
Fine-Tuning DistilBERT for Ethical Content Moderation

Abstract

This project presents a practical approach to building more ethical AI systems by fine-tuning DistilBERT to detect hate speech and offensive language. By addressing class imbalance on the Davidson dataset, the model's recall for hate speech was improved eightfold while transparently documenting precision trade-offs. All results, code, and recommendations are openly shared to support safer AI moderation and responsible data filtering.

William Radiyeh

May 12, 2025

Email: will.rads@outlook.com

Github: github.com/will-rads

	1
1. Introduction	1
2. Dataset & Motivation	2
3. Class Imbalance: The Initial Hurdle	2
4. Model Architecture & Fine-Tuning	3
5. Results after Fine-Tuning (with Frozen Base and Class Weights)	3
6. Additional Experiment: Fine-Tuning with Unfrozen DistilBERT Layers	4
7. Qualitative Check	6
8. Discussion & Limitations	7
9. Use-Case Section: Applications in Ethical AI	8
10. Conclusion	9
Appendix A: Training Details	10
Appendix B: Fairness & Robustness Audit	11
Appendix C: Model Card	12

1. Introduction

Online platforms face increasing pressure to detect and remove harmful content, particularly hate speech. While manual moderation is essential, scalable and reliable AI tools are becoming equally important. Yet, detecting hate speech is complex. It's context-sensitive, evolves rapidly, and often appears in forms difficult for models to catch.

This project applies transfer learning using DistilBERT, a lightweight transformer model, to classify tweets into hate speech, offensive language, or neither. My work highlights how targeted adjustments can lead to significant gains in identifying hate speech, helping pave the way for more ethical and effective AI moderation systems.

2. Dataset & Motivation

This project aimed to build a deep learning model that detects and classifies problematic online language, focusing on distinguishing Hate Speech, Offensive Language, and Neutral content. The motivation lies in the growing prevalence of harmful speech online and the need for scalable moderation tools, especially ones that can help build ethical AI by filtering biased or harmful training data.

We used the widely adopted Davidson et al. Hate Speech and Offensive Language dataset (via Hugging Face: [tdavidson/hate_speech_offensive](#)), which includes approximately 24,000 tweets labeled into three classes:

- **Class 0:** Hate Speech
- **Class 1:** Offensive Language
- **Class 2:** Neither (Neutral)

To ensure class proportions remained consistent, the data was split into **80% training** and **20% validation** using stratified sampling.

3. Class Imbalance: The Initial Hurdle

Initial training of our DistilBERT-based model (frozen base, custom head) immediately revealed a common real-world challenge: class imbalance. In this dataset, hate speech represents a small fraction of examples compared to the more frequent offensive and neutral tweets.

This imbalance had a major impact. The first version of the model achieved a seemingly high 86% overall accuracy (weighted average), but failed to identify hate speech effectively, scoring just 0.07 recall and 0.12 F1-score for that class.

```

--- Detailed Validation Set Evaluation ---
310/310 25s 68ms/step

Classification Report:

```

	precision	recall	f1-score	support
Hate Speech	0.59	0.07	0.12	286
Offensive Language	0.88	0.96	0.92	3838
Neither	0.76	0.66	0.70	833
accuracy			0.86	4957
macro avg	0.74	0.56	0.58	4957
weighted avg	0.84	0.86	0.84	4957

```

Confusion Matrix:
[[ 20 236  30]
 [  9 3686 143]
 [  5 281 547]]

```

Figure 1: Note the high overall accuracy but critically low recall (0.07) and F1-score (0.12) for the "Hate Speech" class.

Why Class Imbalance Hurts Performance on Minority Classes:

Standard classification models optimize for overall accuracy, which can be misleading when class distributions are skewed. Without intervention, the model tends to favor the majority classes, learning little about rare but important ones like hate speech.

Addressing Class Imbalance with Balanced Class Weights:

To tackle this, we applied a widely-used method called balanced class weighting. This technique adjusts the training loss function to penalize mistakes on minority classes more than majority classes.

Specifically, we used scikit-learn's `class_weight='balanced'` option, which automatically calculates weights based on how frequently each class appears in the training set. This encourages the model to focus more on correctly identifying examples of hate speech.

Prior work explores SMOTE, undersampling and class-weighting. For my project, I chose class weighting due to its simplicity, effectiveness, and compatibility with deep learning methods, without needing to change the original data distribution.

