Explanation Consistency Under Data Augmentation An Empirical Study on Sentiment Classification

Johny Tarbouch
https://github.com/JohnyTarbouch/augmented-explained-transformer

Abstract

We study whether synonym-based data augmentation changes the consistency, stability, and quality of post-hoc explanations for transformer sentiment models. Using DistilBERT fine-tuned on the SST-2 dataset, we compare a baseline model against one trained with WordNet synonym augmentation. We evaluate explanation consistency using Integrated Gradients, LIME, and attention weights across three metrics: Kendall's τ, top-k overlap, and cosine similarity. Additionally, we assess faithfulness via AOPC curves and validate explanations through sanity checks. Our results show that synonym augmentation does not significantly improve explanation consistency, IG and attention consistency remain stable, while LIME shows marginal improvement. Both models pass faithfulness and sanity checks equally.

1 Introduction

Transformer models are accurate but ambiguous, motivating post-hoc explanation methods for debugging and trust. However, explanations should be *consistent* under meaning-preserving input changes [Nauta et al., 2023]: if a model predicts the same label for an original sentence and its augmented counterpart, the salient evidence should remain similar. Data augmentation is a widely used technique to improve model robustness [Wei and Zou, 2019, Xie et al., 2020]. However, its effect on explanation consistency, whether a model produces stable explanations for semantically equivalent inputs, remains underexplored.

This report evaluates whether training on augmented data change explanation consistency in DistilBERT sentiment classification. We measure explanation stability for Integrated Gradients [Sundararajan et al., 2017], LIME [Ribeiro et al., 2016], and attention-based explanations [Jain and Wallace, 2019, Wiegreffe and Pinter, 2019], and assess explanation quality with faithfulness (AOPC) and sanity checks. The key research question is:

> *Does training with synonym augmentation make explanations more consistent under input perturbations?*

We focus on *explanation consistency*, measuring whether the model attributes importance to the same words when presented with semantically equivalent sentences. Explanation consistency matters for:

- **Trust**: Users expect that semantically equivalent inputs ("I love this movie" vs "I adore this film") should yield similar explanations. Inconsistent explanations reduce user confidence.
- **Reliability**: Consistent explanations indicate that the model learned the underlying *concept* (positive sentiment) rather than memorizing specific *words*.
- **Real-world applications**: In domains like healthcare, explanation stability is critical. A medical AI should provide similar reasoning.
- **Debugging**: Consistent models are easier to audit and debug (predictable).

Project for Reinforcement Learning lecture

Importantly, consistency measures the *stability of the explanation method*, not whether the model's prediction is correct. If an explanation method yields high consistency, we gain confidence that observed attributions reflect genuine model behavior rather than noise.

2 Related Work

Explanation Methods for NLP. Attention weights have been controversially proposed as explanations [Jain and Wallace, 2019, Wiegreffe and Pinter, 2019], with ongoing debate about their faithfulness. Recent work by Chefer et al. [2021] extends interpretability beyond raw attention weights. Therefore, gradient-based methods like Integrated Gradients [Sundararajan et al., 2017] provide theoretically grounded attributions satisfying axiomatic properties including completeness and sensitivity. LIME [Ribeiro et al., 2016] offers a model-agnostic alternative by fitting local linear approximations to the decision boundary.

Evaluation of Explanations. The ERASER benchmark [DeYoung et al., 2020] introduced standardized evaluation protocols including comprehensiveness and sufficiency metrics. Sanity checks [Adebayo et al., 2018] test whether explanations depend on learned model parameters, identifying cases where gradient-based methods produce constant outputs regardless of model weights.

Data Augmentation. Wei and Zou [2019] proposed simple text augmentation techniques to improve classification task. However, the author argue that transformers dose not need data augmentation.

3 Approach

3.1 Experimental Setup

Dataset. We use SST-2 [Socher et al., 2013], a binary sentiment classification dataset from the GLUE benchmark [Wang et al., 2018], containing movie review labeled as positive or negative.

Augmentation Strategy. For training example, we create an augmented variant by replacing 10% of non-stopword tokens with WordNet synonym replacement [Wei and Zou, 2019].

