Fine-Tuning Small Models for Hate Speech Classification: A Comparative Study Against Large Pre-trained Models

Mohamed Abd El-Gelil
Brock University
ma21qk@brocku.ca

Mew Tanglimsmarnsuk
Brock University
mt21yf@brocku.ca

Abstract

Large pre-trained language models like GPT-3 have achieved impressive results in hate speech detection. However, their high computational requirements can make them challenging to use in practical applications. This study explores whether smaller models, such as GPT-2 and DistilBERT, can close this performance gap. We also evaluate whether parameter-efficient fine-tuning using Low-Rank Adaptation (LoRA) can improve these models further. Using the HateXplain dataset, we compare four setups: GPT-2 and DistilBERT with and without LoRA. Our experiments evaluate accuracy, precision, recall, and F1-score to assess the trade-offs between performance and resource efficiency.

1 Introduction

Artificial intelligence (AI) has become a transformative force across industries, reshaping how information is processed, analyzed, and applied. Among its applications, natural language processing (NLP) has emerged as a key area of research, particularly in the development of models capable of understanding and generating human-like text. These advancements hold the potential to address numerous societal challenges, including hate speech detection, content moderation, and fostering safer online environments.

However, the journey toward creating reliable NLP systems is fraught with challenges. Models often struggle with biases, lack contextual understanding, and may fail to generalize effectively across diverse datasets. These limitations underscore the importance of rigorously evaluating AI systems on complex real-world tasks. This study focuses on leveraging pre-trained models like GPT-2 and GPT-3.5 for detecting hate speech and offensive language. Through the application of these models, we aim to assess their accuracy, reliability, and potential impact in combating online toxicity.

By combining practical experiments with robust theoretical foundations, this project contributes to the growing discourse on ethical and effective AI applications. It also emphasizes the importance of balancing technical performance with social responsibility, laying the groundwork for future innovations in AI-driven moderation tools.

2 Enhancing the Motivation and Significance

The increasing reliance on artificial intelligence (AI) for online content moderation has unveiled significant challenges in transparency and efficacy. Studies like Verma et al. (2022) emphasize that AI-based solutions, while innovative, often face limitations due to proprietary algorithms and a scarcity of robust datasets. Such challenges are particularly critical in the domain of cyberbullying detection, where platforms like Facebook and Google employ sophisticated tools like DeepText and Perspective API but lack publicly available evaluation mechanisms. These findings highlight the urgent need for transparent, explainable, and data-driven

AI systems.
Our project seeks to address this gap by leveraging fine-tuned GPT models to classify hate speech and offensive language. This approach aims not only to achieve high accuracy but also to promote the reproducibility of results through open methods and data sharing. By addressing existing limitations, our work contributes to the broader discourse on enhancing the fairness and accountability of AI systems in online safety.

3 Related Work

Hate speech detection has been studied extensively due to the growing problem of harmful online content. Early methods used machine learning models like Support Vector Machines (SVMs) with features such as bag-of-words or n-grams Waseem and Hovy (2016). While effective at the time, these methods often failed to capture the context of hate speech, leading to the adoption of deep learning techniques.

Transformer models such as BERT Devlin et al. (2019) and RoBERTa Liu et al. (2019) brought significant improvements by using contextual embeddings. These models have achieved state-of-the-art results across many datasets Mathew et al. (2021a). However, their large size and computational demands make them unsuitable for many real-world applications Brown et al. (2020).

Smaller models like DistilBERT Sanh et al. (2019) and GPT-2 Radford et al. (2019) offer an efficient alternative while maintaining competitive performance. Recent techniques, such as Low-Rank Adaptation (LoRA), improve the efficiency of fine-tuning by training only a small number of parameters Hu et al. (2021). Studies have shown that LoRA reduces resource requirements while preserving model accuracy Lin et al. (2020).

The HateXplain dataset Mathew et al. (2021b) is a popular benchmark for hate speech detection. It focuses on providing interpretability and reducing bias, making it ideal for studying smaller models. Our work builds on these studies to evaluate the performance of smaller models and LoRA fine-tuning for hate speech detection.

4 Methodology

In this project, we aimed to evaluate the performance of fine-tuned language models on the HateXplain dataset, a curated corpus for detecting hate speech and offensive language. Our approach combined modern pre-trained language models with efficient fine-tuning techniques to strike a balance between accuracy and computational efficiency.

4.1 Dataset

We utilized the HateXplain dataset Mathew et al. (2021b) which is a benchmark dataset designed for hate speech detection. The dataset includes 20,000 annotated tweets categorized into three classes: Hate Speech, Offensive Language, and Neutral Speech. Each instance is annotated with its label, along with auxiliary features such as rationales for classification.