4. Model Architecture & Fine-Tuning

We used **transfer learning** with the pretrained `distilbert-base-uncased` model as our feature extractor. DistilBERT offers a lighter, faster alternative to BERT, retaining strong language understanding while being more efficient. For this phase, the base layers were **frozen** to preserve their pretrained knowledge while we trained a custom classifier for our task.

Custom Classification Head:

On top of the frozen DistilBERT, we added a feed-forward classification head that takes the `[CLS]` token embedding (representing the input tweet) and passes it through a series of Dense layers:

- **Dense Layer 1:** 256 units
- **Dense Layer 2:** 128 units
- **Dense Layer 3:** 32 units
- **Output Layer:** 3 units (matching our three classes: Hate Speech, Offensive Language, Neither) with a softmax activation, giving us the probability for each class.

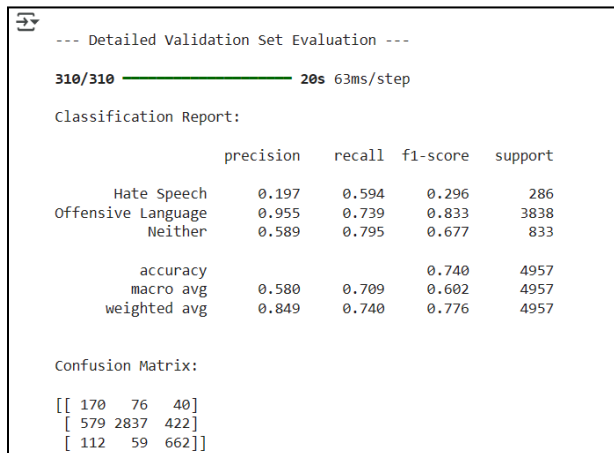
Regularization & Training Techniques: To improve generalization and training stability, we applied the following:

- **He Normal Initialization:** Helps maintain activation variance across layers, especially effective with ReLU-based activations like ReLU and Swish, which we used in the hidden layers.
- **L2 Regularization:** Penalizes large weights to reduce overfitting and encourage simpler representations.
- **Batch Normalization:** Stabilizes training by normalizing layer outputs, enabling higher learning rates and faster convergence.
- **Dropout:** Randomly deactivates neurons during training to prevent co-adaptation and encourage redundancy.
- **Activation Functions:**
 - **LeakyReLU**
 - **Swish**

We trained this setup using the Adam optimizer with a learning rate of $3e-5$, applying sparse categorical cross-entropy loss alongside the balanced class weights previously described.

5. Results after Fine-Tuning (with Frozen Base and Class Weights)

After applying balanced class weights and optimizing our classification head (128 \rightarrow 64 \rightarrow 32 Dense units with He Normal initialization, L2 regularization, Batch Normalization, and Dropout), we retrained the DistilBERT-based model on our dataset.



```

--- Detailed Validation Set Evaluation ---
310/310 20s 63ms/step
Classification Report:

```

	precision	recall	f1-score	support
Hate Speech	0.197	0.594	0.296	286
Offensive Language	0.955	0.739	0.833	3838
Neither	0.589	0.795	0.677	833
accuracy			0.740	4957
macro avg	0.580	0.709	0.602	4957
weighted avg	0.849	0.740	0.776	4957

```

Confusion Matrix:
[[ 170  76  40]
 [ 579 2837 422]
 [ 112  59 662]]

```

Figure 2: Model performance metrics on the validation set after implementing balanced class weights and tuning the classification head. The recall for "Hate Speech" improved significantly to 0.594.

Analysis of Results:

As expected, using balanced class weights considerably reshaped the model's performance:

- **"Hate Speech" (Class 0):**
 - **Recall:** Increased significantly from an initial 0.07 to 0.594, meaning our model now correctly detects nearly 60% of hate speech instances, a notable improvement from previous results.
 - **Precision:** Dropped to 0.197, showing that while the model identifies more hate speech, it also incorrectly flags many posts from the "Offensive Language" (579 instances) and "Neither" (112 instances) categories, as shown in the confusion matrix:
 - **F1-score:** Improved to 0.296, significantly better than the previous 0.12, although precision still needs improvement.
- **"Offensive Language" (Class 1):**
 - Precision remained comparatively high at 0.955.
 - Recall dropped slightly to 0.739 (from an initial 0.96), a trade-off resulting from more aggressively classifying some offensive content as hate speech.
- **"Neither" (Class 2):**
 - Recall remained strong at 0.795 with a moderate precision of 0.589.

