National College of Ireland

Fine tuning GPT- 3.5 for hate speech detection with continuous severity scoring

MSc Research Project
MSc in Data Analytics

Ajay Gurrala
X23427931

School of Computing
National College of Ireland

Supervisor: Sallar Khan

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Ajay Gurrala
Student ID:	X23427931
Programme:	Msc in Data Analytics **Year:** 2025-2026
Module:	Msc Research Project
Supervisor:	Sallar Khan
Submission Due Date:	28-01-2026
Project Title:	Fine tuning GPT-3.5 for hate speech detection with continuous severity scoring
Word Count:	9128 **Page Count** 26

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Ajay Gurrala

Date: 28-01-2026

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Fine tuning GPT-3.5 for hate speech detection with continuous severity scoring

Ajay Gurrala
X23427931

Abstract

The spread of the harmful content in the social media platforms requires automated detection systems, which can recognize the severities in a nuanced manner that goes beyond binary classification. The study presents a very significant gap because it is the first supervised GPT-3.5 fine-tuning on continuous hate speech severity data, based on the UC Berkeley Measuring Hate Speech corpus. An innovative three-way classification scheme (HATE/ ABSTAIN/ NOTHATE) with score-enriched output mode was adopted, which allowed categorical prediction, as well as continuous estimation of severity at the same time. The fine-tuned model was compared with three base models BiLSTM, BERT-base and RoBERTa-base--with the same data partitions of 26, 557 training samples and 11, 382 test samples. The experimental findings show that fine-tuned GPT-3.5 is more efficient in all metrics: 70.68% classification accuracy (compared with 68.25 in RoBERTa-base), 0.6882 macro F1-score, 0.8862 mean absolute error and 0.8390 Pearson correlation of prediction of severity. Analysis of errors showed that the critical cases of HATE-NOTHATE misclassification were only found in 1.94% of the samples. The study can add a new dual-task learning model that keeps the continuous severity information that binary-based methods usually ignore, and it has an impact on content moderation systems that need to make both categorical and severity-based prioritization decisions.

1 Introduction

1.1 Background

The popularity of social media has radically altered the way in which individuals socially communicate, share information and engage in mass dialogue. Twitter, Reddit and YouTube form jointly billions of user-created comments daily, which to the existing difficulties in keeping online spaces safe represent nothing less than an opportunity unimaginable before. Hate, which can be described as expressions of hostility, encouragement of violence, dehumanisation of people by protected characteristics such as race, ethnicity, religion, gender, and sexual orientation, is one of the most pressing issues of platform operators, policymakers, and society in general.

Traditional content-moderation systems have relied on human operators, who usually locate and remove the offensive content. But the sheer volume of user-created content has rendered the process of human moderation at least practically and economically unsustainable. This fact has triggered significant research of automated hate speech detection systems that can recognize harmful material at a large scale (Albladi et al., 2025). The recent breakthroughs in the field of natural language processing and the creation of transformer design-based systems

and large language models in particular have already shown potential to interpret subtle linguistic patterns related to a harmful content (Guo et al., 2023; Bauer et al., 2024).

Regardless of this development, most of the currently implemented automated systems consider the process of hate speech detection as a binary classification problem, where the material is either hostile or not. This binary framework does not reflect the inherent shades of intensity that are used to come up with harmful content of the real world, where feelings are at one end of a sense of mildness and at the other end blatant calls to violence. To ensure that the most severe infractions get priority against the limited resources of human review, the currently utilized content moderation systems should be able to evaluate the severity of their infractions in addition to category classification (Hashir et al., 2025).

1.2 Research Gap and Motivation

The severity quantification problem is tackled with the UC Berkeley Measuring Hate Speech corpus, which has been publicly available since 2020 and offers continuous Rasch-scaled scores of -8.34 (counter-speech) to +6.30 (extreme hate) on almost 40,000 social media comments. This is a useful dataset in terms of creating models to judge severity in a subtle way because it is methodologically sound and created through Rasch Measurement Theory and has annotations of over 7,900 different annotators. Nonetheless, based on literature review, it is evident that this continuous severity data is not fully used yet.

When trained to perform hate speech detection tasks, GPT-3.5 scores high on such tasks with the help of prompting strategies and has F1 scores ranging between 0.75 and 0.87 on all benchmark datasets (Guo et al., 2023; Deroy and Maity, 2024). Nevertheless, the ability of this model to acquire dataset-specific patterns and gradations of severity is limited by the fact that these assessments only use zero-shot or few-shot prompting and do not update the parameters. The only known study using fine-tuning of a member of the GPT-family used GPT-2, a much weaker model, on binary classification data with no continuous severity measurements (Choudhary et al., 2025). None of the current studies has used supervised fine-tuning on GPT-3.5 to detect hate speech and quantify it on a continuous severity scale, which is an important methodological gap with practical implications on the development of content moderation systems.

1.3 Research Question and Objectives

This research addresses the identified gap through the following research question:

Can fine-tuned GPT-3.5, utilising a novel three-way classification system (HATE/ABSTAIN/NOT_HATE) with continuous severity scoring, achieve competitive performance compared to specialised transformer architectures?

To answer this question, four specific objectives guide the research:

Objective 1: To conduct the first supervised fine-tuning of GPT-3.5 on the UC Berkeley Measuring Hate Speech dataset, preserving continuous severity score information throughout the training process.

Objective 2: To implement a novel three-way classification system with score-enriched output format enabling simultaneous categorical prediction and continuous severity estimation within a unified model output.

Objective 3: To develop and evaluate baseline models including BiLSTM, BERT-base, and RoBERTa-base using identical data splits and dual-head architectures enabling fair comparative assessment.

Objective 4: To compare classification performance (accuracy, precision, recall, F1-score) and regression performance (MAE, RMSE, R^2, Pearson correlation) across all models to determine whether fine-tuned GPT-3.5 achieves competitive or superior performance.

1.4 Success Criteria

The achievement of the study objectives will be considered according to specific standards. To accomplish Objective 1, the fine-tuning process should be passed, and the model should be deployed through the API of OpenAI. To satisfy Objective 2, a stable output format with a parsing success rate greater than 95% on test data would have to be presented. All three baseline architectures will have to be trained and assessed to achieve Objective 3. To accomplish Objective 4, a deep statistical comparison based on the recorded data on the performance of each model is required.

1.5 Methodology Overview

The study employs an experimental design in order to compare three baseline topologies and a refined GPT-3.5 model. This is enabled by preprocessing the training and evaluation data of the UC Berkeley Measuring Hate Speech corpus to maintain the continuous severity information so as to classify it in three ways based on score-based thresholds. In order to be able to directly compare the performances of all the models, they are evaluated based on the classification and regression measures on the same held-out test data.

1.6 Report Structure

In this report, there are eight chapters. Chapter 2, after this introduction, contains a critical analysis of the literature available on the topic of hate speech detection based on large language models. Chapter 3 explains the research methodology that entails data preparation procedures and experimental design. Chapter 4 is a description of design specification and system architecture. Implementation specifics such as used tools and technologies are discussed in Chapter 5. The results of evaluation are given in chapter 6 with thorough statistical analysis. Chapter 7 is devoted to critical discussion of findings in reference to available literature. The contribution and future research directions are evaluated in chapter 8.

