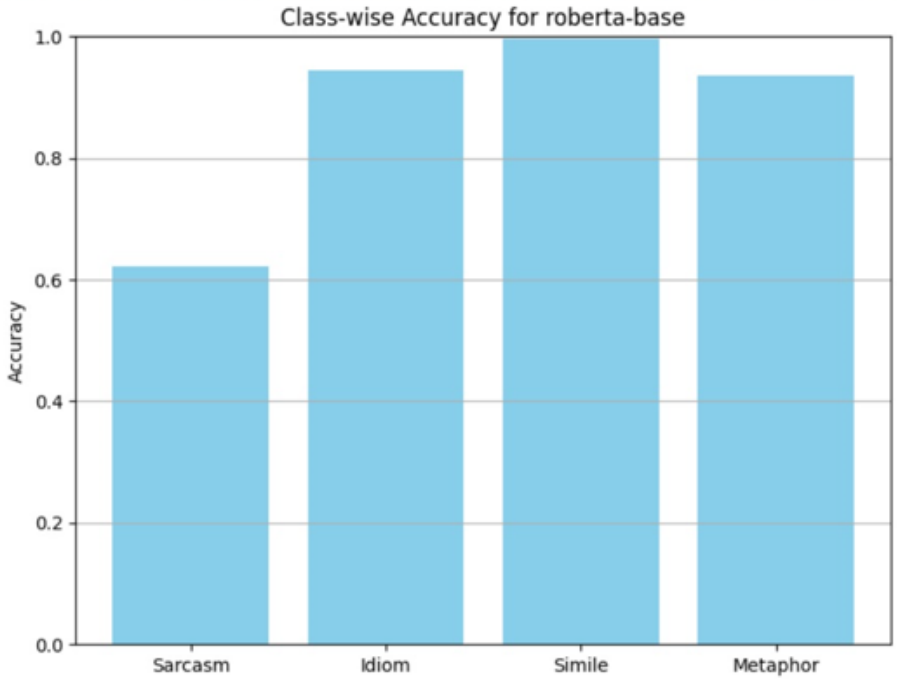
Overall Accuracy: While the difference is minimal, RoBERTa marginally outperforms BERT in accuracy. This indicates that both models are comparably capable of performing the classification task.

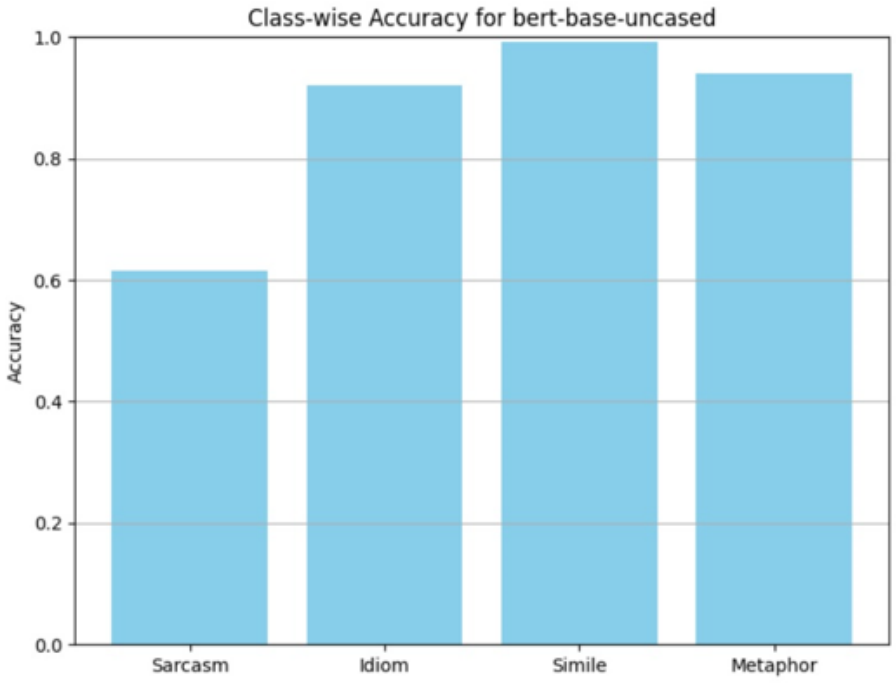
```
Accuracy for roberta-base: 0.79

Accuracy for bert-base-uncased: 0.78
```

Class-wise Accuracy: The accuracies for each class are summarized below:

Class	RoBERTa-base	BERT-base-uncased
Sarcasm	62.1%	61.5%
Idiom	94.4%	92%
Simile	99.6%	99.2%
Metaphor	93.5%	94%





Both models struggle with Sarcasm, achieving the lowest accuracy in this category. This could indicate that sarcasm detection is inherently difficult due to its reliance on context and subtleties in language.

Idiom, Simile, and Metaphor show high accuracies, with BERT slightly outperforming RoBERTa for Metaphor.

Precision, Recall, and F1-Score: The detailed precision, recall, and F1-scores for both models are summarized below:

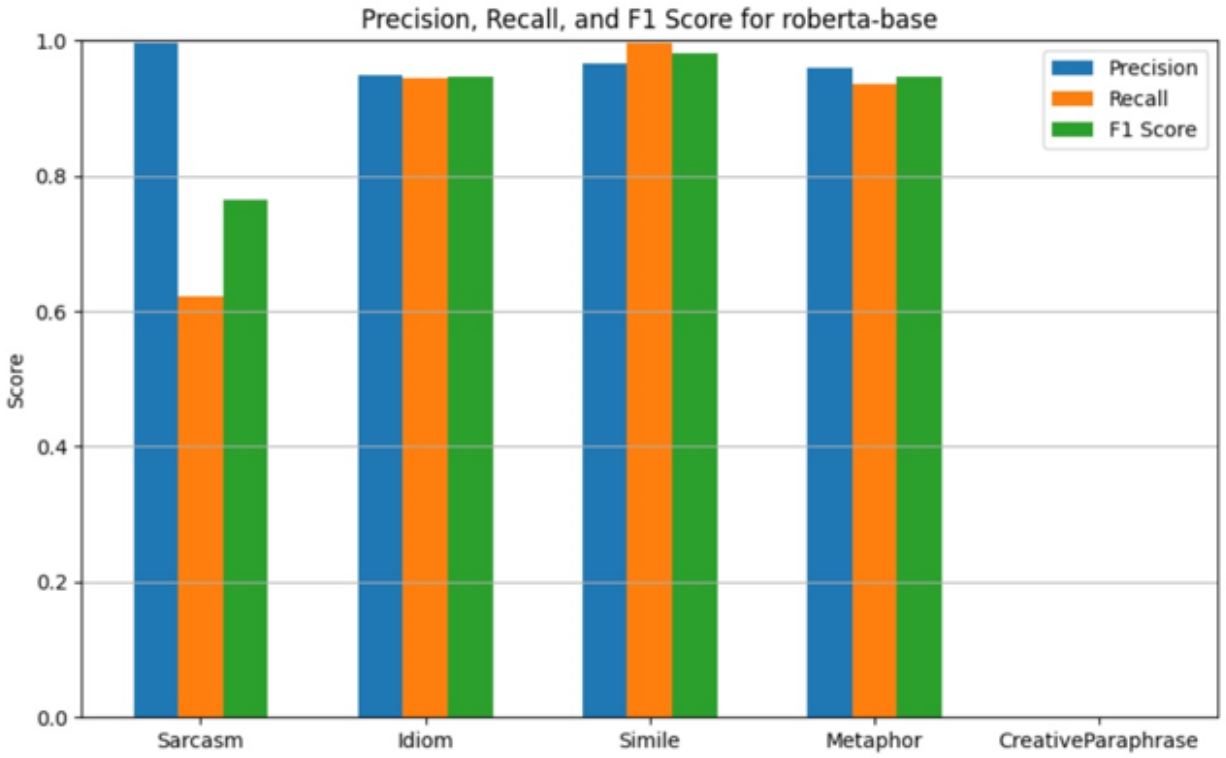
Class	Model	Precision	Recall	F1-Score
Sarcasm	RoBERTa-base	1.00	0.62	0.77
	BERT-base-uncased	0.99	0.61	0.76
Idiom	RoBERTa-base	0.95	0.94	0.95
	BERT-base-uncased	0.95	0.92	0.93
Simile	RoBERTa-base	0.97	1.00	0.98
	BERT-base-uncased	0.98	0.99	0.99
Metaphor	RoBERTa-base	0.96	0.94	0.95
	BERT-base-uncased	0.94	0.94	0.94
Creative Paraphrase	Both Models	-	-	-