Overall Accuracy: Decreased from 0.86 to 0.74, but this is a more realistic measure of model performance across all classes. These results show that class weighting successfully increased sensitivity to hate speech. Although precision remains low, recall has improved dramatically, making the model more

suitable for content moderation tasks where missing hate speech (false negatives) is riskier than catching false positives.

6. Additional Experiment: Fine-Tuning with Unfrozen DistilBERT Layers

To explore whether further gains could be made by adapting DistilBERT's internal language representations to our dataset, we conducted an additional fine-tuning experiment.

Methodology:

The following adjustments were made for this experiment:

- Starting from our previously optimized model (with class weights and a frozen base), we set the DistilBERT model to `trainable = True` and selectively unfroze the top two transformer blocks (layers 4 and 5). The lower layers and embeddings remained frozen to retain general language understanding and reduce the risk of overfitting.
- The model was recompiled using the AdamW optimizer with a reduced learning rate (`2e-5`) and `weight_decay=0.01`, then trained for 5 more epochs with the same training and validation sets.

Results of Fine-Tuning:

The fine-tuning process lasted 5 epochs, during which the learning rate was adaptively adjusted, as seen in the training log:

```
Starting fine-tuning training (Stage 2)...
Epoch 1/5
1240/1240 ----- 106s 74ms/step - accuracy: 0.6449 - loss: 0.8334 - val_accuracy: 0.7289 - val_loss: 0.7441 - learning_rate: 2.0000e-05
Epoch 2/5
1240/1240 ----- 86s 69ms/step - accuracy: 0.6516 - loss: 0.8197 - val_accuracy: 0.7359 - val_loss: 0.7330 - learning_rate: 2.0000e-05
Epoch 3/5
1239/1240 ----- 0s 56ms/step - accuracy: 0.6532 - loss: 0.8345
Epoch 3: ReduceLROnPlateau reducing learning rate to 9.999999747378752e-06.
1240/1240 ----- 142s 69ms/step - accuracy: 0.6532 - loss: 0.8345 - val_accuracy: 0.7192 - val_loss: 0.7686 - learning_rate: 2.0000e-05
Epoch 4/5
1239/1240 ----- 0s 56ms/step - accuracy: 0.6667 - loss: 0.8081
Epoch 4: ReduceLROnPlateau reducing learning rate to 4.999999873689376e-06.
1240/1240 ----- 142s 69ms/step - accuracy: 0.6667 - loss: 0.8081 - val_accuracy: 0.7289 - val_loss: 0.7474 - learning_rate: 1.0000e-05
Epoch 5/5
1239/1240 ----- 0s 56ms/step - accuracy: 0.6647 - loss: 0.8284
Epoch 5: ReduceLROnPlateau reducing learning rate to 2.499999936844688e-06.
1240/1240 ----- 142s 69ms/step - accuracy: 0.6647 - loss: 0.8284 - val_accuracy: 0.7242 - val_loss: 0.7566 - learning_rate: 5.0000e-06
```

Figure 3: Training log snippet during the fine-tuning phase with top DistilBERT layers unfrozen.

After fine-tuning, the model's performance on the validation set was re-evaluated:

Classification Report (after fine-tuning):				
	precision	recall	f1-score	support
Hate Speech	0.182	0.615	0.281	286
Offensive Language	0.955	0.714	0.817	3838
Neither	0.602	0.810	0.691	833
accuracy			0.724	4957
macro avg	0.580	0.713	0.596	4957
weighted avg	0.851	0.724	0.765	4957
Confusion Matrix (after fine-tuning):				
[[176 72 38]				
[691 2739 408]				
[100 58 675]]				

Figure 4: Performance metrics after attempting to fine-tune with the top 2 DistilBERT transformer blocks unfrozen. See Appendix A for training log & hyper-params.

Analysis of Fine-Tuning Experiment:

Comparing these results (Figure 4) with our best frozen-base model (Figure 2):

- **"Hate Speech" (Class 0):**
 - Recall improved slightly (0.594 → 0.615).
 - Precision decreased (0.197 → 0.182).
 - F1-score slightly declined (0.296 → 0.281).
- **"Offensive Language" (Class 1):**
 - Recall slightly decreased (0.739 → 0.714).
 - F1-score slightly decreased (0.833 → 0.817).
- **Overall Accuracy:**
Decreased slightly (0.740 → 0.724).