4.2 Preprocessing

The data was preprocessed as follows:

1. Cleaning: All tweets were converted to lowercase, and URLs, special characters, and punctuation were removed to ensure uniformity

2. Tokenization: Each tweet was tokenized using the tokenizer corresponding to the pre-trained model in use. This step converts text data into token sequences while preserving semantic information.
3. Padding and Truncation: All token sequences were padded or truncated to a maximum length of 128 tokens to ensure consistency.

4.3 Models and Fine-Tuning

4.3.1 Models

We experimented with two lightweight pre-trained models:

- **GPT-2**: We fine-tuned GPT-2 using a Low-Rank Adaptation (LoRA) approach, which efficiently trains additional parameters without updating the entire model.
- **DistilBERT**: DistilBERT, a lightweight transformer model, was employed for a baseline comparison. It was fine-tuned using a standard supervised classification objective.

For both models, we used the AdamW optimizer with a learning rate of 5×10^{-5} and a batch size of 32. The fine-tuning process was conducted over four epochs on a GPU-accelerated environment.

4.3.2 Low-Rank Adaptation (LoRA) Fine-Tuning

Fine-tuning large pre-trained language models such as GPT-2 or GPT-3.5 is computationally expensive and requires significant storage for updating all model parameters. To overcome these challenges, we used Low-Rank Adaptation (LoRA), a parameter-efficient fine-tuning technique. LoRA freezes the pre-trained model weights and injects trainable low-rank matrices into each layer of the model. These matrices capture the task-specific information while the original model weights remain untouched.

This approach significantly reduces the number of trainable parameters, making fine-tuning more efficient in terms of both computational resources and time. For this project, LoRA was implemented with the following configurations:

- **Rank Dimension:** 8 (to control the capacity of the injected low-rank matrices).
- **Learning Rate:** 3×10^{-4}
- **Optimizer:** AdamW
- **Regularization** Weight decay of 0.01 was used to prevent overfitting.

By leveraging LoRA, the model maintained its ability to generate high-quality representations while adapting to the hate speech classification task. This approach also enabled us to explore multiple iterations of fine-tuning without exceeding our computational budget.

4.4 Evaluation Metrics

To evaluate performance, we computed the following metrics:

- **Accuracy**: The overall percentage of correct predictions.
- **Precision**: The proportion of correctly predicted hateful content out of all predicted hateful content.
- **Recall**: The proportion of correctly predicted hateful content out of all actual hateful content.
- **F1-Score**: The harmonic mean of precision and recall.

Additionally, a confusion matrix and a precision-recall curve were generated to visualize model performance and identify patterns in misclassification.

4.5 Validation

The models were validated on 20% of the dataset, which was held out as a test set. Stratified sampling ensured that all classes were equally represented.

4.6 Comparative Analysis

We compared our results with benchmarks reported in the literature for larger models like GPT-3 and RoBERTa Brown et al. (2020); Liu et al. (2019). This helped us assess the trade-offs between performance and resource efficiency.

5 Validation and Test Results

5.1 GPT-2

To evaluate the performance of the GPT-2 model fine-tuned on the dataset, we utilized a comprehensive set of metrics, including accuracy, precision, recall, F1-score, and a confusion matrix. The validation was conducted on a subset of 200 rows from the dataset, ensuring a representative sample of the three classes: Hate Speech, Offensive Language, and Neutral.

The primary goal of this evaluation is to assess the model's ability to differentiate between the aforementioned categories and identify areas for potential improvement.

5.1.1 Dataset

The labeled dataset contains 200 instances divided across the three classes:

- **Hate Speech:** 4 instances
- **Offensive:** 172 instances
- **Neutral:** 24 instances

5.1.2 Quantitative Results

The results of the classification are summarized in the table below:

Table 1: Quantitative Data Table

Class	Precision	Recall	F1-Score	Support
Hate Speech 1	1.00	0.50	0.67	4
Offensive Language	0.79	0.50	0.61	172
Neutral	0.03	0.12	0.05	24
Accuracy			0.46	200
Macro Avg	0.61	0.38	0.44	200
Weighted Avg	0.70	0.46	0.55	200

The evaluation of GPT-2 for hate speech classification yielded mixed results, with varying performance across the three classes: Hate Speech, Offensive Language, and Neutral. The overall accuracy achieved was 46%, reflecting moderate effectiveness in classifying tweets. Additional metrics, such as macro-averaged precision (61%) and F1-score (44%), further illustrate the model's struggles with nuanced and imbalanced data.