Models. We compare two configurations:

- **Baseline**: DistilBERT [Sanh et al., 2019], the standard HuggingFace modely [1].
- **Augmented (5 seeds)**: The same architecture further fine-tuned on augmented data.

All experiments are run with 5 **random seeds** (13, 21, 42, 1337, 2024) and aggregated. Consistency metrics are computed on 200 samples per seed. We fine-tuned the model using 3 **epochs** and a **batch size** of 32 with an RTX 3060 Ti.

3.2 Explanation Methods

We use three explanation approaches (**however, our main focus is IG**):

Integrated Gradients (IG): IG computes attributions by integrating gradients along a path from a baseline input to the actual input:

$$\mathrm{IG}_i(x) = (x_i - x'_i) \cdot \int_{\alpha=0}^{1} \frac{\partial F(x' + \alpha(x - x'))}{\partial x_i} d\alpha \qquad s_t = \sum_{d=1}^{768} \mathrm{IG}_{t,d} \tag{1}$$

where x' is a baseline (padding tokens, see Figure 1), F is the model output, and i indexes each embedding dimension. IG satisfies the completeness axiom: attributions sum to $F(x) - F(x')$.

In practice, we approximate the integral via Riemann summation with $n = 50$ steps. Since each token is represented by a 768-dimensional embedding, we obtain a 768-dimensional attribution

[1] https://huggingface.co/distilbert/distilbert-base-uncased-finetuned-sst-2-english

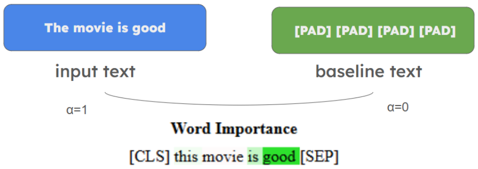


Figure 1: Integrated Gradients intuition (baseline to input interpolation).

vector per token. To get a single importance score per token t, we sum over embedding dimensions (right equation). Finally, we aggregate subword scores to word-level (`["amaz", "##ing"]` → `"amazing"`).

LIME: generates perturbations by randomly masking words, collects model predictions, and fits a weighted linear model to approximate local behavior. Word weights indicate importance.

Attention: While attention's status as explanation is debated [Jain and Wallace, 2019], it provides an interesting comparison point. We extract attention weights from the `[CLS]` token to all other tokens:

$$\mathrm{Attn}_i = \frac{1}{H}\sum_{h=1}^{H} \alpha^{(h)}_{[\mathrm{CLS}],i} \tag{2}$$

3.3 Consistency Metrics

For each input, we create a synonym-augmented version and compute explanations for both.
Token Alignment: To compare original and augmented sentences, we align tokens by exact matches using SequenceMatcher and then by WordNet synonym matching for leftover tokens. Metrics are computed only on aligned tokens, **otherwise**, the metrics will be meaningless, then compute:

Kendall's τ. Rank correlation measuring agreement in importance ordering:

$$\tau = \frac{C - D}{C + D} \tag{3}$$

where C is the same relative ordering in both rankings and D is the opposite ordering. $\tau = 1$ indicates identical ranking.

Top-k Overlap. Fraction of shared tokens among the top-k most important:

$$\text{Top-k} = \frac{|A_k \cap B_k|}{k} \tag{4}$$

where A_k and B_k are the top-k tokens in each explanation.

Cosine Similarity. Agreement in attribution magnitude and direction:

$$\cos(\mathbf{a}, \mathbf{b}) = \frac{\mathbf{a} \cdot \mathbf{b}}{\|\mathbf{a}\|\|\mathbf{b}\|} \tag{5}$$

3.4 Faithfulness Evaluation

We evaluate using AOPC(Area Over Perturbation Curve) [DeYoung et al., 2020], to measure whether the tokens identified as important by the explanation method actually influence the models predictions.