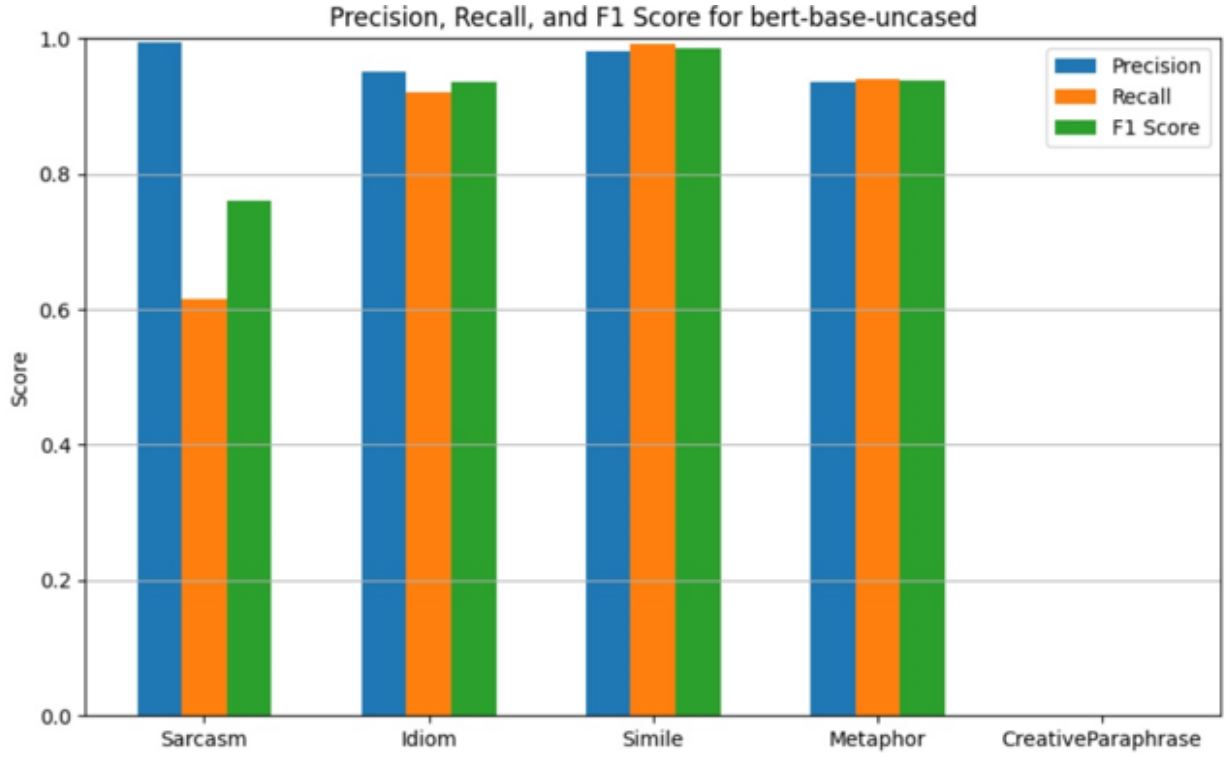
Classification Report for roberta-base:

	precision	recall	f1-score	support
Sarcasm	1.00	0.62	0.77	750
Idiom	0.95	0.94	0.95	250
Simile	0.97	1.00	0.98	250
Metaphor	0.96	0.94	0.95	248
CreativeParaphrase	0.00	0.00	0.00	0
accuracy			0.79	1498
macro avg	0.77	0.70	0.73	1498
weighted avg	0.98	0.79	0.86	1498

Classification Report for bert-base-uncased:

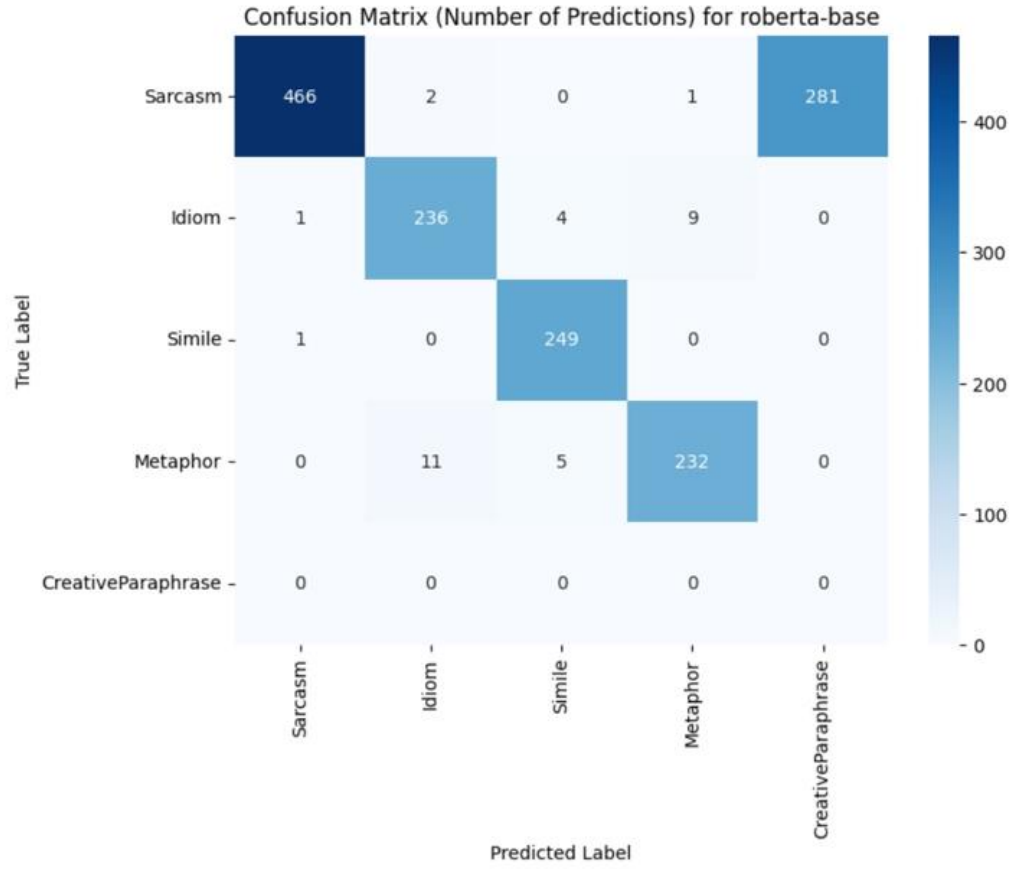
	precision	recall	f1-score	support
Sarcasm	0.99	0.61	0.76	750
Idiom	0.95	0.92	0.93	250
Simile	0.98	0.99	0.99	250
Metaphor	0.94	0.94	0.94	248
CreativeParaphrase	0.00	0.00	0.00	0
accuracy			0.78	1498
macro avg	0.77	0.69	0.72	1498
weighted avg	0.97	0.78	0.86	1498