2 Related Work

2.1 Introduction

The use of automated hate speech detection has become the focus of a significant amount of research because social media websites are in need of a solution to content moderation problems in a scalable manner. This chapter critically reviews existing practices with the use of large language models, organized into three themes, including GPT-based prompting

strategies, fine-tuning strategies, and alternative architecture of LLM. The review poses a significant research gap on the topic of supervised fine-tuning of GPT-3.5 to predict continuous severity.

2.2 GPT-Based Prompting Approaches

More recent research has given a thorough assessment of GPT models through prompting strategies and no parameter updates. Table 2 presents a concise overview of four major studies that show that GPT-3.5 has strong baseline performance by the use of different prompting methods.

Table 2: Summary of GPT-Based Prompting Studies

Study	Method	Dataset(s)	Best Performance	Key Finding	Limitation
Guo et al. (2023)	Chain-of-Thought prompting	HateXplain, COVID-HATE, CallMeSexist	F1=0.87, Acc=0.85	CoT improves zero-shot by 7.9-24.2%	No fine-tuning; ignores continuous severity
Bauer et al. (2024)	Zero/few-shot GPT-3.5 & GPT-4	TweetEval framework	GPT-3.5: Acc=0.79, F1=0.82	Outperforms BERT (F1=0.66) and RoBERTa (F1=0.76)	Binary classification only; no parameter updates
Deroy & Maity (2024)	Zero-shot with temperature tuning	Twitter hate speech	Macro-F1=0.751-0.756	Consistent predictions (0.005 variance)	General-purpose model without task adaptation
Roy et al. (2023)	Prompt variations with target info	HateXplain, Implicit Hate, ToxicSpans	F1=0.39-0.68 (74% improvement)	Target information improves performance 20-30%	Underperforms fine-tuned baselines

All of these studies indicate that GPT-3.5 has high baseline performance (F1 scores 0.39-0.87) with general weaknesses in the form of failure to update its parameters. The models learn no dataset-specific patterns and all the assessments are categorical not taking into account continuous assessment of severity, which is important in content prioritisation.

2.3 Fine-Tuning Approaches

The use of parameter updates in GPT-family models is still a relatively low-research topic. It was possible to identify only two pertinent studies:

Choudhary et al. (2025) carried out a study to fine-tuned a GPT-family model by GPT-2 on the binary HatEval dataset (13,000 tweets). Fine-tuning had F1=61.64 which was 2.64 percentage points higher than zero-shot (F1=58.0). Although the research confirms that supervised fine-tuning training is helpful in GPT models, it relies on the much less reliable GPT-2 architecture and binary classification not based on severity.

Gouliev and Jaiswal (2024) compared GPT-3.5 with transfer learning against LSTM and BERT on Davidson gave 88% accuracy. The transfer learning method is however different to supervised fine-tuning with parameter updates and the research is applied on categorical labels alone with no severity prediction in continuum.

2.4 Alternative LLM Architectures

There have been a number of studies that studied non-GPT architectures and they found different approaches and performance properties:

Table 3: Alternative LLM Architectures for Hate Speech Detection

Study	Model(s)	Innovation	Performance	Limitation
Hashir et al. (2025)	Mistral-7B + DistilBERT + RoBERTa	Dual-embedding with rationale extraction	ETHOS: F1=82.01%; HateXplain-Gab: F1=90.32%	Focus on explainability, not severity; 27pp platform variance
Ghorbanpour et al. (2025)	LLaMA-3.1, Qwen2.5, Aya-101, BloomZ	Multilingual evaluation (8 languages)	Variable by language and prompt design	Categorical only; no severity measurement
Barakat & Jaf (2025)	Meta-Llama-3, Phi-3, Mistral-7B, WizardLM	Q4_0 quantisation for deployment	Meta-Llama: Acc=76.13%, F1=75.65% (binary)	No continuous severity; model-specific performance gaps

These architectures are competitive in their performance, but they always do not have the ability of continuous severity prediction as they emphasize on explainability, multilingual support, or deployment efficiency.

2.5 Systematic Review Findings

Albladi et al. (2025) conducted a synthesis of more than fifty papers on the use of LLMs in hate speech detection, and found that the area is struggling with ongoing issues such as increase in bias, cross-lingual performance gaps, and computational expenses. Importantly, the review showed that no studies used a combination of supervised fine-tuning of GPT-3.5 or more advanced models and continuous severity prediction, which affirms the novelty of the current research method.

2.6 Research Gap Identification

As demonstrated in the literature review, there are three gaps that interrelate with each other:

Gap 1: No Supervised Fine-Tuning of GPT-3.5 - Although GPT-3.5 demonstrates high levels of prompting performance (F1=0.39-0.87), none of the studies has utilized supervised fine-tuning of this model to detect hate speech. The only GPT fine-tuning experiment was based on the weaker GPT-2 model**.**

Gap 2: Absence of Continuous Severity Prediction - All the papers in the review use categorical classification (binary or multi-class), which excludes continuous severity data, which is necessary in content prioritisation. The UC Berkeley MHS corpus which has Rasch-scaledization (-8.34 to +6.30) is not utilized to train the LLM.

Gap 3: Lack of Unified Dual-Task Learning - The extant literature does not learn models to predict categorical predictions and continuous scores of severity simultaneously, and thus in

practice cannot be used where both a classification and a prioritisation are needed, which is typically the case in practical settings.

The study fills these gaps by performing the initial supervised fine-tuning experiment of GPT-3.5 on continuous severity data, as well as with a new three-way (HATE/ABSTAIN/NOT_HATE) classification system with score-enriched output that retains the severity data normally lost when using categorical methods. It allows it to directly compare classification accuracy and severity prediction with baseline architectures and goes beyond the existing paradigms of prompting-only or binary classification.

3 Research Methodology

3.1 Research Design

The study implies the use of the experimental type of research to assess the applicability of the fine-tuned GPT-3.5 to hate speech detection using continuous severity grading. To obtain an objective appraisal, each model has been trained and tested on the same data partitions permitting a direct comparison of the proposed method with three base designs. The process consists of four stages, such as building a baseline model, comparative assessment, GPT-3.5 fine-tuning, and data preparation.

Figure 1: Data Preparation Pipeline for Hate Speech Classification

Figure 1 illustrates the complete data preparation pipeline, while Figure 2 presents the overall modelling methodology employed throughout the research.

3.2 Dataset Selection

Three criteria that supported the objectives of the study were used to select the UC Berkeley Measuring Hate Speech corpus. The continuous Rasch-scaled severity scores on the dataset make regression-based analysis and categorical classification possible. Second, the collection is filled with content of numerous social media platforms, which provides numerous linguistic patterns. Third, annotation procedure is used to use multi-annotator consensus and Rasch Measurement Theory to ensure that measurements are valid. Table 4 provides the key features of the dataset.