Comprehensiveness: We progressively remove the top-ranked tokens (by IG attribution) and measure the drop in prediction probability. If explanations are faithful, removing important tokens should significantly decrease confidence:

$$\mathrm{AOPC}_{\mathrm{comp}} = \frac{1}{|F|}\sum_{f \in F} (p_{\mathrm{orig}} - p_{\mathrm{masked}}) \tag{6}$$

where F is the set of removal fractions (10%, 20%, ...), p_{orig} is the original prediction probability, and p_{masked} is the probability after masking the top fraction of tokens.

Sufficiency: Conversely, we keep only the top-ranked tokens and mask all others. If the selected tokens capture the model's reasoning, the prediction should remain stable:

$$\text{AOPC}_{\text{suff}} = \frac{1}{|F|} \sum_{f \in F} (p_{\text{orig}} - p_{\text{kept}}) \tag{7}$$

where p_{kept} is the probability when only the top fraction of tokens is retained.

Lower AOPC indicates the attributed tokens are sufficient for the prediction. We compare against a **random baseline** where tokens are removed/kept in random order, rather than by importance ranking.

3.5 Sanity Checks

Following Adebayo et al. [2018], we test whether IG depends on learned weights by progressively randomizing layers from the classifier head backward toward earlier transformer layers and measuring IG divergence from the original trained model.
If IG is faithful, explanations should become less similar as more layers are randomized, since random weights encode no learned signal. Let $E(x;\theta)$ denote the IG explanation for input x with trained parameters θ, and $\theta^{(k)}$ the parameters after randomizing the top k layers. We measureÖ

$$s_k(x) = \text{Sim}\Big(E(x;\theta), E(x;\theta^{(k)})\Big), \tag{8}$$

where Sim is Kendall's τ, top-k, or cosine-sim. A faithful method should show $s_k \downarrow$ as k increases.

4 Experiments and Results

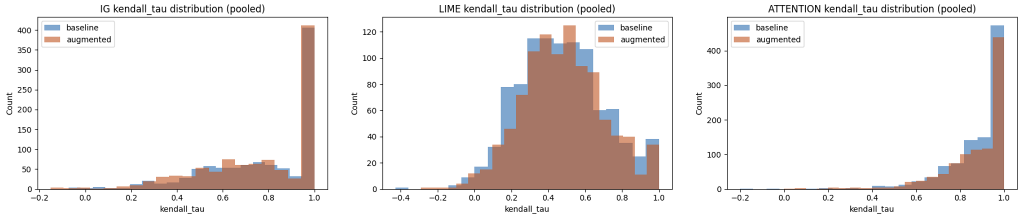
4.1 Consistency Results

Table 1: Explanation consistency metrics (mean ± std over 5 seeds). Higher is better for all metrics.

Method	**Model**	**Kendall's τ**	**Top-k Overlap**	**Cosine Sim.**
IG	Baseline	$\mathbf{0.794 \pm 0.010}$	$\mathbf{0.908 \pm 0.003}$	0.939 ± 0.006
	Augmented	0.786 ± 0.009	0.906 ± 0.005	$\mathbf{0.940 \pm 0.007}$
LIME	Baseline	0.471 ± 0.006	0.776 ± 0.004	0.856 ± 0.010
	Augmented	$\mathbf{0.486 \pm 0.021}$	$\mathbf{0.783 \pm 0.010}$	$\mathbf{0.864 \pm 0.013}$
Attention	Baseline	$\mathbf{0.892 \pm 0.006}$	$\mathbf{0.955 \pm 0.002}$	0.974 ± 0.002
	Augmented	0.888 ± 0.005	0.952 ± 0.002	$\mathbf{0.975 \pm 0.003}$

Table 1 shows the main consistency results. Key observations:

- **IG consistency** is high for both models, with the baseline slightly outperforming the augmented model.
- **LIME** shows the only improvement with augmentation: Kendall's τ increases (+3.4%), though this remains the lowest among methods.
- **Attention** has the highest consistency, with minimal difference between models.