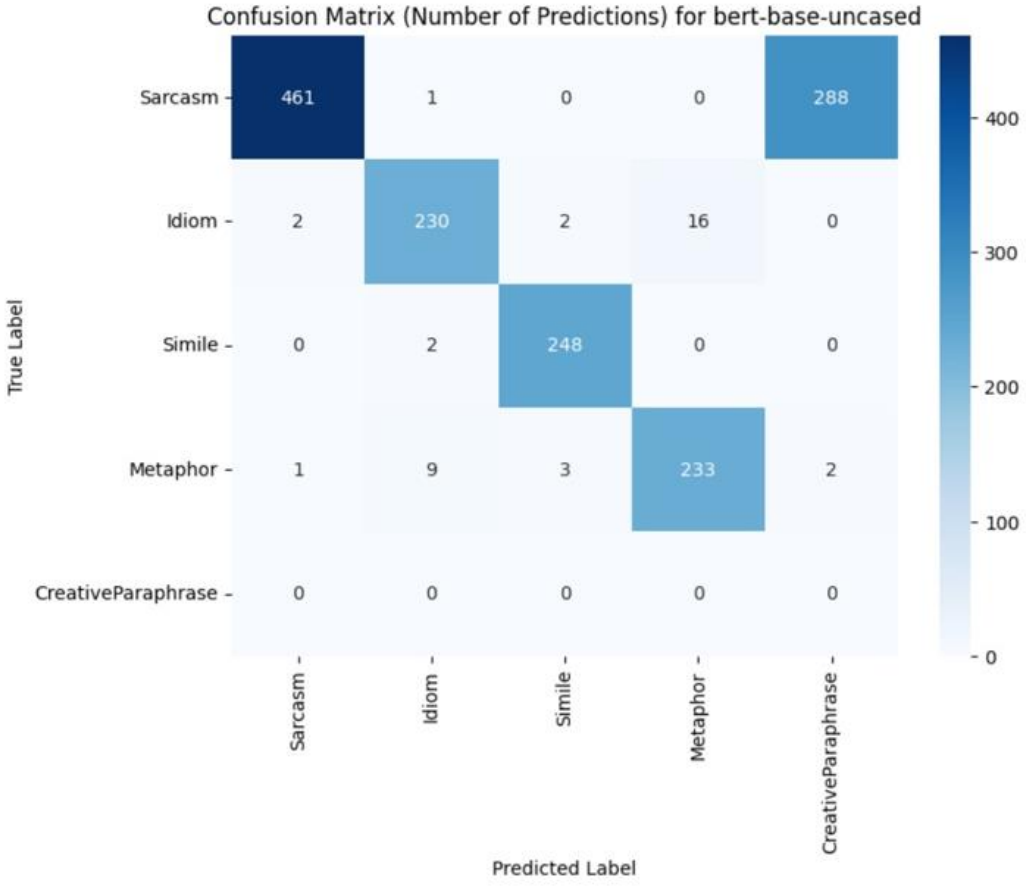




Both models perform equivalently across most metrics and achieve consistently high precision and recall for Idiom, Simile and Metaphor, leading to strong F1-scores. However, Sarcasm recall is low for both models, highlighting a difficulty in correctly identifying all sarcastic samples.

Confusion Matrix: Both models have difficulties distinguishing Sarcasm from Creative Paraphrase, possibly due to overlapping linguistic features.





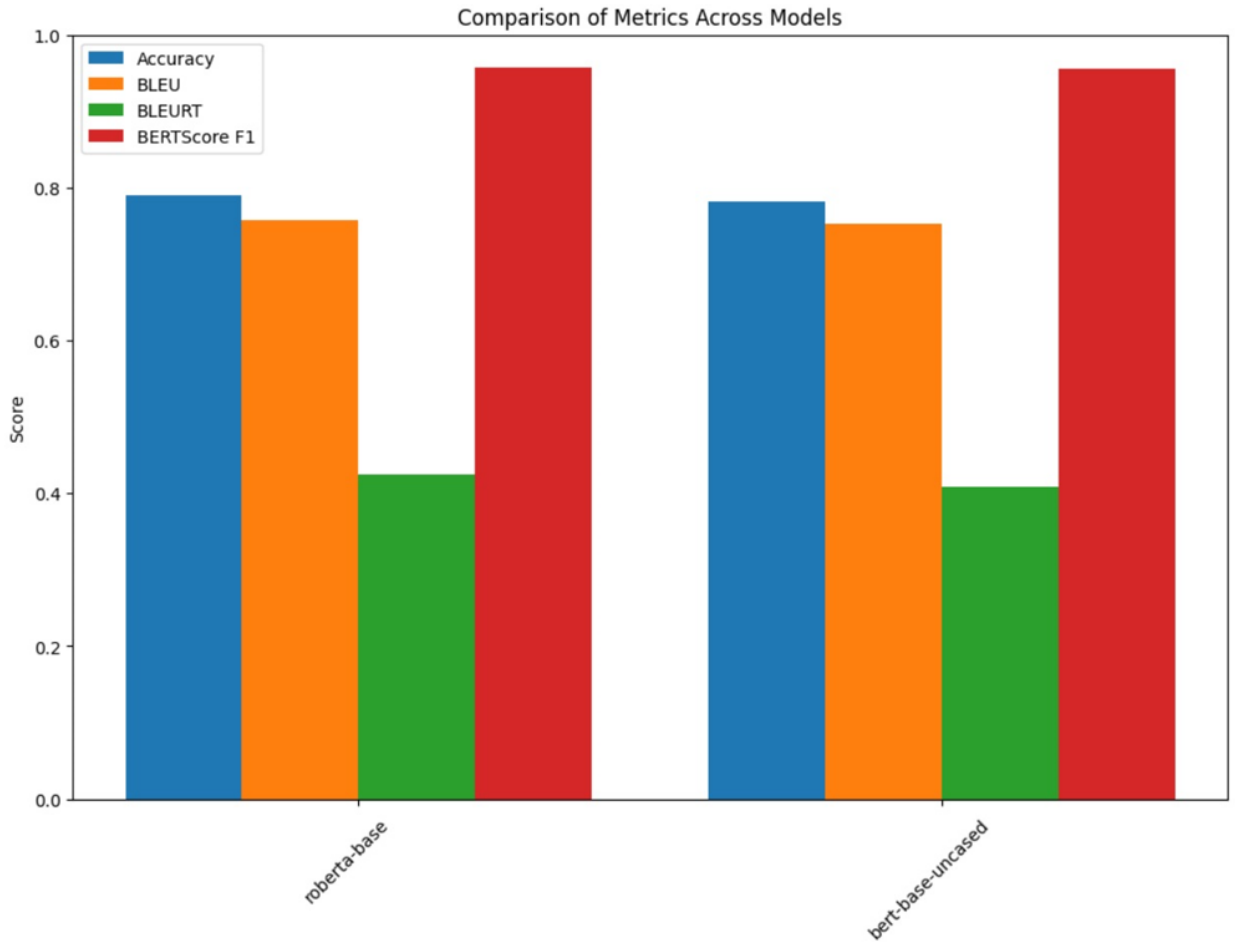
Evaluation Metrics: The following metrics were used to evaluate generative performance:

Metric	RoBERTa-base	BERT-base-uncased
BLEU	0.7580	0.7525
BLEURT	0.4254	0.4091
BERTScore F1	0.9571	0.9557

BLEU: RoBERTa performs slightly better, indicating better n-gram overlap with references.

BLEURT: Both models perform similarly, with RoBERTa again scoring slightly higher.

BERTScore F1: Both models exhibit near-identical semantic similarity performance, with scores exceeding 95%.



Misclassified Examples: Total Misclassified Examples:

- **RoBERTa-base:** 315 out of 1498 (21%).
- **BERT-base-uncased:** 326 out of 1498 (21.8%).

Misclassifications primarily involve Sarcasm being confused with Creative Paraphrase. A common theme in the errors is subtle linguistic ambiguity or context reliance, making it challenging for the models to correctly classify.

Conclusion: Both RoBERTa-base and BERT-base-uncased perform competitively, with RoBERTa exhibiting a slight edge in overall metrics. However, both models struggle with Sarcasm, indicating room for improvement in handling linguistically subtle categories.

8. Results and Observations of the evaluation of figurative language understanding using Holmes' probing task:

Below is a detailed analysis of the evaluation of a pre-trained RoBERTa-base model in distinguishing between sarcastic and literal interpretations of sentences by analyzing their log-likelihood scores. The contrastive task provides insight into the model's ability to capture nuances of figurative language, specifically sarcasm.

```
Mean Difference: 0.3166
Accuracy: 0.4628
Mean Confidence: 1.7093
Sarcastic Mean: -0.8895
Literal Mean: -1.2061
T-test P-value: 0.0012
Wilcoxon P-value: 0.3932
Sarcastic Std Dev: 2.5545
Literal Std Dev: 3.2047
```

Mean Difference: 0.3166

The average log-likelihood difference between sarcastic and literal sentences is 0.3166, suggesting a slight overall preference for sarcastic sentences by the RoBERTa model. Positive values indicate higher likelihood scores for sarcastic sentences, meaning the model tends to rate sarcastic sentences as more plausible than literal ones on average. The small mean difference implies that the model is not strongly biased toward either class but does slightly favor sarcasm.

Accuracy: 0.4628

The model correctly identified sarcastic sentences as more likely 46.28% of the time. Since the ground truth assumes sarcastic sentences should always be preferred, this score is below random chance (50%). The accuracy suggests that the model struggles to distinguish sarcasm effectively. It does not consistently prefer sarcastic over literal sentences, which may indicate limitations in its ability to detect sarcasm nuances.

Mean Confidence: 1.7093

This metric reflects the average absolute difference in likelihood scores between sarcastic and literal sentences, providing a measure of how confidently the model distinguishes between the two. A mean confidence of 1.7093 indicates moderate separation between sarcastic and literal sentences in terms of likelihood. However, this confidence is not necessarily aligned with correctness, as seen from the accuracy metric.

Sarcastic Mean: -0.8895 / Literal Mean: -1.2061

The average log-likelihood scores for sarcastic sentences and literal sentences are -0.8895 and -1.2061, respectively. Since log-likelihoods are negative (logarithms of probabilities), higher values (closer to 0) indicate greater likelihood. The model rates sarcastic sentences as slightly more likely than literal ones, which aligns with the positive mean difference.

T-test P-value: 0.0012

The t-test assesses whether the mean log-likelihood difference between sarcastic and literal sentences is statistically significant. A p-value of 0.0012 indicates strong evidence against the null hypothesis (no difference). The significant p-value confirms that the model’s preference for sarcastic sentences is not due to random chance, though the magnitude of this preference is small.

Wilcoxon P-value: 0.3932

The Wilcoxon signed-rank test is a non-parametric test comparing paired data (sarcastic vs. literal scores). A p-value of 0.3932 suggests no statistically significant difference. This result contrasts with the t-test and may indicate that while there is a slight mean preference, the individual score distributions for sarcastic and literal sentences overlap substantially.

Sarcastic Std Dev: 2.5545 / Literal Std Dev: 3.2047

The standard deviation measures variability in the likelihood scores. Literal sentences have a slightly higher standard deviation, indicating more variation in their scores compared to sarcastic sentences. The larger variability in literal sentence scores could suggest that the model is less consistent in evaluating literal statements, potentially due to the diverse nature of literal explanations compared to sarcastic ones.

Conclusion: The model demonstrates a slight preference for sarcastic sentences, with significant but small differences in log-likelihood scores.

Accuracy below random chance highlights challenges in sarcasm detection, suggesting that the model does not consistently distinguish sarcasm from literal language.

Confidence scores indicate that the model differentiates sarcastic and literal sentences to some extent, but this separation does not necessarily translate into accurate detection.

Statistical significance tests reveal mixed results, with the t-test showing significance but the Wilcoxon test indicating no strong difference.

Variability in scores suggests that the model may handle sarcasm more consistently than literal language, though its overall performance remains suboptimal.