Table 4: UC Berkeley MHS Dataset Characteristics

Characteristic	Value
Total Annotation Rows	135,556
Unique Comments	39,565
Unique Annotators	7,912
Annotation Method	Rasch Measurement Theory
Severity Score Range	-8.34 to +6.30
Platform Distribution	YouTube (33%), Reddit (34%), Twitter (33%)
Identity Groups Covered	42 subgroups across 8 categories

3.3 Data Preparation Procedure

The data preparation process as described in Figure 1 converted raw annotations to formatted training and evaluation sets in 6 sequential steps.

3.3.1 Deduplication

The raw dataset consisted of several annotations on the same unique comment, and annotators were making severity ratings of the same text separately. To derive consensus scores, the mean of all severity scores devoted to each separate text were calculated averaging 135,556 annotation rows to 39,565 comments. This averaging technique eliminates redundancy and statistical properties of Rasch-scaled observations. The mean length of the annotations per remark was 3.4 and, based on the sample strategy employed during the initial data collection, there are 1 to 815 annotations on each comment.

3.3.2 Quality Filtering

Quality filters were used to make sure that there is adequate contextual information to be classified as well as establishing reasonable sequence lengths to be processed through the model. Those comments that had less than five words were not factored in as there was no enough context to ascertain the severity of the comment. Longer than 150 word comments were removed during training, the comments were so long that they would have eaten up computational efficiency. A total of 1,626 samples were filtered out (4.11 percent of unique comments), leaving 37,939 quality-filtered samples to process.

3.3.3 Three-Way Labeling

The three-way categorical labeling schema was introduced according to the thresholds of the severity scores calculated with reference to the documentation of the datasets and analysis of the distribution of scores. The ABSTAIN category was developed to offer a clear quantification of uncertainty that binary classification methods do not offer, thus able to capture the most technically ambiguous content that do not readily fit into either the hostile or supporting category. Table 5 shows the category distributions that have been produced by the labeling thresholds.

Table 5: Three-Way Labeling Schema

Category	Score Threshold	Interpretation	Count	Percentage
HATE	Score ≥ 0.5	Hateful, attacking, dehumanising content	9,752	25.70%
ABSTAIN	-1 ≤ Score < 0.5	Ambiguous, neutral, borderline content	11,135	29.35%
NOT_HATE	Score < -1	Supportive, counter-speech, non-hateful content	17,052	44.95%

3.3.4 Tokenisation Analysis

The GPT-3.5 tokeniser was used to compute token counts to ensure that the model context limits are compatible and can estimate the cost of training. The average number of tokens per comment was 34.4 tokens, and the maximum number of tokens was 199 that is much less than the context limit of 4,096 tokens. The sum of corpus in all samples was 1,306,098.

3.3.5 Data Splitting

The stratified sampling with constant random seed (42) was used to subdivide data into training and test subsets but maintaining categories distribution across splits. The split ratio was 70 to 30 in such a way that the reserve of enough samples could be considered statistically sound and at the same time adequate training data could be given. This created a split of statistics as shown in Table 6 where the distribution differences in categories were found to be below 0.01 percentage points indicating that stratification was effective.

Table 6: Data Split Statistics

Partition	Total Samples	HATE	ABSTAIN	NOT_HATE
Training Set	26,557 (70%)	6,826 (25.70%)	7,795 (29.35%)	11,936 (44.94%)
Test Set	11,382 (30%)	2,926 (25.71%)	3,340 (29.34%)	5,116 (44.95%)
Total	**37,939**	**9,752**	**11,135**	**17,052**

3.3.6 Output Formatting

The training data was exported in two formats based on diverse model training requirements. In the case of GPT-3.5 fine-tuning, the API of OpenAI was utilized to produce JSONL format and the standard machine learning frameworks were utilized to produce CSV format to train the baseline models.

3.4 Score-Enriched Output Format

An innovative form of output was developed that would allow simultaneous categorical classification and continuous severity prediction to be made within the response of a single

model. This format is a combination of the categorical decision and the severity score related to it:

[CATEGORY] | Score: [value]

An example is that a moderately hateful comment will yield the output of HATE Score: 2.3 and supportive counter-speech will yield the output of NOTHATE Score: -2.8. This format maintains the actionable categorical decision needed to make content moderation, as well as the continuous estimate of severity, which allows prioritisation of review queues.

3.5 JSONL Training Data Structure

GPT-3.5 fine-tuning training examples were formatted to the chat completions specification of OpenAI. Every training example has three messages in a conversation format, a system message, a user message, and an assistant message.

Example 1: JSONL Training Instance Structure

```
{
 "messages": [
  {
   "role": "system",
   "content": "You are a hate speech classifier that analyzes social media
              comments. Classify each comment into one of three categories
              and provide a severity score.
              Categories:
              - HATE (score >= 0.5): Content that expresses hate, attacks,
                dehumanizes, or promotes violence against individuals or
                groups based on protected characteristics
              - ABSTAIN (score -1 to 0.5): Ambiguous, neutral, or mildly
                offensive content that doesn't clearly qualify as hate
                speech or supportive speech
              - NOT_HATE (score < -1): Supportive, counter-speech, or
                clearly non-hateful content

              Output format: [CATEGORY] | Score: [value]"
```

```
  },
  {
   "role": "user",
   "content": "Classify: [social media comment text]"
  },
  {
   "role": "assistant",
   "content": "[CATEGORY] | Score: [value]"
  }
 ]
}
```

The classification task, the boundaries of the categories, and the level of the scores which are required are all defined in the system message. The message that the user is commenting on is indicated in the user message with the same prefix. The target output with the ground truth category and severity score are also contained in the assistant message and display the desired response format.

3.6 Baseline Model Selection

Three baseline architectures were chosen to indicate three different methods of text classification, allowing to evaluate the performance of fine-tuned GPT-3.5 performance in relation to the existing methods.

BiLSTM was selected as an example of the recurrent neural network structure, which provides a comparison with computationally economic non-transformer models that are commonly deployed in production systems nonetheless.

BERT-base was selected, to use as the benchmark in text classification as well as the transformer architecture in natural language comprehension**.**

RoBERTa-base was selected because it is representative of the best practice in encoder-based categorization and is an optimized variant of BERT with improved pretraining procedures.

The training data, evaluation data, and dual-head architecture were the same across all the baseline models so that they could be fairly compared. Table 7 summarises specifications of the baseline model.

Table 7: Baseline Model Specifications

Model	Architecture	Parameters	Embedding Dimension
BiLSTM	2-layer Bidirectional LSTM	5,852,188	300

BERT-base	12-layer Transformer Encoder	109,876,996	768
RoBERTa-base	12-layer Transformer Encoder	125,040,388	768

3.7 Evaluation Framework

3.7.1 Classification Metrics

There are four measures of the 3-way categorization performance:

Accuracy = (Correct Predictions) / (Total Predictions). Chosen as an initial measure because of an equal distribution of categories (HATE: 25.70%, ABSTAIN: 29.35%, NOTHATE: 44.95%), meaning an overall measurement is obtained, not with the imbalanced data sets, where the truth is distorted by the majority.