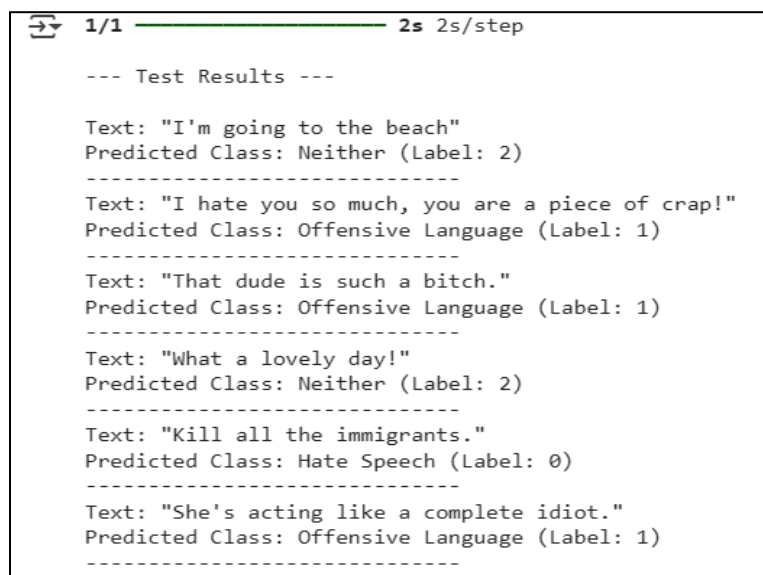
Conclusion of Experiment:

This fine-tuning attempt did not lead to significant improvements and, in fact, slightly degraded key metrics. These results suggest the previously frozen DistilBERT layers already provided strong contextual understanding. Given these outcomes, our primary model remains the one with the fully frozen DistilBERT base (Section 5, Figure 2).

7. Qualitative Check

To complement quantitative metrics, we tested the best-performing model (frozen base + class weights) on a small set of manually selected tweets

The following examples illustrate the model's predictions:



```

1/1 2s 2s/step

--- Test Results ---

Text: "I'm going to the beach"
Predicted Class: Neither (Label: 2)
-----
Text: "I hate you so much, you are a piece of crap!"
Predicted Class: Offensive Language (Label: 1)
-----
Text: "That dude is such a bitch."
Predicted Class: Offensive Language (Label: 1)
-----
Text: "What a lovely day!"
Predicted Class: Neither (Label: 2)
-----
Text: "Kill all the immigrants."
Predicted Class: Hate Speech (Label: 0)
-----
Text: "She's acting like a complete idiot."
Predicted Class: Offensive Language (Label: 1)
-----

```

Figure 5: Qualitative test results on a small set of manually selected tweets using the model with a frozen DistilBERT base and class weights.

Assessment:

The model handled this diverse set well, especially in correctly classifying hate speech and distinguishing it from general offensiveness. It also avoided mislabeling neutral or positive statements. While this is only

a small sample, it suggests that the improvements seen in recall meaningfully translate to real examples. A larger-scale qualitative audit would be required for deployment, but this early check supports the model's practical utility.

8. Discussion & Limitations

This project successfully developed a DistilBERT-based model to classify tweets into Hate Speech, Offensive Language, and Neutral. Our main achievement was significantly improving the detection of hate speech, a minority class, by applying balanced class weights and designing a regularized classification head. Our best model (frozen base, tuned head) raised hate speech recall from 0.07 to 0.594, a meaningful gain for content moderation.

Interpreting the Results: The qualitative check (Figure 5) confirmed that these improvements extended to real-world examples. Although overall accuracy dropped slightly (from 0.86 to 0.74), this is expected when shifting focus toward underrepresented classes. In practical terms, catching more hate speech, even at the expense of false positives, can be preferable to missing harmful content altogether.

Our attempt to fine-tune two of DistilBERT's top transformer layers didn't improve results. Minor gains in recall were offset by drops in precision and accuracy, suggesting the frozen base was already extracting effective features for our dataset. More aggressive or granular fine-tuning may be needed to benefit from unfreezing layers.