9. Results and Observations of the second training of the models (After data augmentation):

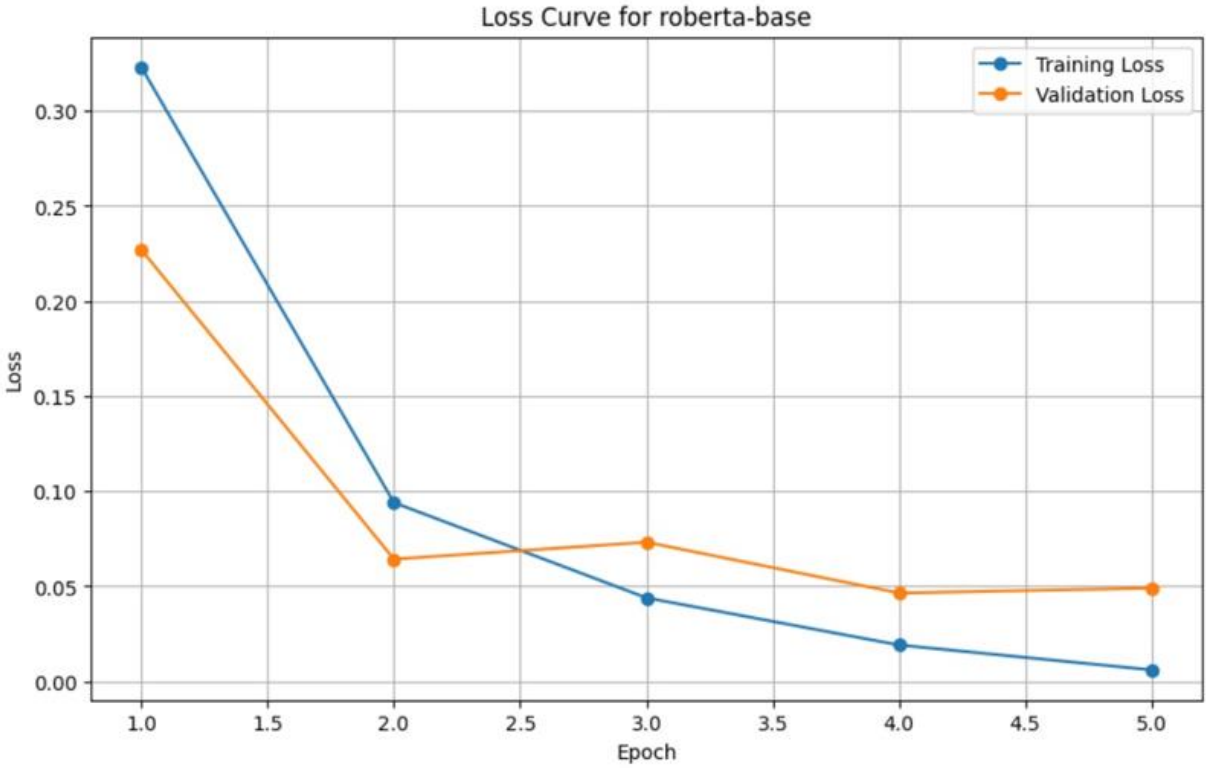
Both RoBERTa-base and BERT-base-uncased, trained and tested for multi-class classification of language figurative expressions on the FLUTE dataset after data augmentation. The focus is to analyze the impact of augmentation on performance metrics and compare the results with the previous experiments conducted without augmentation. Below is a detailed analysis of the performance metrics for each model.

Training and Validation Loss: Both models exhibit a smooth decline in training and validation loss across epochs, indicating effective convergence after data augmentation.

RoBERTa-base:

- Training Loss: Starts at 0.3230 and decreases to 0.0060.
- Validation Loss: Starts at 0.2269 and decreases to 0.0490.

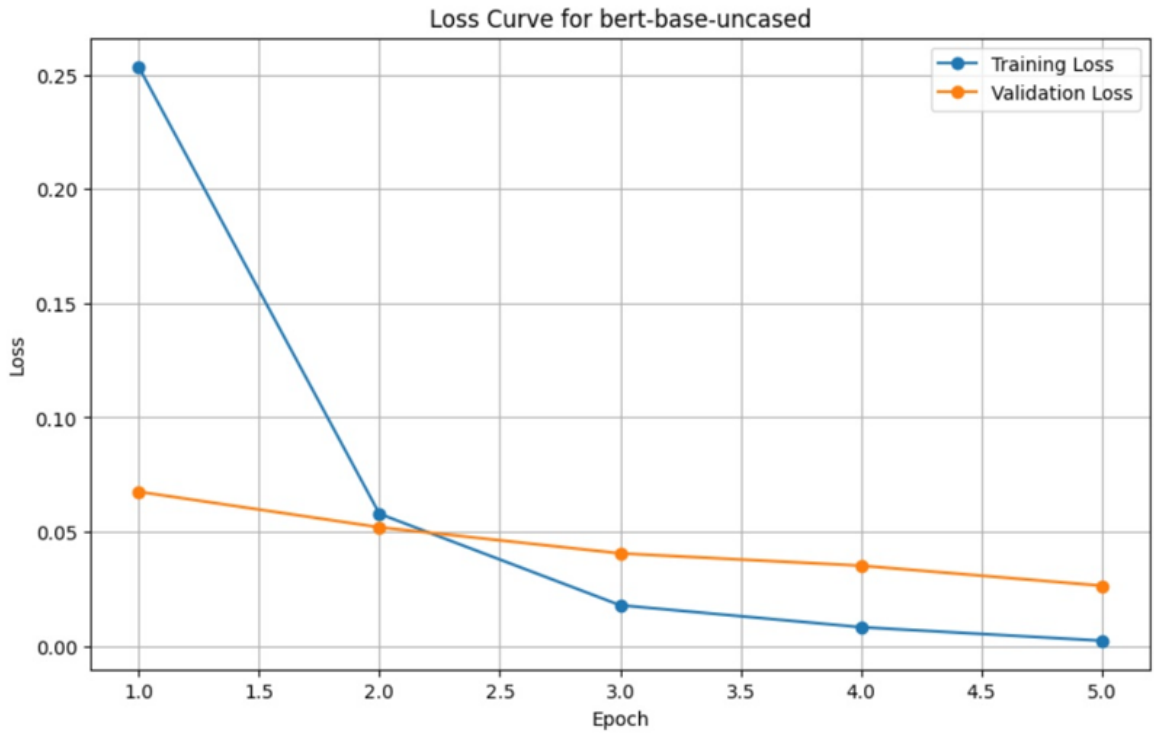
Epoch	Training Loss	Validation Loss
1	0.323000	0.226981
2	0.094000	0.064279
3	0.043900	0.073166
4	0.019100	0.046444
5	0.006000	0.049006



BERT-base-uncased:

- Training Loss: Starts at 0.2535 and decreases to 0.0024.
- Validation Loss: Starts at 0.067 and decreases to 0.0264.

Epoch	Training Loss	Validation Loss
1	0.253500	0.067551
2	0.057900	0.052010
3	0.017900	0.040593
4	0.008300	0.035216
5	0.002400	0.026445



Data augmentation appears to help both models converge faster and generalize better on the validation set. This is reflected in the tighter alignment between training and validation loss curves.

Overall Accuracy: While the difference is minimal, RoBERTa marginally outperforms BERT in accuracy. This indicates that both models are comparably capable of performing the classification task.