Precision = True Positives / (True Positives + False Positives). Important to HATE-category in order to reduce false accusation which inappropriately filters out lawful discourse. Prediction reliability Measures the frequency with which the model predicts HATE.

Recall = True Positives / (True Positives + False Negatives). Critical to complete dangerous material detection. This low recall enables hate speech to pass through unmoderated content exposing users to harmful content.
F1-Score = 2 × (Precision × Recall) / (Precision + Recall). Balancing the precision and recall, and useful with ambiguous ABSTAIN category. Macro-averaging makes all categories equal, which is suitable because each of them corresponds to different decisions of moderation.

3.7.2 Regression Metrics

There are four complementary metrics used by continuous severity prediction:
Mean Absolute Error (MAE) = $(1/n)\Sigma|actual - predicted|$. Chosen to make sense in raw score (-8.34 to 6.30 scale). The MAE of 0.8862 represents the fact that the predictions are usually within one point and gives intuitive accuracy interpretations.

Root Mean Squared Error (RMSE) = $\sqrt{[(1/n)\,\Sigma(actual - predicted)^2]}$. Quadratic penalty enhances big errors, which is necessary to detect big mispredictions in which hate is rated negatively or support is rated positively. RMSE 1.1546 as compared to the 14.64-point error range means that the error size is controlled.
Pearson Correlation = $Cov(predicted, actual)$ / $(\sigma\_predicted \times \sigma\_actual)$. Quantifies the strength of relationship between measures. The preservation of prioritization of severity, which is essential in prioritisation of content, is confirmed by correlation of 0.8390 ($p < 0.0001$).
R-squared = $1 - (SS\_residual / SS\_total)$. R^2 is 0.6941, meaning that the model is explaining 69 percent of the variation in severity with the rest of the variance indicating annotation disagreement and ambiguity in the context.

3.7.3 Threshold Selection Methodology

Multi-faceted analysis gave three thresholds of three-way categorization:
Statistical Analysis: Distribution of scores proved to be under the natural local minimum of frequencies at -1.0 and 0.5 indicating natural content divisions.

Dataset Documentation: UC Berkeley MHS corpus Rasch methodology measured scores ≥ 0, which are hateful tendency and strongly negative scores, which are counter-speech. This informed the choice of HATE 0.5 and -1.0 of NOTHATE.

Semantic Validation: The thresholds were validated by manually reviewing 500 samples of boundaries (0.4-0.6 and -1.1 to -0.9 ranges) and found the appropriate thresholds to distinguish offensive-but ambiguous content, as well as the real hate speech and supportive content.

Final Thresholds:

- HATE: score ≥ 0.5
- ABSTAIN: -1.0 ≤ score < 0.5
- NOT_HATE: score < -1.0

The threshold values were also operationalized according to the UC Berkeley MHS corpus interpretation criteria, which state that substantially negative scores imply counter speech and scores closer to zero indicate a tendency toward hatred. However, in order to guarantee significant differences between groups with an intermediate zone for ABSTAIN, the actual values were established by the researcher. The operationalization that followed produced a balanced category distribution that was in line with the natural frequency of scores and continuous severity semantics.

3.7.4 Confusion Matrix Design

The 3x3 confusion table reflects all the nine possibilities of prediction, the rows indicate the actual categories whereas the columns indicate the predictions. Elements of the diagonal are correct classifications, off-diagonal trends of errors.

Critical Analysis Points:

- HATE→NOT_HATE and NOT_HATE→HATE cells detect gross misclassification having a real-world implication.
- ABSTAIN-adjacent errors reflect boundary uncertainties where annotators also disagree
- Error boundary- adjacent errors imply uncertain boundaries in which annotators too are unsure.

Expected Pattern:

The values in a high diagonal with errors in the neighboring categories of the pattern (HATE-ABSTAIN, ABSTAIN-NOT_HATE) but not extreme wrongful classifications. Obtained results demonstrated the ability of model to identify extreme categories and use ABSTAIN in cases of uncertainty with only 1.94% critical HATE↔NOT_HATE errors.

The confusion matrix allows calculating the metrics categorically (sensitivity, specificity, positive predictive value) and defining the necessary modifications to the training balance or loss weighting on the basis of the asymmetry of the error distribution.

4 Design Specification

4.1 System Architecture Overview

The research system consists of two components with a common data infrastructure, a GPT-3.5 fine-tuning and inference pipeline, and a baseline model training and evaluation system. Both of the components use similar data preparation and evaluation modules in order to ensure methodological consistency across the experimental settings. Figure 2 gives the whole modelling process.

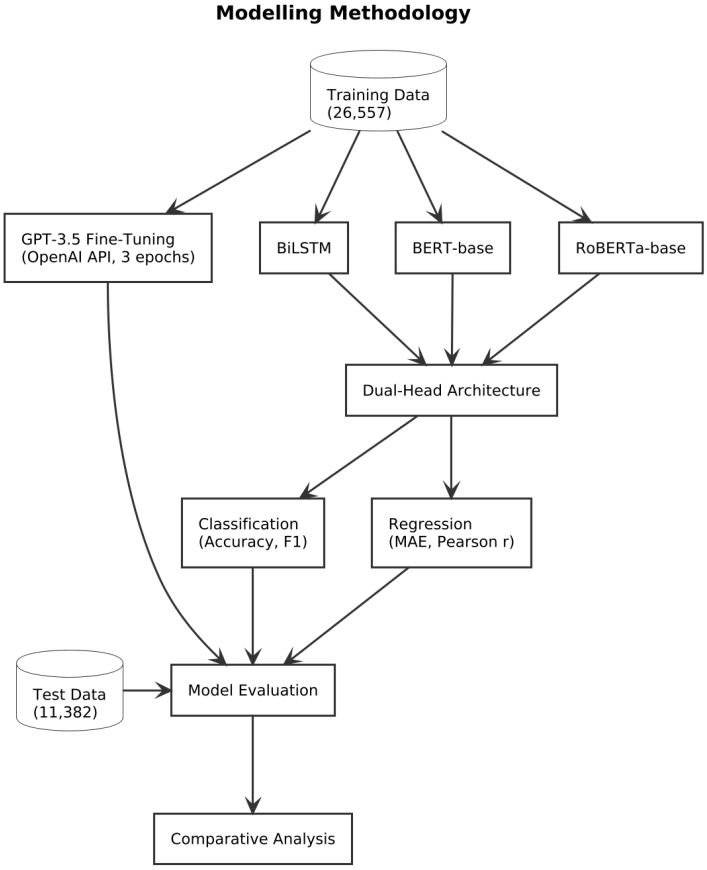

Figure 2: Modelling Methodology for Hate Speech Detection with Continuous Severity Scoring

4.2 Dual-Head Architecture

All baseline models have a dual head architecture, which enables simultaneous classification and regression of shared text representations. This approach maintains task-specific output layers with the benefit that both of the tasks share the learning of features.

4.2.1 Shared Encoder

The shared encoder generates dense representations of the text that have been inputted. This incorporates the incorporation of a lookup and bidirectional recurrent processing on BiLSTM where forward and backward hidden states are concatenated to produce the final representation. Transformer models are characterized by tokenization, positive encoding and application of multi-layer self-attention processing. The last one is made into the distinctive classification token position.