Limitations:

- **Low Precision for "Hate Speech":** Precision (0.197) for hate speech remains low, leading to many false positives. Practically, this necessitates human moderation to prevent over-censorship.
- **Nuance and Context:** Automated detection remains challenging due to subtleties in human language, such as sarcasm, irony, slang, and context-specific usage, posing inherent difficulties for the model.
- **Dataset Bias:** Our model reflects biases and definitions from the "Davidson et al." dataset. Performance might differ significantly across different platforms or demographics.
- **Scope of Harmful Content:** The model only targets three categories. It doesn't address other harmful behaviors, like misinformation or subtle radicalization.
- **Evolving Language:** Online speech continuously evolves, requiring periodic retraining or updates to maintain model accuracy.

Future Work & Next Steps:

To enhance precision and further improve recall, future steps include:

- **Data Augmentation:** Applying synonym replacement to the "Hate Speech" class could diversify training examples, improving the model's ability to distinguish subtle language variations. While beyond this project's immediate scope, this strategy is supported by prior coursework and could significantly enhance minority-class detection.
- **Hyperparameter Tuning:** Conducting more thorough tuning of the classification head and fine-tuning strategies.
- **Advanced Architectures:** Experimenting with alternative designs for the custom classification head or additional transformer layers.
- **Ensemble Methods:** Combining predictions from multiple models to boost overall performance.

9. Use-Case Section: Applications in Ethical AI

The improved sensitivity to hate speech in our DistilBERT-based model has important applications in building **ethical AI systems** and **safer digital spaces**. Though the model isn't ready for deployment without further tuning, its core functionality is already valuable across several domains.

1. Pre-filtering Training Data for AI Models

- **Challenge:** Large-scale datasets often contain harmful or biased content. If used unfiltered, these flaws are inherited by AI systems.
- **Solution:** Our model can serve as an **automated filter** to screen datasets before they're used for training, helping reduce downstream bias in large language models and other AI systems.

2. Supporting Content Moderation

- **Challenge:** Human moderators can't keep up with the volume of online content.
- **Solution:** The model can act as a **triage tool**, flagging high-risk posts for human review. It's not a replacement for moderators but can significantly speed up moderation workflows.

3. Monitoring Sociopolitical Events

- **Challenge:** Hate speech often spikes during elections, protests, or crises.
- **Solution:** The model can be deployed for **real-time monitoring**, helping NGOs, researchers, or media watchdogs track hate speech trends and respond quickly.

4. Building Specialized Filters

- **Challenge:** Online abuse can target race, gender, religion, and more, but generic models might miss these nuances.
- **Solution:** The model can be retrained on targeted datasets to detect specific types of abuse, such as misogynistic or racial hate speech, enabling more nuanced interventions.

Ethical Deployment Considerations: As with any AI moderation tool, precision and bias must be carefully managed. Given our model's tendency to over-flag hate speech, it should always be paired with human oversight, appeal mechanisms, and transparent documentation. Ongoing updates are also necessary to keep up with changing language patterns and cultural contexts.

10. Conclusion

This project successfully developed a DistilBERT-based model for classifying harmful language, with a strong focus on improving hate speech detection. Through class weighting and a well-regularized head architecture, we raised hate speech recall from **0.07 to 0.594** which is a crucial step toward ethical content moderation.

While precision remains low and fine-tuning deeper layers of DistilBERT yielded no additional benefit, the frozen-base model proved both efficient and effective for our dataset. A qualitative check supported these gains, showing real-world improvement in classification behavior.

Key limitations include dataset bias, limited generalizability, and the need for human oversight due to false positives. Future work should explore synonym-based data augmentation, more advanced fine-tuning, and specialized retraining for different types of abuse. Overall, this model serves as a strong foundation for building safer and more responsible AI tools.