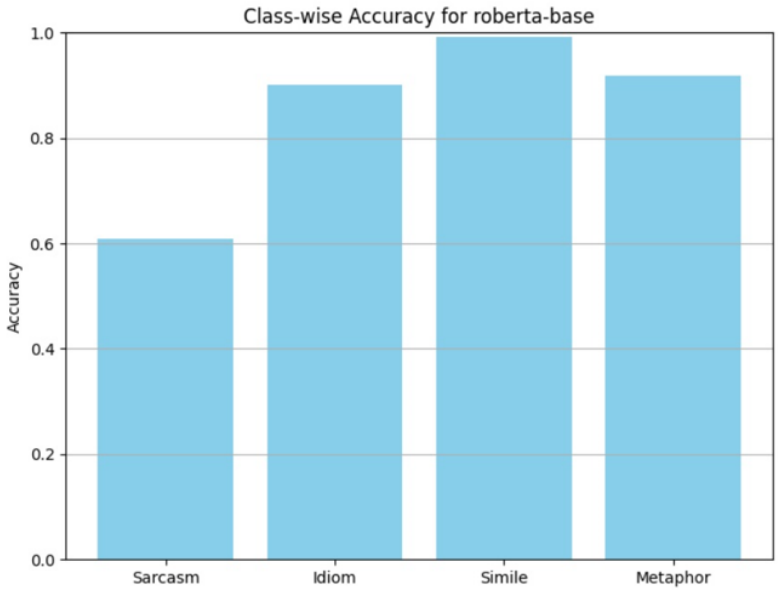
Accuracy for roberta-base: 0.77

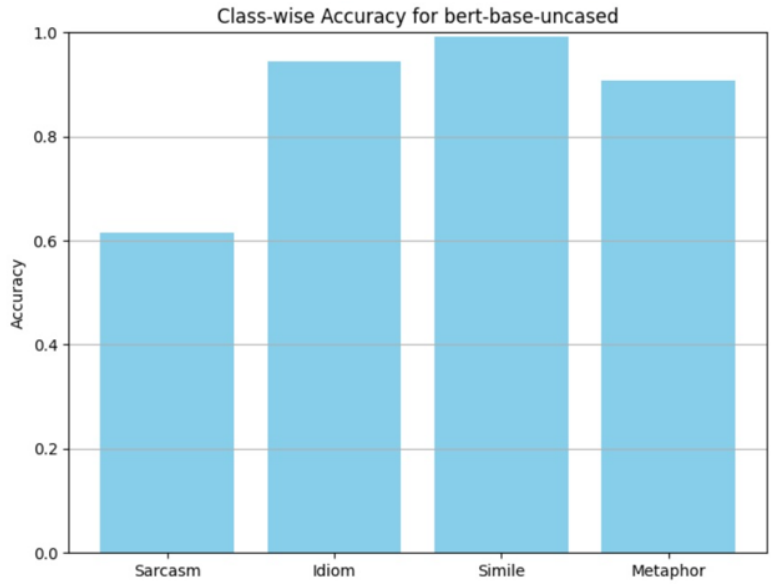
Accuracy for bert-base-uncased: 0.78

The slight decrease in accuracy for both models after data augmentation suggests that the augmented data may have introduced noise or examples that the models found challenging to generalize from. This highlights that while augmentation can enrich the dataset, it can also alter the data distribution in a way that negatively impacts performance. Possible causes could be the augmentation quality or the class balance which should be checked and modified again.

Class-wise Accuracy: The accuracies for each class are summarized below:

Class	RoBERTa-base	BERT-base-uncased
Sarcasm	60.9%	61.6%
Idiom	90%	94.4%
Simile	99.2%	99.2%
Metaphor	91.9%	90.7%





Both models show high accuracy for Idiom, Simile, and Metaphor, maintaining consistent performance across these categories even after augmentation.

Performance on Sarcasm remains low for both models, indicating that sarcasm detection remains challenging even with augmented data.

Precision, Recall, and F1-Score: The detailed precision, recall, and F1-scores for both models are summarized below:

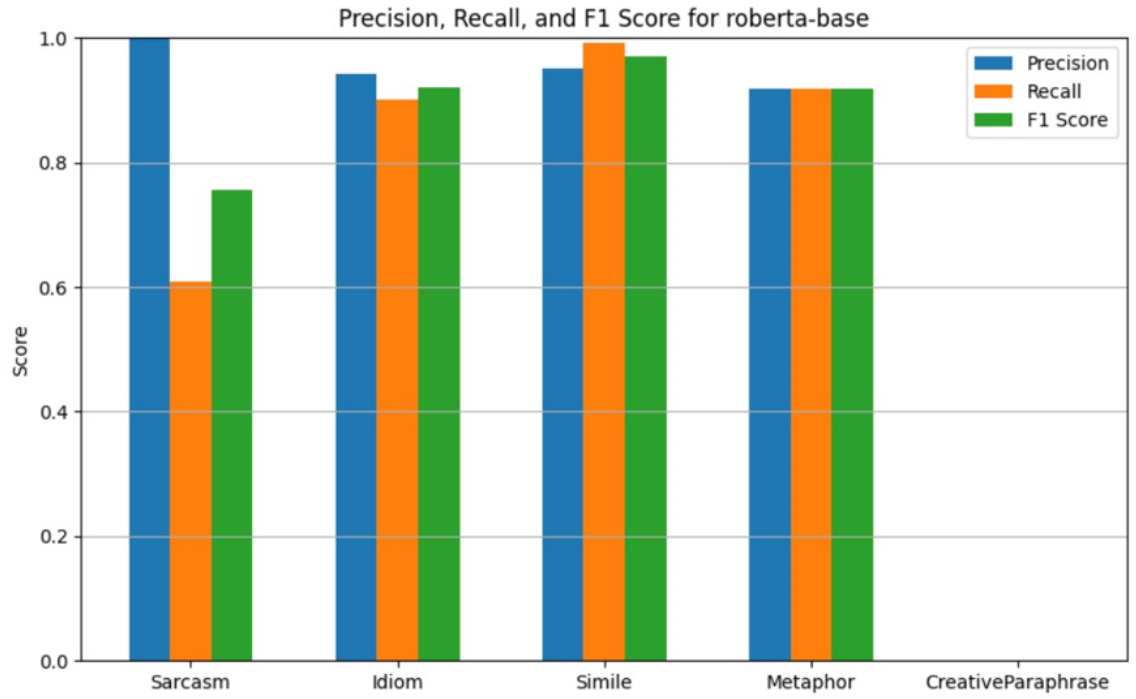
Class	Model	Precision	Recall	F1-Score
Sarcasm	RoBERTa-base	1.00	0.61	0.76
	BERT-base-uncased	1.00	0.62	0.76
Idiom	RoBERTa-base	0.94	0.90	0.92
	BERT-base-uncased	0.92	0.94	0.93
Simile	RoBERTa-base	0.95	0.99	0.97
	BERT-base-uncased	0.97	0.99	0.98
Metaphor	RoBERTa-base	0.92	0.92	0.92
	BERT-base-uncased	0.94	0.91	0.92
Creative Paraphrase	Both Models	-	-	-