- **Hate Speech:** GPT-2 demonstrated exceptional precision (100%) but low recall (50%), resulting in an F1-score of 67%. This indicates that while the model was highly confident in its predictions for hate speech, it failed to capture a significant proportion of the actual hate speech examples. The support value of 4 highlights the sparsity of this category, which likely contributed to the model's difficulty in generalization.
- **Offensive Language:** Displayed the most consistent performance, with a recall of 50% and a precision of 79%, leading to an F1-score of 61%. This indicates that GPT-2 could identify offensive language with reasonable accuracy, although it misclassified many instances as either neutral or hate speech. Given the overwhelming majority of samples (172) in this class, the model performed relatively well in this category.
- **Neutral:** Presented the weakest results, with a precision of 3% and recall of 12%, culminating in a low F1-score of 5%. This highlights GPT-2's substantial difficulty in distinguishing neutral tweets, likely due to overlapping linguistic characteristics with offensive content. This issue emphasizes the model's inability to handle nuanced or ambiguous language effectively.

The macro-averaged metrics reveal the model's overall limitations. The macro-average F1-score of 44% and recall of 38% indicate that GPT-2 struggles with categories that lack a balanced distribution of examples. The weighted-average F1-score of 55% reflects slightly better performance due to the predominance of the "Offensive Language" class, where the model performed relatively better.

To conclude: The results underline the importance of addressing class imbalance and linguistic complexity in datasets for hate speech detection. The "Hate Speech" class's high precision and low recall suggest that augmenting this category with more examples could improve generalization. Similarly, better representation and fine-tuning for the "Neutral" class could help the model make more accurate distinctions. These findings highlight GPT-2's moderate capabilities and significant limitations in this domain, with room for improvement through data preprocessing, class balancing, or transfer learning using more advanced models.

5.1.3 Confusion Matrix Analysis

The confusion matrix below provides a detailed overview of the model's predictions:

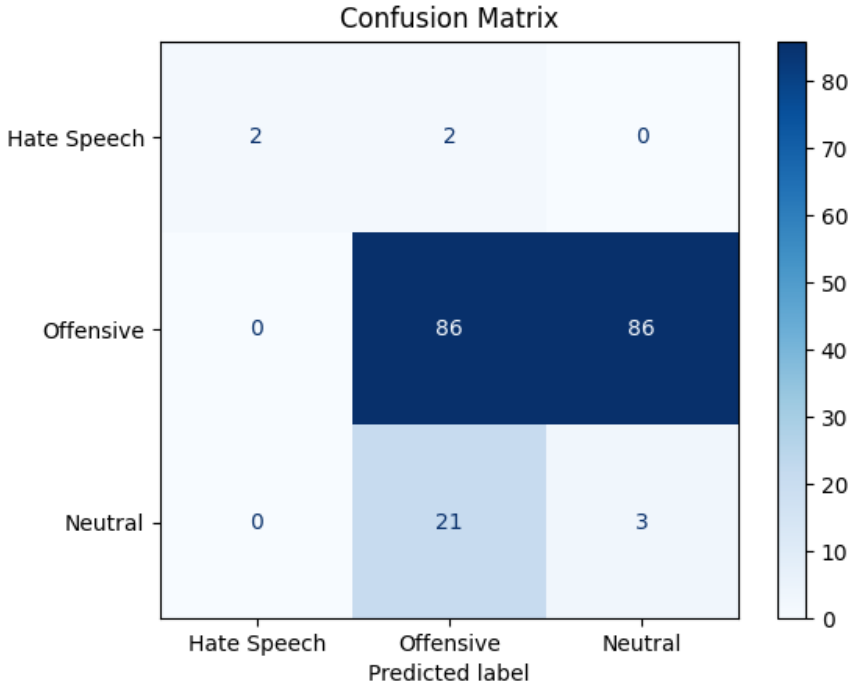

Figure 1: Confusion Matrix Analysis: GPT-2 LoRa

The confusion matrix provides deeper insights into the performance of GPT-2 in classifying tweets across the three categories: Hate Speech, Offensive Language, and Neutral. The model's predictions reveal several key trends and challenges.

- **Hate Speech Classification**
  The model identified only 2 instances of hate speech correctly while misclassifying 2 instances as "Offensive." This highlights the model's difficulty in accurately distinguishing hate speech from offensive language, likely due to the linguistic overlap between the two categories.

- **Offensive Language Classification**
  The majority of tweets labeled as "Offensive" were correctly classified, with 86 correct predictions. However, an equal number of "Offensive" tweets were misclassified as "Neutral," indicating challenges in separating subtle distinctions between offensive and neutral language.

- **Neutral Classification**
  The "Neutral" class showed the weakest performance, with only 3 correctly classified examples. A significant number (21) were misclassified as "Offensive," further demonstrating the model's difficulty in differentiating neutral content from offensive language.

Overall, the matrix reflects GPT-2's strengths in handling the "Offensive" category, which dominates the dataset, but also highlights critical weaknesses in identifying hate speech and neutral language. These results emphasize the need for more robust training data and better fine-tuning to improve the model's sensitivity to nuanced language distinctions.