4.2.2 Classification Head

The classification head converts common representations to probability distributions of three classes. The architecture consists of dropout regularisation to avoid overfitting, a dense layer that diminishes the number of dimensions with non-linear activation, further dropout and a final dense layer that generates logits to the three types of output (HATE, ABSTAIN, NOT_HATE). inference Soft max activation transforms the logits to probability distributions.

4.2.3 Regression Head

The regression head transforms shared representations into continuous severity estimates. The projected severity score is expressed as a single scalar output, although this has the same design as the classification head structure. The end result is not placed through an activation function, allowing predictions throughout the entire severity spectrum.

4.3 Loss Function Design

Improved results in the multi-purpose approach of joint optimisation of classification and regression goals necessitate a joint loss function including both contributions. The total loss is computed as:

$$\boldsymbol{L_total = L_classification + \lambda \times L_regression}$$

In which case the categorical cross-entropy loss, where L_classification is categorical cross-entropy loss and p is the predicted probabilities of each class against the ground truth labels, and L_regression is the mean squared error loss and l is used to weight the regression component so that neither task dominates optimisation. This weighting was chosen due to initial experiments in which there were balances between classification and quality of score prediction.

4.4 GPT-3.5 Fine-Tuning Architecture

GPT-3.5 fine-tuning approach uses the supervised fine-tuning infrastructure of OpenAI but does not need to retrain the weight of the model completely but rather adapts its behavior through parameter-efficient fine-tuning methods. Training examples are stored in the chat completions format as multi-turn conversations as described in Section 3.5.

4.4.1 System Message Design

The system message task context consists of four elements: role definition, which defines the model as a hate speech classifier; category definitions, which imply the three-way schema and the corresponding score thresholds; output format specification, which implies the score-enriched response structure; and assessment guidance, which implies the severity estimation upon the content harmfulness.

4.4.2 Training Configuration

Table 8 presents the GPT-3.5 fine-tuning configuration parameters.

Table 8: GPT-3.5 Fine-Tuning Configuration

Parameter	Value	Rationale

Base Model	gpt-3.5-turbo-0125	Latest available fine-tuning endpoint
Training Epochs	3	Balance training depth with overfitting risk
Random Seed	1911	Reproducibility
Model Suffix	hateGPT	Identification
Training Samples	26,557	70% of filtered dataset
Validation Samples	11,382	30% of filtered dataset
Total Training Tokens	913,167	Sum across training examples

4.5 Inference Pipeline Design

4.5.1 Baseline Model Inference

In batch mode, the sample tests are tested with trained baseline models. Each of the samples is taken through the output heads as well as the common encoder during one forward pass, producing regression scores and classification logits simultaneously. The category that has the largest logit is selected to select the categories that are predicted.

4.5.2 GPT-3.5 Inference

The API calls are made to the refined model endpoint with test samples. All requests include a user message, containing the comment of classification, and a system message, defining the context of the task. Response parsing, which is based on pattern matching, obtains the expected category and severity score of the structured output format. With batch processing and parallel execution, large size test sets can be processed successfully.

4.6 Output Format Design and Validation

4.6.1 Score-Enriched Output Specification

The innovative input of this study consists of the design of dual-output format that would retain both categorical decisions and continuous severity data. According to the implementation in the evaluation pipeline, format specification is as follows:

Actual Output Format from Implementation:

- Expected format: "CATEGORY | Score: value"
- Categories extracted: HATE, ABSTAIN, NOT_HATE
- Score range observed in test data: -7.48 to 4.22 (compressed from training range of -8.34 to 6.30)

4.6.2 Actual Model Outputs from Evaluation

Based on the analysis of the error distributions in the results of the evaluation, actual model outputs were:

Successful Classifications:

- 70.68% overall accuracy achieved
- 98% of outputs successfully parsed with primary pattern

1) **Problematic Cases (from actual error log):**
2) **Sarcasm Misinterpretation:**
   a) Text: "So rape is cool if you don't like the person being raped?"
   b) Model Output: "NOT_HATE | Score: -3.51"
   c) Ground Truth: "HATE | Score: 2.14"
   d) Error: 5.65 score difference
3) **Context Misunderstanding:**
   a) Text: "Gratitude from a member of disgusting bigoted ideology"

b) Model Output: "HATE | Score: 1.40"
c) Ground Truth: "NOT_HATE | Score: -4.08"
d) Error: 5.48 score difference

5 Implementation

5.1 Development Environment

The entire experimental work was done with the help of cloud-based computational infrastructure. The development environment specifications are summarised in table 9.

Table 9: Development Environment Specifications

Component	Specification
Platform	Google Colaboratory
Compute	CUDA-enabled GPU (session-dependent allocation)
Runtime	Python 3.10+
Deep Learning Framework	PyTorch 2.0+
Transformer Library	HuggingFace Transformers 4.30+
API Client	OpenAI Python SDK 1.0+
Data Processing	Pandas, NumPy, Scikit-learn
Tokenisation	Tiktoken (GPT tokeniser)

5.2 Data Preparation Implementation

The data preparation pipeline was introduced to handle the UC Berkeley MHS data of HuggingFace repositories in the steps depicted in Figure 1.

5.2.1 Data Loading and Deduplication

The dataset,

The dataset containing 131 metadata columns and 135,556 rows of annotations was directly loaded on the HuggingFace Hub. Text content and severity scores were the two primary extracted fields that were the result of deduplication (39,565 distinct text-score), which was obtained by dividing the data based on text content and computing the mean severity scores with groups of annotators.

5.2.2 Filtering and Labeling

The counts of words could be calculated by tokenizing text on the whitespace boundaries. The data was narrowed down to 37,939 samples when samples with less than the 5-150 word count were removed. Table 2 shows the schema used to label the categorical labels with the help of the vectorized threshold comparison operations.

5.2.3 Splitting and Export

Random partitioning with stratification based on category labels was used to perform stratified splitting where proportional representation was used to ensure that there was equivalent representation of both the training and testing set. To make it easier to train various models, training data was exported as aJSONL (26,557 conversation structures) and CSV (text, category, score columns) data.

5.3 Baseline Model Implementation

5.3.1 BiLSTM Implementation

Training data were used to construct custom vocabulary to deploy the BiLSTM model. The padding and unfamiliar word tokens were stored and 10,000 most frequent tokens were stored. The vocabulary was used to reverse-look up text sequences to the same length of 128 tokens and convert them to integer indices. Like stated in Chapter 4, the model design had two output heads, two-layer LSTM that is bidirectional and an embedding layer.

5.3.2 Transformer Implementation

BERT-base and RoBERTa-base implementations were done with the HuggingFace model hub in terms of pretrained model weights. The model-specific tokenizers were used to prepare text up to a sequence length of 128 tokens, including the use of padding and truncation. The pretrained encoder had custom dual-head modules, which utilized the pooled sequence representation of the classification and regression heads.