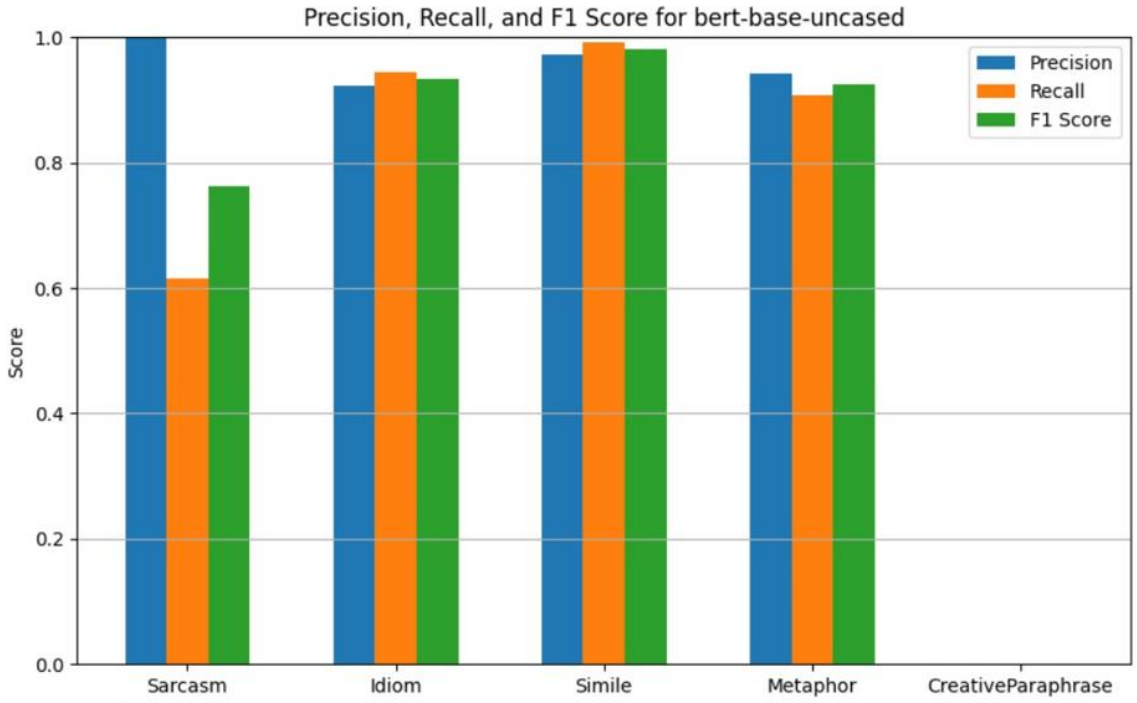
Classification Report for roberta-base:

	precision	recall	f1-score	support
Sarcasm	1.00	0.61	0.76	750
Idiom	0.94	0.90	0.92	250
Simile	0.95	0.99	0.97	250
Metaphor	0.92	0.92	0.92	248
CreativeParaphrase	0.00	0.00	0.00	0
accuracy			0.77	1498
macro avg	0.76	0.68	0.71	1498
weighted avg	0.97	0.77	0.85	1498

Classification Report for bert-base-uncased:

	precision	recall	f1-score	support
Sarcasm	1.00	0.62	0.76	750
Idiom	0.92	0.94	0.93	250
Simile	0.97	0.99	0.98	250
Metaphor	0.94	0.91	0.92	248
CreativeParaphrase	0.00	0.00	0.00	0
accuracy			0.78	1498
macro avg	0.77	0.69	0.72	1498
weighted avg	0.97	0.78	0.85	1498