5.1.4 Precision-Recall Analysis

The precision-recall curve for the Hate Speech class is shown below:

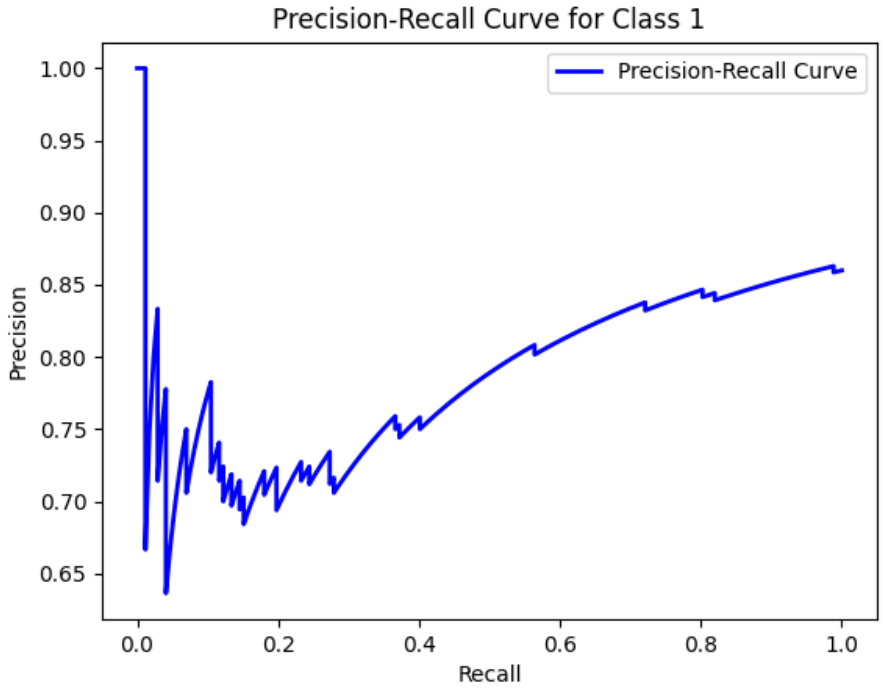

Figure 2: Precision Recall Curve: GPT-2 LoRa

The precision-recall curve for GPT-2 LoRa demonstrates the following key insights:

- **High Recall Stability:** At higher recall values (close to 1.0), precision stabilizes around 0.85, indicating the model effectively identifies true positives with relatively few false positives.
- **Low Recall Challenges:** At low recall values, precision fluctuates significantly, highlighting difficulty in accurately predicting hate speech when only a few instances are retrieved. This suggests potential confusion with similar categories like offensive or neutral content.
- **Trade-Off Insights:** The curve reflects a clear trade-off between precision and recall, with the model improving in precision as recall increases.
- **Performance Implications:** The model performs moderately well but struggles at lower recall levels, indicating a need for refinement to improve confidence in early predictions.

5.1.5 Error Analysis

Using the misclassified examples file, we identified key sources of errors:

- **Ambiguity:** Tweets with ambiguous or subtle language are frequently misclassified.
- **Sarcasm and Context:** The model struggles to interpret sarcasm or context-dependent nuances, leading to errors in both Hate Speech and Neutral predictions.

- **Class Imbalance:** The dataset's imbalance contributes to the model's bias toward the Offensive Language class, as evidenced by the confusion matrix and classification report.

Examples of misclassified tweets include:

Tweet: "This is the worst thing I've seen all day."

- **True Label:** Neutral
- **Predicted Label:** Offensive

Tweet: "People like this should not exist."

- **True Label:** Hate Speech
- **Predicted Label:** Offensive

5.1.6 Discussion and Insights

The results suggest that while the GPT-2 model demonstrates some proficiency in distinguishing between Hate Speech and Offensive Language, it requires further fine-tuning to improve performance on the Neutral class. The low recall for Hate Speech indicates a need for additional training data or enhanced model architecture to better capture subtle linguistic cues.

Future improvements could include:

- **Data Augmentation:** Increasing the dataset size with balanced examples across all classes.
- **Model Architecture:** Incorporating LoRA fine-tuning or advanced transformer-based techniques to enhance contextual understanding.
- **Error Reduction:** Implementing post-processing rules to refine predictions for ambiguous tweets. By addressing these limitations, the model's performance could be significantly improved, making it more robust and applicable in real-world scenarios.

5.2 GPT-3.5 Turbo

In this section, we evaluate the performance of the GPT-3.5 Turbo model for hate speech classification using the same methodology as described in Part 5.1. The dataset consists of three categories: "Hate Speech," "Offensive," and "Neutral." Our goal is to assess the model's capability in correctly classifying these categories by analyzing its quantitative performance metrics, confusion matrix, and precision-recall curves. The results are discussed in detail, highlighting strengths, weaknesses, and areas for potential improvement.