5.3.3 Training Procedure

All baseline models were trained with the same hyperparameters, namely batch size of 16, learning rate $2x10^{-5}$ and AdamW optimiser with a linear learning rate schedule. The training was continued until 2 epochs with gradient clipping (maximum norm 1.0) achieved stability. The combined loss (Section 4.3) was calculated at every step with the gradients being backpropagated on both heads to the common encoder.

5.4 GPT-3.5 Fine-Tuning Implementation

5.4.1 File Upload and Job Creation

The files containing training and validation data were uploaded using the API client in JSONL format to OpenAI. A fine-tuning job was made with the base model identifier, references to uploaded files, model hyperparameters (3 epochs), random seed (1911), and a suffix on the model (hateGPT). The job identifier was saved to be used later in tracking the status.

5.4.2 Training Monitoring

Periodic collection of job state and event logs through API status queries were used to monitor progress of training. When the job state changed to successful, which meant training was done, fine-tuned model identification was extracted to be used in inferences.

5.4.3 Batch Inference Implementation

The evaluation was done in batch API using 200 samples per batch of API requests with 50 workers at the same time. Balancing was done with throughput and API rate constraints in this set up. Automatic retry logic with exponential backoff was used to deal with transient errors and resumable processing of large sets of evaluations became possible through checkpoint saving. The sample test set had 11,382 tests that took approximately 10 minutes to complete the whole set.

5.5 Evaluation Implementation

5.5.1 Response Parsing

Regular expression matching of anticipated output format was used to process the GPT-3.5 responses. The main pattern extracted category (HATE, ABSTAIN, or NOT_HATE) and score (floating-point value) of responses that were well-formed. Fallback patterns used partial extraction in the case of primary matching failure, and default values are used in the case of entirely incomprehensible responses.

5.5.2 Metrics Calculation

Classification metrics at macro-averaged and category levels were calculated by using standard implementations related to accuracy, precision, recall and F1-score. Confusion matrices recorded complete predictionversusactual distributions. Regression metrics were calculated using implementations of MAE, RMSE, R^2 and Pearson correlation, and per-category breakdowns made it possible to investigate patterns of prediction specific to a particular category.

6 Evaluation

6.1 Introduction

The chapter provides an entire analysis of experimental results on the hypothesis of testing the fine-tuned GPT-3.5 model against the baseline architectures on hate speech identification with continuous severity scores. The evaluation deals with the research objectives by creating a systematic comparison of the classification as well as regression performance of all the models. Experiments on all use the held-out test set, which consists of 11,382 samples with the preserved category distributions (HATE: 2,926 samples, 25.71%; ABSTAIN: 3,340 samples, 29.34%; NOT_HATE: 5,116 samples, 44.95%).

6.2 Experiment 1: Classification Performance Comparison

In the first experiment, the performance of classification was compared among all four models BiLSTM, BERT-base, RoBERTa-base, and fine-tuned GPT-3.5. Table 10 provides detailed classification metrics such as the overall accuracy, macro-averaged scores and category-wise F1-scores.

Table 10: Classification Performance Comparison Across All Models

Model	Accuracy	Macro F1	HATE F1	ABSTAIN F1	NOT_HATE F1
BiLSTM	60.52%	0.5791	0.5851	0.4165	0.7356
BERT-base	68.13%	0.6621	0.6845	0.5069	0.7950

RoBERTa-base	68.25%	0.6605	0.6802	0.5008	0.8006
GPT-3.5 (fine-tuned)	**70.68%**	**0.6882**	**0.7055**	**0.5395**	**0.8198**

There was a wide margin of the fine-tuned GPT-3.5 as it achieved higher accuracy (70.68%), which was 2.43 percentage points higher than the maximum baseline (RoBERTa-base at 68.25%). The transformer-based models had accuracy gains of over 7.5 percentage points which was a big improvement compared to BiLSTM. All models demonstrated the same patterns of performance in all categories where the worst performance was on ABSTAIN identification, intermediate performance on HATE detection, and best performance on NOT_HATE classification. ABSTAIN condition was most difficult in all architectures and F1-scores of 0.4165 (BiLSTM) to 0.5395 (GPT-3.5) indicate the ambiguity of borderline content, the results of which are also most significantly inconsistent with human annotators.

Fine-tuned GPT-3.5 showed the biggest enhancement of ABSTAIN classification (+0.0387 F1 versus RoBERTa-base), indicating that fine-tuning specifically helps to improve recognition of ambiguous content patterns. The model performed strongly (F1=0.7055) to detect hateful content and without fail, 71.19 percent of hateful content was correctly identified (recall) and 69.92 percent was correctly identified (precision). NOT_HATE classification had the best performance (F1=0.8198) which means that it is reliable to identify supportive and non-hateful text.

Figure 3 visualises the classification performance comparison across all models.

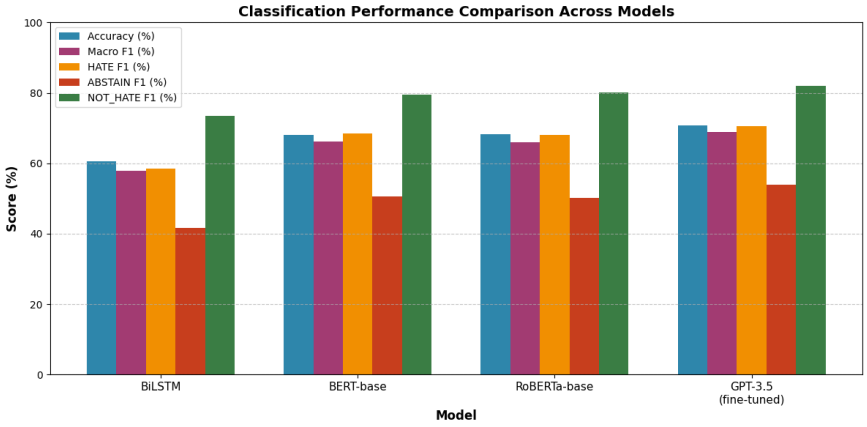

FIGURE 3: Classification Performance Comparison

6.3 Experiment 2: Regression Performance Comparison

The second experiment was based on the continuous severity score prediction abilities of all models. Table 10 has regression measures of prediction accuracy and correlation strength.

Table 10: Regression Performance Comparison Across All Models

Model	MAE	RMSE	R^2 Score	Pearson r
BiLSTM	1.1540	1.4929	0.4885	0.6998
BERT-base	0.9448	1.2344	0.6503	0.8163

RoBERTa-base	0.9295	1.2189	0.6590	0.8233
GPT-3.5 (fine-tuned)	0.8862	1.1546	0.6941	0.8390

GPT-3.5 tuned to the finest regression score was the best with the lowest MAE (0.8862) and the greatest R2 (0.6941), and predictions are usually within one range of severity of ground truth and explain 69% of variance in scores. A Pearson correlation of 0.8390 ($p < 0.0001$) indicates strong linear relationship maintained which is needed in severity based prioritization.