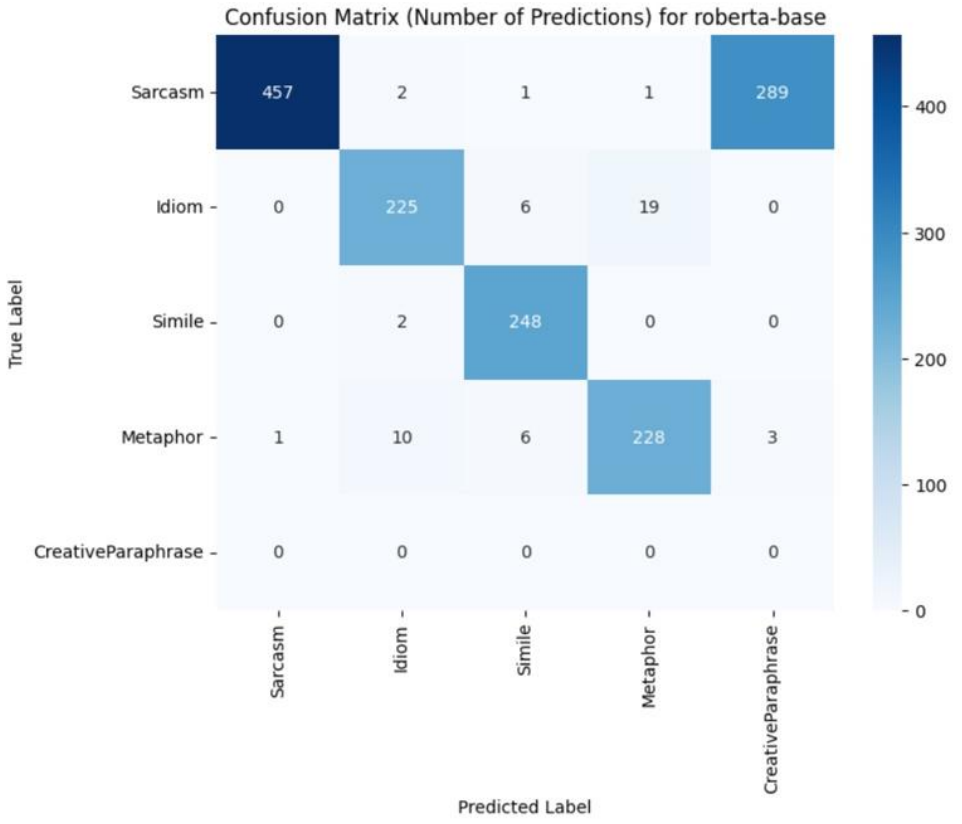


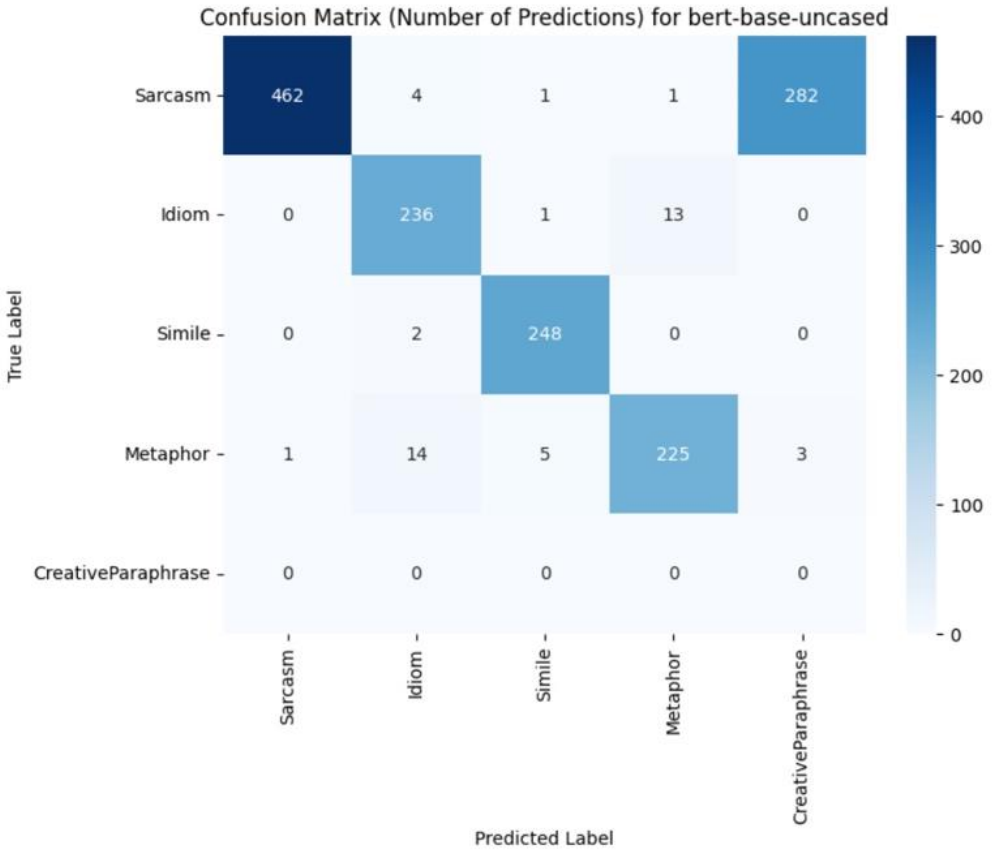


Precision for both models remains strong across all classes, especially for Sarcasm, suggesting low false positive rates. Recall for Sarcasm remains low, indicating the models still struggle to capture all sarcastic instances.

The F1-scores for Idiom, Simile, and Metaphor remain above 0.90, reflecting strong overall performance in these categories

Confusion Matrix: Despite data augmentation, the models continue to confuse Sarcasm with Creative Paraphrase, possibly due to overlapping linguistic patterns in these categories.





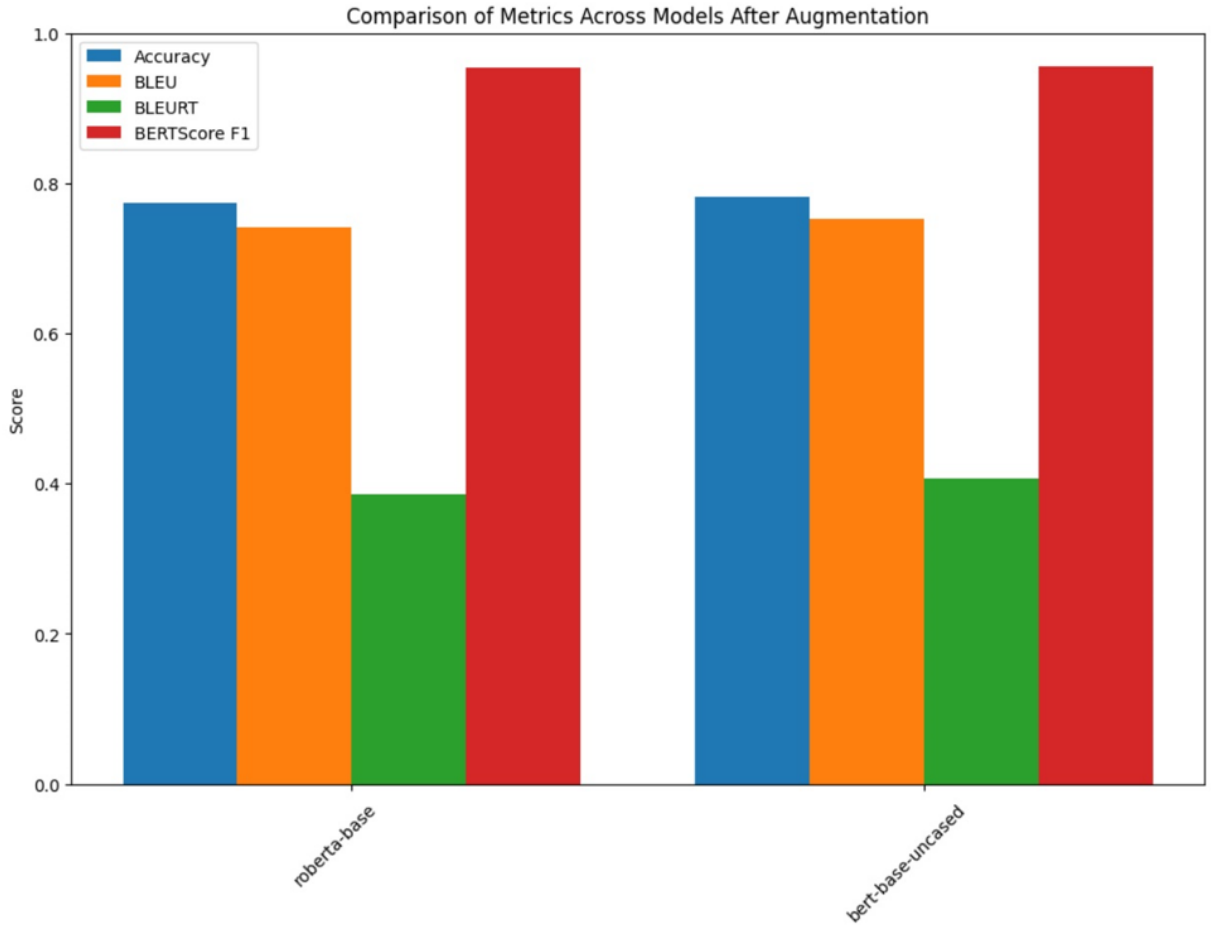
Evaluation Metrics: The following metrics were used to evaluate generative performance:

Metric	RoBERTa-base	BERT-base-uncased
BLEU	0.7411	0.7528
BLEURT	0.3851	0.4061
BERTScore F1	0.9538	0.9555

BLEU: Scores remain relatively stable, with BERT marginally outperforming RoBERTa.

BLEURT: Both models perform similarly, with BERT again scoring slightly higher.

BERTScore F1: Remains consistently high, demonstrating that the models capture the semantic essence of the figurative language effectively.



Misclassified Examples: Total Misclassified Examples:

- **RoBERTa-base:** 340 out of 1498 (22.7%).
- **BERT-base-uncased:** 327 out of 1498 (21.8%).

Even after data augmentation, misclassifications primarily involve Sarcasm being confused with Creative Paraphrase. A common theme in the errors is subtle linguistic ambiguity or context reliance, making it challenging for the models to correctly classify.

Conclusion: Both RoBERTa-base and BERT-base-uncased perform competitively, with BERT-base-uncased exhibiting a slight edge in overall metrics after augmentation. However, both models struggle with Sarcasm, indicating room for improvement in handling linguistically subtle categories.

9. Conclusion up to this stage:

While augmentation added diversity, the accuracy slightly decreased for both models, emphasizing the importance of evaluating augmentation methods for their ability to improve model generalization. Possible reasons for this failure can be considered as below:

- **Augmentation Quality:** If the augmented examples are not well-aligned with the original data distribution, the models might struggle to learn effectively.
- **Class Balance:** Augmentation might have unintentionally emphasized certain classes more than others, affecting the models' ability to generalize across all categories.