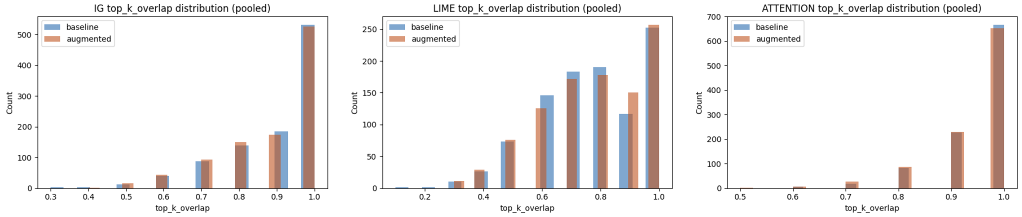


(a) IG Kendall's τ

(b) LIME Kendall's τ

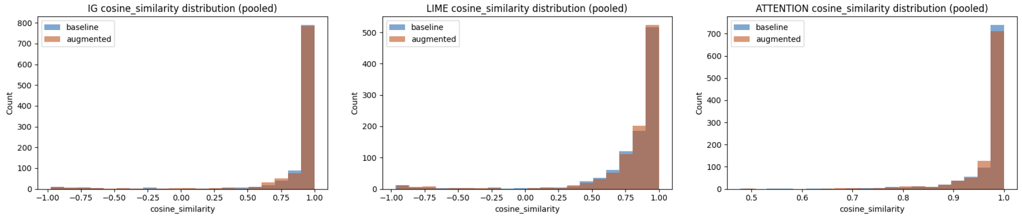
(c) Attention Kendall's τ

Figure 2: Kendall's τ distributions (pooled across 5 seeds). IG and Attention show high consistency with peaks near 1.0. LIME shows broader distributions with lower overall consistency.



(a) IG Top-k Overlap (b) LIME Top-k Overlap (c) Attention Top-k Overlap

Figure 3: Top-k overlap distributions (pooled across 5 seeds). Attention achieves best overlap (> 0.95), while LIME shows more variability.



(a) IG Cosine Similarity (b) LIME Cosine Similarity (c) Attention Cosine Similarity

Figure 4: Cosine similarity distributions (pooled across 5 seeds). All methods show strong similarity, with Attention achieving the highest values (> 0.97).

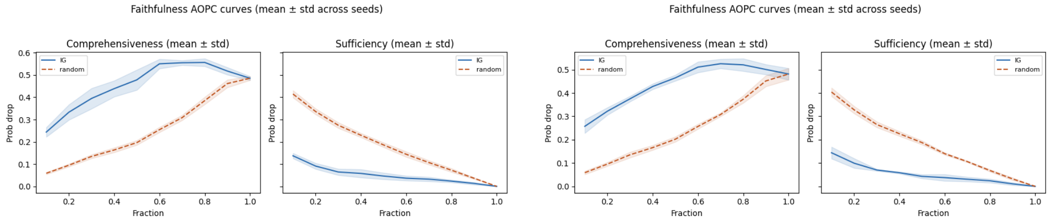
4.2 Faithfulness Results

Table 2: Faithfulness metrics (AOPC, higher comprehensiveness and lower sufficiency is better).

Model	AOPC Comp.	AOPC Suff.	Random Comp.
Baseline	0.455 ± 0.020	0.050 ± 0.009	0.254 ± 0.007
Augmented	0.439 ± 0.012	0.052 ± 0.007	0.253 ± 0.005

Faithfulness metrics (Table 2) are nearly identical between models. Both achieve:

- **High comprehensiveness**: removing top IG tokens drops prediction confidence.
- **Low sufficiency**: top tokens are nearly sufficient for the prediction.
- **Better than random**: IG substantially outperforms random token selection.



(a) Baseline (b) Augmented

Figure 5: AOPC curves: probability drop. Both models show similar faithfulness patterns.

4.3 Sanity Check Results

Figure 6 shows the sanity check results. As model layers are progressively randomized:

- IG explanations degrade appropriately (Kendall's τ drops from 1.0 to near 0).
- Both models pass the sanity check, confirming IG depends on learned parameters.
- Cosine similarity drops faster than rank correlation, indicating magnitude is more sensitive than ordering.