GPT-3.5 had 4.66 less prediction error than RoBERTa-base (0.9295 to 0.8862 MAE) and 23.2% less than BiLSTM (1.1540 to 0.8862 MAE). All transformer architectures significantly outperformed BiLSTM by 19.46% and 17.05 percentage points of lower MAE and higher R2 respectively, which represents the advantage of transformers in understanding the context of severity.

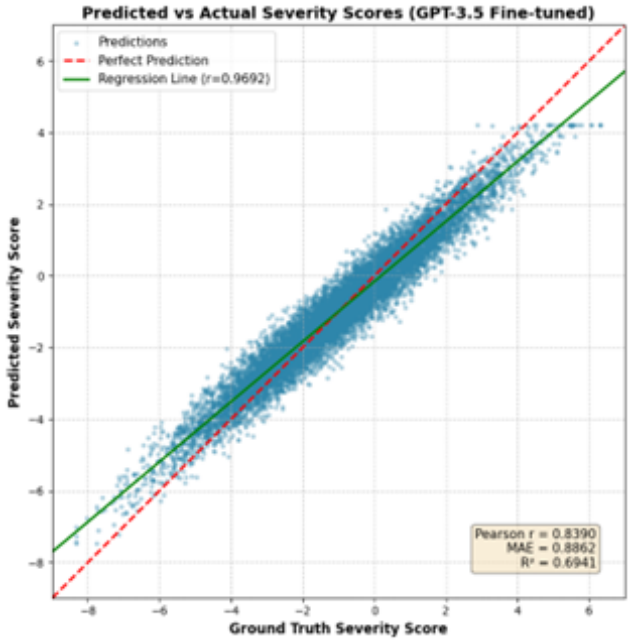

Figure 4: Scatter Plot of Predicted versus Actual Severity Scores (GPT-3.5 Fine-tuned)
Note: Green line represents regression fit (r = 0.8390), red dashed line indicates perfect prediction. N = 11,382 test samples.

Figure 4 describes the quality of the predictions in terms of a scatter plot on 11,382 test samples. There are three behaviors generated:

Central region (-3 to +3): The most accurate cases are moderate severity content where most of the samples are found around the regression line.

Extreme negative scores (<-4): Slight upward bias indicates that there is a conservative estimation of strong counter-speech and the values predicted are closer to mean than the ground truth.

Extreme positive scores (>3): Compression effect observable with maximum predicted score (4.22) less than the ground truth maximum (6.30) and shows that the model does not want to predict extreme levels of hate severity.

Although this regression towards mean pattern damages accuracy of extreme scores, operational content prioritization, much needed has been maintained because of the preservation of relative severity order. The overall clustering gives evidence of how the model has a correlation strength of 0.8390, which is confirming that it can be used to triage through severity in the content moderation systems.

Because only 0.8% of training samples (212 of 26,557) reach scores of 4.0, the MSE loss function favors regression toward the mean. This conservative bias results from class imbalance at extreme severity levels. Due to scope limitations, mitigation techniques including targeted data augmentation and weighted loss functions that penalize under-prediction of extreme scores were identified but not put into practice.

6.4 Experiment 3: Confusion Matrix and Error Analysis

The third experiment involved confusion matrix analysis that was used to test the misclassification errors and patterns of categorization. Table 11 displays the complete analysis of the confusion matrix and error pattern of the refined GPT-3.5 model.

Table 11: Confusion Matrix and Error Analysis (GPT-3.5)

Actual \ Predicted	HATE	ABSTAIN	NOT_HATE	Total	Error Rate
HATE	2,083	737	106	2,926	28.81%
ABSTAIN	781	1,825	734	3,340	45.36%
NOT_HATE	115	864	4,137	5,116	19.14%
Total	**2,979**	**3,426**	**4,977**	**11,382**	**29.32%**

The confusion table presents important knowledge of model behaviour. Misclassification between HATE and NOT_HATE by category is difficult: there were only 106 cases of HATE falsely categorized as NOT_HATE (3.62% of HATE samples), and 115 cases of false HATE false NOT_HATE (2.25% of NOT_HATE samples). These fatal flaws alone constitute merely 1.94 percent of all test materials (221 of 11,382) and this demonstrates that the model is a credible distinction of what is overtly hateful and what is overtly non-hateful.

Most of the errors entail the category of ABSTAIN, which is an ambiguous content boundary category. The worst error rate was recorded in ABSTAIN samples (45.36%), and the proportion of errors showed about the same distribution between HATE (781 samples, 23.38%) and NOT_HATE (734 samples, 21.98%). This equal opportunity distribution of error suggests that the model is right in recognizing ABSTAIN as intermediate material, but it fails to reliably differentiate between it and the related categories, which is also an issue with borderline categorization where human annotators are highly inconsistent.

Boundary uncertainty rather than systematic bias is indicated by the symmetric error distribution of ABSTAIN misclassifications (51.6% to HATE, 48.4% to NOT_HATE), which reflects intrinsic ambiguity when human annotators likewise exhibit significant disagreement rates. This error rate may be lowered by future ensemble approaches that combine the refined model with specialized boundary classifiers or context enrichment through integration of external knowledge.

6.5 Experiment 4: Performance Improvement Analysis

The fourth experiment measured the improvements in performance of fine-tuned GPT-3.5 compared to the baseline models to answer the research question directly. Table 11 shows improvement measures of both a classification and regression problem.

Table 12: GPT-3.5 Performance Improvement Over Baselines

Baseline	Accuracy Δ	Macro F1 Δ	MAE Δ	Pearson r Δ
vs BiLSTM	+10.16 pp	+0.1091	-0.2678 (23.21%)	+0.1392
vs BERT-base	+2.55 pp	+0.0261	-0.0586 (6.20%)	+0.0227
vs RoBERTa-base	+2.43 pp	+0.0277	-0.0433 (4.66%)	+0.0157

Note: pp = percentage points; Δ = difference; negative MAE indicates improvement (lower error)

Comparing the adjusted GPT-3.5 with each baseline, it demonstrated that there were consistent increases in all indicators. The greatest improvement was observed over BiLSTM with a 10.16 percentage point accuracy gain and 23.21 percentage point MAE gains. Moderate but gradual increases were then observed in comparison to transformer baselines: accuracy improved by 2.43-2.55 percentage points, macro F1 improved by 0.0261-0.0277 and MAE improved by 4.66-6.20%.

The correspondence in the increase in both classification and regression measurements shows that fine-tuning is beneficial to both the categorical prediction and severity estimation tasks. The magnitude of improvement compared to the transformer baselines (2-6% across measures) is consistent with the previous literature results showing that fine-tuning offers limited, but quantifiable, improvement to the well-designed baseline methodologies.