Appendix A: Training Details

Model Architecture

Model: "functional"			
Layer (type)	Output Shape	Param #	Connected to
input_ids (InputLayer)	(None, 128)	0	-
attention_mask (InputLayer)	(None, 128)	0	-
lambda (Lambda)	(None, 768)	0	input_ids[0][0], attention_mask[0..
dense (Dense)	(None, 256)	196,864	lambda[0][0]
leaky_re_lu (LeakyReLU)	(None, 256)	0	dense[0][0]
batch_normalization (BatchNormalizatio..	(None, 256)	1,024	leaky_re_lu[0][0]
dropout (Dropout)	(None, 256)	0	batch_normalizat..
dense_1 (Dense)	(None, 128)	32,896	dropout[0][0]
leaky_re_lu_1 (LeakyReLU)	(None, 128)	0	dense_1[0][0]
batch_normalizatio.. (BatchNormalizatio..	(None, 128)	512	leaky_re_lu_1[0]..
dropout_1 (Dropout)	(None, 128)	0	batch_normalizat..
dense_2 (Dense)	(None, 32)	4,128	dropout_1[0][0]
batch_normalizatio.. (BatchNormalizatio..	(None, 32)	128	dense_2[0][0]
dropout_2 (Dropout)	(None, 32)	0	batch_normalizat..
dense_3 (Dense)	(None, 3)	99	dropout_2[0][0]
Total params: 235,651 (920.51 KB)			
Trainable params: 234,819 (917.26 KB)			
Non-trainable params: 832 (3.25 KB)			

Training Hyperparameters

Parameter	Value
Epochs	15 (early stopped at 92% val accuracy)
Batch size	16
Learning rate	3e-5
Optimizer	Adam

Weight decay	1e-4 (via L2 regularization in classifier layers)
Early stopping	Yes (custom callback at <code>val_accuracy ≥ 0.92</code>)
Frozen layers	Yes (pretrained DistilBERT frozen, custom classifier trained)
Hardware	Google Colab, T4 GPU

Carbon Emission:

Training emissions: 0.0273 kg CO₂

Note: Carbon emissions were measured for a single representative training run on Google Colab (T4 GPU). Actual values may vary depending on hardware and model configuration.

Sustainability

Model training was tracked using CodeCarbon, with total emissions estimated at **0.0273 kg CO₂**.

This is roughly equivalent to the energy used by a low-wattage light bulb running for 3 hours. While relatively low, it still underscores the cumulative environmental cost of large-scale model development and justifies the project's lightweight, resource-efficient approach.

Appendix B: Fairness & Robustness Audit

We evaluated the classifier's performance across gendered-pronoun slices using the `fairlearn MetricFrame` with macro-averaged precision and recall. Tweets were grouped by the presence of **he/him** vs. **she/her** pronouns..

	accuracy	precision	recall
sensitive_feature_0			
he/him	0.529954	0.51854	0.571862
she/her	0.810734	0.46706	0.579985

Table B.1 – Slice-level performance metrics for gendered pronoun groups

Interpretation:

The model's accuracy is substantially higher for tweets containing "she/her" pronouns (0.81) compared to "he/him" (0.53). However, macro-averaged precision and recall are similar across both groups, with slightly higher recall for "she/her". This suggests that while the model is better at **overall classification** for "she/her" tweets, **precision** remains relatively low for both groups. These results indicate possible differences in model confidence or data distribution between slices and highlight the importance of continuous fairness monitoring.

Note: Metrics are based on a single representative run; minor variations may occur across runs due to random initialization and data shuffling.

Appendix C: Model Card

This appendix summarizes the key details from the public Hugging Face Model Card for this project.

Model Name

Fine-Tuning DistilBERT for Ethical Content Moderation

Model Description

A fine-tuned DistilBERT model for automated detection of hate speech, offensive language, and neutral content in short texts (e.g., tweets). The model is trained on the Davidson hate speech and offensive language dataset, with a focus on ethical AI and responsible content moderation.

Intended Uses & Limitations

- **Intended Use:**
For academic research and prototyping of ethical AI systems in content moderation.
- **Not Intended For:**
Direct deployment in critical or real-world moderation systems without further validation and bias assessment.

Training and Evaluation Data

- **Dataset:** Davidson et al. (2017) “Hate Speech and Offensive Language” dataset, cleaned and preprocessed.
- **Task:** Multi-class classification (hate speech, offensive, neither).

Training Procedure

- **Model:** DistilBERT base, frozen transformer; custom classifier head trained.
- **Hyperparameters:**
 - Learning rate: 3e-5
 - Epochs: 15 (with early stopping)
 - Batch size: 16

- Optimizer: Adam
- **Hardware:** Google Colab, T4 GPU

Performance

- **Best validation accuracy:** 74.0%
- **See Appendix A for full training results.**

Sustainability & Fairness

- **Training emissions:** 0.0273 kg CO₂ (measured via CodeCarbon)
- **Fairness assessment:**
Performed targeted evaluation for gender pronoun bias. (Results and details in Appendix B.)

Model Availability

- **Public repository:**
[Hugging Face Model Card](#)