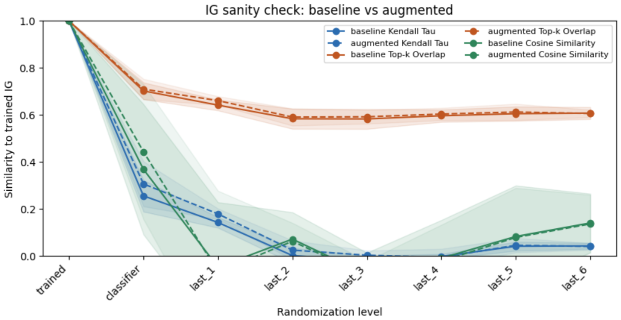


Figure 6: Progressive layer randomization reduces IG similarity, both models pass the sanity check.

5 Discussion of Results

5.1 Comparison of Explanation Methods

We employ three distinct categories of explanation methods to ensure a comprehensive analysis. Their differences explain the variance in our results:

- **Integrated Gradients (Gradient-based):** IG attributes importance by integrating gradients from a neutral baseline to the real input in embedding space. This yields *path-aware* attributions that are deterministic, sensitive to learned weights, and satisfy the completeness axiom. In practice, IG is less noisy than perturbation methods and more faithful than attention because it directly measures output sensitivity.
- **LIME (Perturbation-based):** approximates the model locally by sampling random perturbations (masked inputs) and fitting a linear model. *Why discrepancies occur:* LIME is stochastic, the random sampling introduces inherent variance, explaining why it consistently yields lower stability scores compared to deterministic methods.
- **Attention (Intrinsic):** uses the model's internal attention weights. *Why high consistency:* Attention reflects architectural information flow rather than causal influence on the output. It is structurally stable, leading to the highest consistency scores, though its faithfulness as an explanation is debated.

Why results differ: LIME's perturbation-based nature may explain the marginal LIME improvement (+3.4% Kendall's τ). It generates local explanations by randomly masking words and observing prediction changes. A model trained on augmented data may have learned more robust internal representations that generalize better to LIME's perturbation distribution, whereas gradient-based and intrinsic methods remained largely unaffected by the subtle synonym changes.

5.2 Implications for Trust and Reliability

We argued 1 that explanation consistency is essential for user trust. Our results suggest that:

- **Trust**: Both baseline and augmented models achieve high consistency, meaning users can reasonably trust that explanations for semantically similar inputs will be similar. However, augmentation does not further improve this trust.
- **Reliability**: The high baseline consistency suggests that pretrained DistilBERT already **learned robust sentiment** *concepts* rather than memorizing specific words. Augmentation does not enhance this property.

Therefore, the original model (without augmentation) already provides a trustworthy explanation, with no need for further fine-tuning on augmented data.

5.3 Why Didn't Augmentation Help?

- **Augmentation intensity**: Synonym replacement might be too small a change to noticeably affect the model's internal representations.
- **Baseline stability**: The pretrained DistilBERT model already achieves high consistency, leaving limited room for improvement. DistilBERT is pretrained on large, **semantically and grammatically** correct corpora, so it already learns robust patterns, **synonym replacement can introduce unnatural phrasing** that slightly degrades this correctness
- **Synonym limitation**: WordNet synonyms provide new sentences, but may not create sufficiently diverse linguistic patterns to encourage more robust representations.

5.4 Faithfulness and Sanity Checks

Both models have similar faithfulness and outperform random baselines, indicating IG highlights truly important tokens. Sanity checks (Fig. 6) confirm IG depends on learned weights: progressive layer randomization degrades explanations similarly for both models, suggesting augmentation does not measurably alter representations detectable by gradient-based attribution.

5.5 Limitations & Future Work

We used single dataset (SST-2) and model architecture (DistilBERT), with only WordNet synonym augmentation, and a fixed augmentation fraction (10%).

For future work, test stronger augmentation: Combine synonym replacement with back-translation, to get new, diverse, semantically and grammatically correct data. Additionally, consistency aware training objectives.