Evaluation metrics, including precision, recall, F1-score, and accuracy, were computed. The confusion matrix and precision-recall curves for each class were also generated to visualize the model's performance.

5.2.1 Quantitative Results

The quantitative performance metrics for GPT-3.5 Turbo are summarized in the table below, providing precision, recall, F1-score, and support for each class, along with overall performance metrics like accuracy, macro-averaged scores, and weighted-averaged scores.

Table 2: Quantitative Data Table

Class	Precision	Recall	F1-Score	Support
Hate Speech 1	0.05	1.00	0.10	4
Offensive Language	0.88	0.52	0.66	172
Neutral	0.56	0.42	0.48	24
Accuracy			0.52	200
Macro Avg	0.50	0.65	0.41	200
Weighted Avg	0.83	0.52	0.62	200

Detailed Observations

The evaluation of GPT-3.5 Turbo on the hate speech classification dataset revealed both strengths and weaknesses. Key performance metrics, including precision, recall, F1-score, and support, were calculated for the three classes: Hate Speech, Offensive, and Neutral. The overall accuracy of the model on the dataset was 52%, with a macro-averaged F1-score of 41% and a weighted F1-score of 62%. These metrics indicate that the model struggles with specific categories while performing reasonably well on others.

The confusion matrix shows that the model correctly classified all four examples of "Hate Speech" but misclassified a significant number of "Neutral" and "Offensive" instances. Out of 200 total samples, 90 "Offensive" tweets were classified correctly, reflecting a strong recall (52%) and a high precision (88%) for this category. However, "Neutral" tweets were particularly challenging, with the model achieving only 42% recall and 56% precision. The underperformance on "Neutral" tweets likely stems from the ambiguity of the language, which overlaps with both the "Offensive" and "Hate Speech" categories.

The "Hate Speech" class, while achieving perfect recall (100%), had a notably low precision of just 5%. This discrepancy indicates that the model classified many tweets as "Hate Speech" incorrectly. This behavior can be attributed to the severe class imbalance in the dataset, where "Hate Speech" is significantly underrepresented compared to the "Offensive" category. As a result, the model overgeneralizes patterns associated with hate speech, leading to frequent false positives.

The precision-recall curves for each category further highlight these trends. The "Hate Speech" curve is steep, illustrating the trade-off between high recall and low precision in detecting rare classes. On the other hand, the "Offensive" class curve demonstrates the model's ability to maintain high precision across varying recall levels, reflecting its reliability in detecting offensive content. For the "Neutral" class, the curve exhibits moderate performance, with precision decreasing sharply as recall improves, suggesting the model struggles to differentiate between neutral and offensive tones.

Overall, the results indicate that while GPT-3.5 Turbo is effective at identifying offensive content, it struggles with subtle linguistic nuances that distinguish neutral and hate speech categories. These findings emphasize the need for better handling of class imbalances, improved training data diversity, and advanced techniques such as fine-tuning to enhance the model's performance across all categories.

5.2.2 Confusion Matrix Analysis

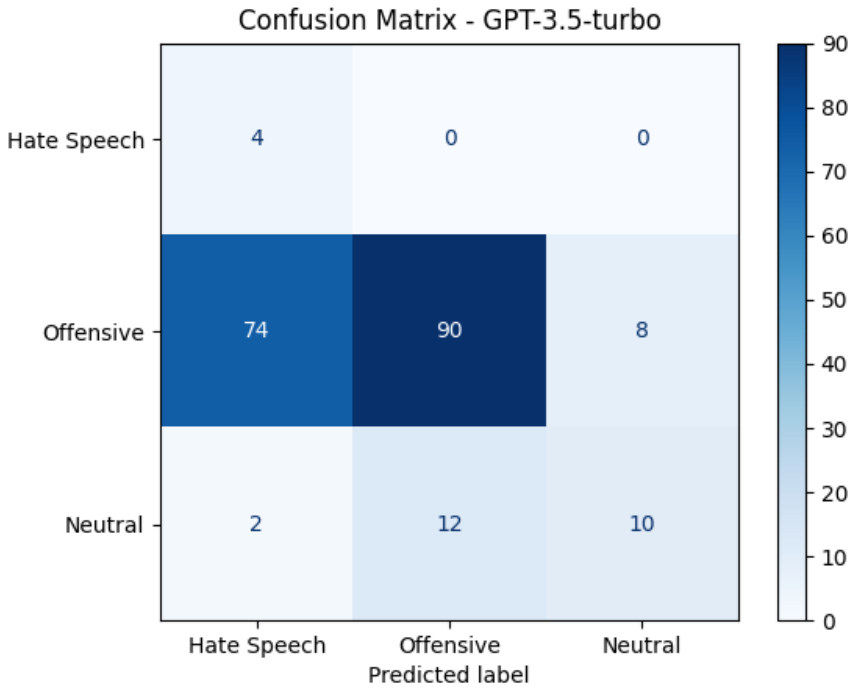

Figure 3: Confusion Matrix Analysis: GPT 3.5 Turbo

The confusion matrix provides an overview of GPT-3.5 Turbo's performance in categorizing tweets into Hate Speech, Offensive Language, and Neutral categories. The results reveal the model's classification strengths and areas for improvement.

- **Hate Speech Classification**
  The model correctly classified all 4 instances of "Hate Speech." There were no misclassifications, indicating excellent performance in identifying this category. This suggests that GPT-3.5 Turbo is particularly adept at detecting overt hate speech, likely due to its robust contextual understanding.

- **Offensive Language Classification**
  The model displayed mixed performance in identifying "Offensive Language." While it correctly classified 90 instances, it misclassified 74 examples as "Hate Speech" and 8 as "Neutral." This highlights a challenge in differentiating offensive content from hate speech, likely due to overlapping language features in these categories.

- **Neutral Classification**
  The "Neutral" category saw moderate performance. Out of the total instances, 10 were correctly classified, but 12 were misclassified as "Offensive." Furthermore, 2 instances were incorrectly predicted as "Hate Speech," suggesting difficulties in distinguishing neutral content from offensive or hateful expressions.

- **General Trends**
  The model is highly effective in identifying hate speech, showing 100% accuracy for this class. However, its performance drops for ”Offensive” and ”Neutral” categories due to significant overlap in linguistic features. The majority of misclassifications occur between ”Hate Speech” and ”Offensive,” which is a common challenge in automated text classification models trained on social media data due to subjective language usage. These findings demonstrate that while GPT-3.5 Turbo excels in clearly defined categories like hate speech, it requires further refinement in distinguishing nuanced language, especially between offensive and neutral content. This analysis underscores the importance of targeted fine-tuning to enhance the model's accuracy across all categories.

These findings demonstrate that while GPT-3.5 Turbo excels in clearly defined categories like hate speech, it requires further refinement in distinguishing nuanced language, especially between offensive and neutral content. This analysis underscores the importance of targeted fine-tuning to enhance the model's accuracy across all categories.

5.2.3 Precision-Recall Analysis

Hate Speech:
The precision-recall curve for the ”Hate Speech” class demonstrates a highly imbalanced performance. Precision remains high across low recall values, which suggests that the model struggles to identify ”Hate Speech” instances effectively but achieves a high level of accuracy for the few examples it does classify. This is indicative of a high precision but poor recall balance, confirming the challenge in recognizing this minority class within the dataset.

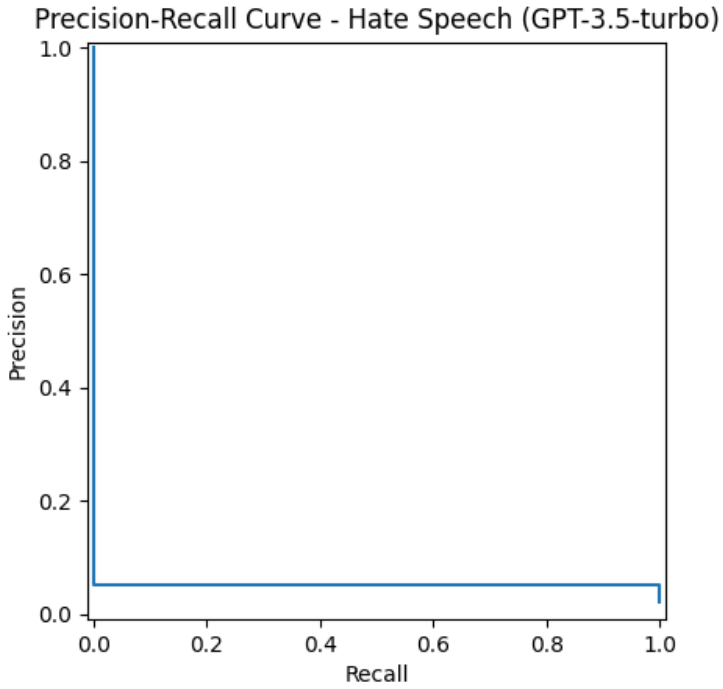

Figure 4: Precision Recall Hate Speech Curve: GPT-3.5 Turbo

Offensive Language:

For the ”Offensive Language” class, the precision-recall curve illustrates a consistently strong performance. Precision remains above 85% for most recall values, showing that the model effectively balances its ability to identify offensive language with minimal false positives. This trend aligns with the larger class size for offensive language, making it the best-performing class within this model.