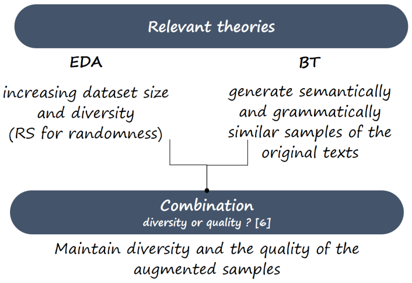

Relevant theories
EDA
BT
increasing dataset size and diversity (RS for randomness)
generate semantically and grammatically similar samples of the original texts
Combination
diversity or quality ? [6]
Maintain diversity and the quality of the augmented samples

Figure 7: Future work: Diverse and semantically correct data.

6 Conclusion

Our contributions include:

- A systematic empirical study comparing explanation consistency between baseline and augmented DistilBERT models on SST-2.
- Evaluation across three explanation methods (IG, LIME, Attention) and three consistency metrics (Kendall's τ, top-k overlap, cosine similarity).
- Faithfulness validation using AOPC curves and sanity checks via layer randomization.
- A meaningful negative result showing that simple synonym augmentation does not substantially improve explanation quality.

References

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. *Advances in neural information processing systems*, 31, 2018.

Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 782–791, 2021.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C Wallace. Eraser: A benchmark to evaluate rationalized nlp models. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 4443–4458, 2020.

Sarthak Jain and Byron C Wallace. Attention is not explanation. *arXiv preprint arXiv:1902.10186*, 2019.

Meike Nauta, Jan Trienes, Shreyasi Pathak, Elisa Nguyen, Michelle Peters, Yasmin Schmitt, Jörg Schlötterer, Maurice Van Keulen, and Christin Seifert. From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai. *ACM Computing Surveys*, 55(13s):1–42, 2023.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642, 2013.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR, 2017.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP workshop BlackboxNLP: Analyzing and interpreting neural networks for NLP*, pages 353–355, 2018.

Jason Wei and Kai Zou. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. *arXiv preprint arXiv:1901.11196*, 2019.

Sarah Wiegreffe and Yuval Pinter. Attention is not not explanation. *arXiv preprint arXiv:1908.04626*, 2019.

Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. Unsupervised data augmentation for consistency training. *Advances in neural information processing systems*, 33:6256–6268, 2020.

Checklist

1. General points:
   (a) Do the main claims made in the abstract and introduction accurately reflect your contributions and scope? [Yes]
   (b) Did you cite all relevant related work? [Yes]
   (c) Did you describe the limitations of your work? [Yes]
   (d) Did you include a discussion of future work? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments (e.g. for benchmarks)...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] GitHub repository linked in author information.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] Configs available in repository.
   (c) Did you run at least 10 repetitions of your method? [No] We ran 5 seeds due to computational/time constraints.
   (d) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] All tables include standard deviations across seeds.
   (e) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] SST-2, DistilBERT, and all methods are cited.
   (b) Did you make sure the license of the assets permits usage? [Yes]
   (c) Did you reference the assets directly within your code and repository? [Yes]

7 Appendix

Validation Accuracy. Baseline: **91.06%**, Augmented: **90.25%** on the SST-2 validation split.

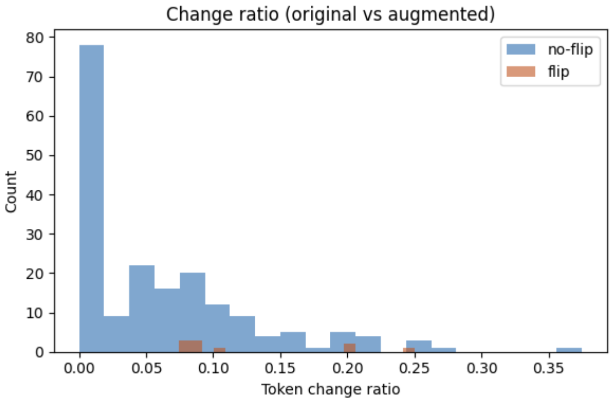


Figure 8: Flip-label analysis: distribution of token change ratios for flip vs no-flip cases.