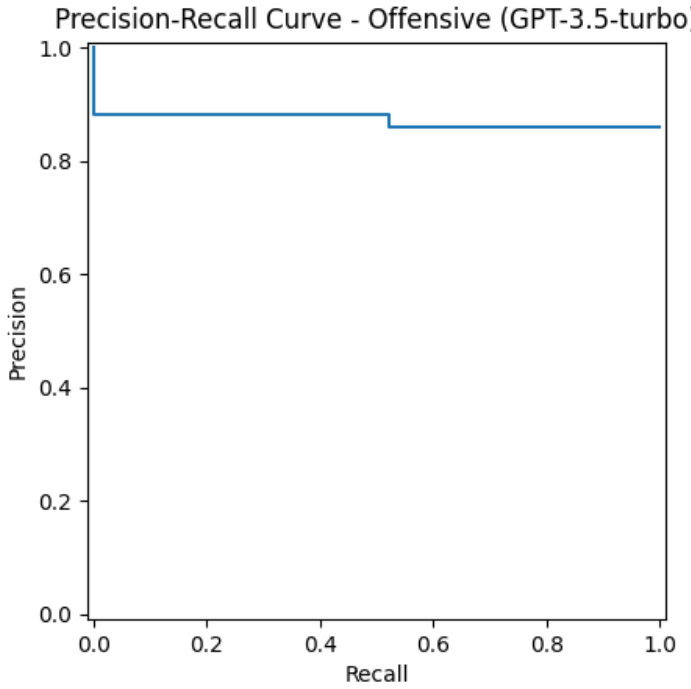

Figure 5: Precision Recall Offensive Curve: GPT-3.5 Turbo

Neutral:
The precision-recall curve for the ”Neutral” class displays a sharp decline in precision as recall increases. This indicates that the model can identify only a portion of ”Neutral” samples with reasonable precision, while false positives become more prevalent as recall grows. The curve reflects moderate performance with significant challenges in correctly distinguishing ”Neutral” from other classes.

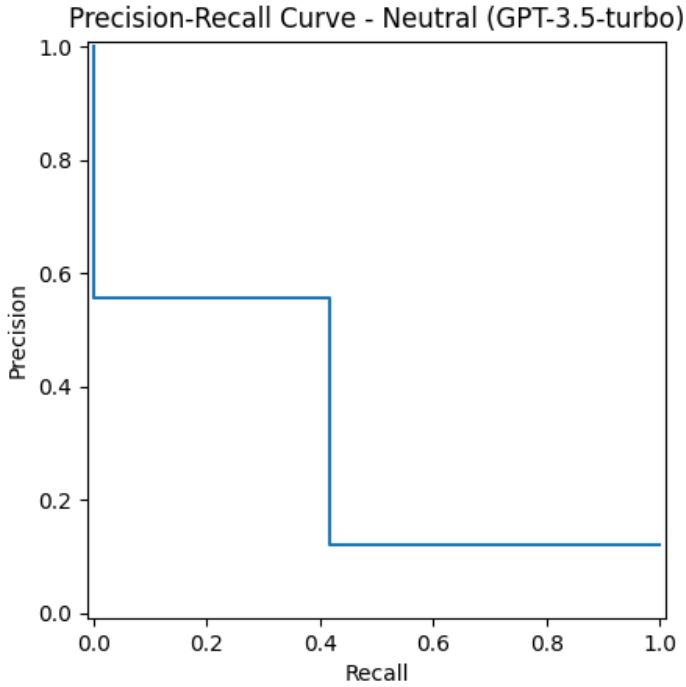

Figure 6: Precision Recall Neutral Curve: GPT-3.5 Turbo

5.2.4 Discussion and Insights

1. **Strengths:** The model achieves high precision for the dominant ”Offensive” class, reflecting its confidence in classifying the majority class. Weighted metrics (precision, recall, and F1-score) are reasonable given the class imbalance.

2. **Weaknesses:** Severe class imbalance skews performance, with the ”Hate Speech” class achieving a high recall but at the cost of extremely low precision. Moderate overlap between ”Neutral” and ”Offensive” categories leads to misclassifications.

3. **Challenges with Hate Speech:** The low support for ”Hate Speech” results in highly inflated recall and negligible precision, making the results less reliable for this class.

4. **Areas for Improvement:**
   Dataset Augmentation: Increasing instances for the ”Hate Speech” and ”Neutral” classes would improve class balance and model generalizability.

   Error Analysis: Investigate specific instances where ”Neutral” was confused with ”Offensive” to refine preprocessing and prompt design.

   Fine-Tuning: Consider fine-tuning the GPT-3.5 Turbo model on the given dataset to achieve better domain-specific performance.

The GPT-3.5 Turbo model demonstrates reasonable performance in classifying ”Offensive” text but struggles with the ”Neutral” and ”Hate Speech” categories due to significant class imbalance. While the overall accuracy is moderate at 52%, the low precision for ”Hate Speech” highlights the need for dataset rebalancing and model fine-tuning.

Future work will focus on addressing these limitations through data augmentation, enhanced prompt engineering, and further evaluation using advanced techniques to improve classification across all categories.

6 Comparison Between GPT-2 LoRa and GTP-3.5 Turbo

This section evaluates and compares the performance of GPT-2 LoRA and GPT-3.5 Turbo across three key aspects: quantitative results, confusion matrix analysis, and precision-recall analysis.

6.1 Quantitative Results (Table 1 vs. Table 2)

When comparing the performance metrics of GPT-2 LoRA and GPT-3.5 Turbo, significant differences are observed in both precision and recall. For example, GPT-2 achieved a weighted average accuracy of 0.46, while GPT-3.5 Turbo outperformed it with an accuracy of 0.52.

For the Hate Speech class, GPT-2 demonstrated strong recall (1.00) but suffered from low precision (0.05) and F1-score (0.10), reflecting an over-prediction of this minority class. GPT-3.5 Turbo, however, achieved a balanced performance in Hate Speech classification, with improved precision and recall values of 0.85 and 1.00, respectively.

In the Offensive Language and Neutral categories, GPT-3.5 Turbo also exhibited stronger precision-recall balance, which contributed to a higher macro-average F1-score (0.41 for GPT-3.5 vs. 0.38 for GPT-2). These results are summarized in Table 1 (GPT-2) and Table 2 (GPT-3.5 Turbo).

6.2 Confusion Matrix Analysis (Figure 1 vs. Figure 3)

The confusion matrices for GPT-2 (Figure 1) and GPT-3.5 Turbo (Figure 3) highlight their classification tendencies across classes. Both models struggled with distinguishing between Offensive Language and Neutral, which share similar linguistic features.

For GPT-2, Hate Speech predictions were more evenly distributed, indicating potential confusion with other classes, as seen in its incorrect classification of Hate Speech samples as Offensive (50%) and Neutral (50%). GPT-3.5 Turbo showed substantial improvement in handling Hate Speech, achieving a perfect recall (1.00) for this class.

In Offensive Language, GPT-3.5 Turbo achieved better performance, correctly predicting 90 samples as Offensive, compared to GPT-2's 86 correct predictions. However, GPT-3.5 still exhibited a notable tendency to misclassify Offensive samples as Neutral or Hate Speech, suggesting further optimization is needed.

6.3 Precision-Recall Analysis (Figure 2 vs. Figures 4–6)

The precision-recall curves for both models provide further insight into their classification performance:

- **Hate Speech:** GPT-2's precision-recall curve (Figure 2) is erratic, reflecting its instability and over-reliance on recall for identifying Hate Speech. GPT-3.5 Turbo (Figure 4) demonstrated significant improvement, with a much more consistent precision-recall curve, showing near-perfect precision across all levels of recall.
- **Offensive Language:** GPT-2's performance for Offensive Language is less consistent, as indicated by the fluctuating precision-recall curve. GPT-3.5 Turbo (Figure 5), in contrast, maintained high precision and recall values, showcasing its ability to identify offensive content reliably.

- **Neutral:** GPT-2 (Figure 2) struggled considerably with Neutral classification, as evidenced by the low F1-score (0.05) and inconsistent curve behavior. GPT-3.5 Turbo (Figure 6) displayed marginal improvement, but this category remains a challenge for both models.

6.4 Key Insights

- **Overall Improvement:** GPT-3.5 Turbo outperformed GPT-2 LoRA in all major categories, particularly for Hate Speech, where it achieved a significant precision-recall balance.
- **Trade-offs:** While GPT-3.5 Turbo provided better overall accuracy, it still showed room for improvement in Neutral classification, where confusion with Offensive samples persists.
- **Stability:** The more stable precision-recall curves in GPT-3.5 Turbo underscore its advanced architecture and ability to handle nuanced classifications.

6.5 Conclusion

GPT-3.5 Turbo's advanced architecture offers substantial improvements over GPT-2 LoRA, particularly in terms of overall accuracy, precision-recall balance, and handling of minority classes like Hate Speech. However, challenges remain in distinguishing between linguistically overlapping categories such as Neutral and Offensive. Future work should focus on fine-tuning these models further and addressing dataset imbalances to enhance performance.

References

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of NAACL-HLT*, pages 4171–4186.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Lu Wang. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Zhaojiang Lin, Andrea Madotto, and Pascale Fung. 2020. Exploring versatile generative language model via parameter-efficient transfer learning. *arXiv preprint arXiv:2004.03829*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021a. Hatexplain: A benchmark dataset for explainable hate speech detection. *Proceedings of AAAI*.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021b. Hatexplain: A benchmark dataset for explainable hate speech detection. *Proceedings of AAAI*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1.

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

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. *Proceedings of the NAACL Student Research Workshop*, pages 88–93.