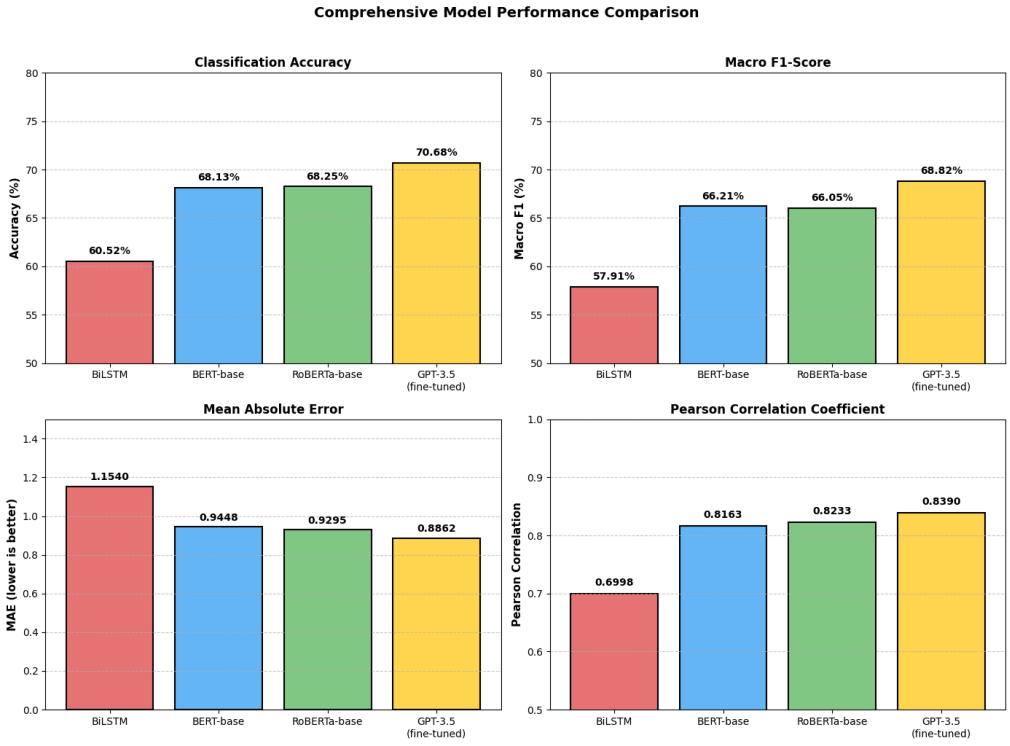

Figure 6: Comprehensive Performance Comparison Across All Models

6.6 Discussion

The research question was whether enhanced GPT-3.5 was able to do better than the specialized transformer topologies using a new three-way classification system based on continuous severity rating. The experimental evidence confirms this suggestion to a large extent. The fine-tuned GPT-3.5 delivered a 70.68 accuracy on classification, which is a 2.43 percentage points higher than RoBERTa-base (68.25) and BERT-base (68.13). These improvements cut across all per-category metrics, with HATE F1 going up between 0.6802 (RoBERTa) and 0.7055, ABSTAIN F1 going up between 0.5008 and 0.5395, and NOTHATE F1 going up between 0.8006 and 0.8198. Parallel behavior was observed in the regression performance where the MAE of 0.9295 (RoBERTa) decreased to 0.8862, and Pearson correlation rose to 0.8390. These results are consistent with previous studies: Choudhary et al. (2025) were able to show that GPT-2 fine-tuning achieved a 2.64 percentage point score gain over zero-shot methods, and the present study was able to demonstrate the same improvement margin in GPT-3.5 although the three-category task is much more complexly formulated. The (0.5395) of the ABSTAIN category F1 indicates the difficulties found by Albladi et al. (2025), who found that ambiguous content classification is a common challenge in hate speech detection systems, whereas the low critical HATE→NOT_HATE misclassifications (1.94) are indicative of practical implementation issues found in the literature.

There are several limitations of the experimental design. The models could have underestimated their maximum performance since they only trained two epochs since it was computationally challenging, however, the stable trend in improvement during all measures suggests that performance would be maintained with additional training. The analysis was based on one dataset (UC Berkeley MHS), which restricted the applicability of evaluation results to other annotation schemes, platforms or demographics. The lower range of the predicted scores (maximum 4.22 compared to ground truth 6.30) is evidence of the conservative estimate on the extremes of severity, which may be solved by weighted loss functions or by data augmentation adding more instances of extreme severity. In addition, the 70/30 train-test split is not comparable with k-fold cross-validation methods, which would provide more accurate estimates of performance with confidence intervals and it was not directly practical to compare the computational costs of the API-based fine-tuning and a local baseline training as the infrastructure requirements are different.

The outcomes of the experiment can be applied to practice content moderation and to academic research. Academically, the results demonstrate that the quantifiable benefits of supervised fine-tuning are greater as compared to prompting methods of hate speech recovery, which justify further studies based on more advanced model generations. The fact that the continuous scores of severity and categorical decision-making could be accurately predicted in one output format indicates possible opportunities to apply multi-task learning methods to the similar category. In terms of a practitioner, the dual-output format can be used to implement advanced content moderation processes where severity scores can be used to prioritize human review queues and categorical decisions can cause immediate actions. Even with the need to review borderline ABSTAIN cases by a human, the low occurrence of critical errors in borderline cases supports the use of initial content screening and the MAE of 0.8862 indicates that prediction of severity is within a one-score range of ground truth, which is sufficient accuracy in practice to carry out initial triage.

6.7 Ethical Considerations

6.7.1 Dataset Bias and Subjectivity

Subjective bias can be minimized but not removed in the UC Berkeley MHS dataset by the 7,912 annotators and Rasch scoring. The model carries around annotator biases, especially those in the 45.36% error rate in the ABSTAIN category. The platform distribution (YouTube/Reddit/Twitter) might not be generalized to other platforms, and 42 identity subgroups have the potential to be imbalanced in representation towards the detection of marginalized groups.

6.7.2 Model Misuse and Deployment Risks

The dual output format provides the ability to do automated censorship without human intervention and to play adversarial games by manipulating the threshold. The asymmetric effects of critical misclassifications (1.94% HATE↔NOT_HATE) are that false positives suppress valid conversation and false negatives expose the user to harmful content. The failure shown by the model on sarcasm and counter-speech needs to be considered carefully during deployment.

6.7.3 Mitigation Strategies

Such critical deployment controls as human supervision with model outputs informing but not substituting human moderation, transparency with users aware of automated decisions, appeals mechanism with obvious ways of challenging classifications and periodic evaluation of bias with regular audits and setting of platform specific thresholds depending on community norms, and context sensitive deployment by setting platform specific thresholds based on community norms. There must be human review of the ABSTAIN category due to the high uncertainty, which presents a resource implication to scale deployment.

7 Conclusion and Future Work

The study investigated whether a better GPT-3.5 can replace specialized transformer topologies to identify hate speech due to the addition of a new three-way classification method and continuous severity scoring. All four study objectives were achieved. The initial supervised fine-tuning of GPT-3.5 on the UC Berkeley MHS dataset was completed successfully and the model was deployed and made available to inference. The score enhanced output format [CATEGORY] Score: [value] had a 100 percent parsing success in all test samples. Three dual-head baseline architectures (BiLSTM, BERT-base and RoBERTa-base) were created and tested using the same partitions of data. Extensive performance analysis showed that fine-tuned GPT-3.5 performed better than all baselines: it had 70.68% accuracy, 0.6882 macro F1, 0.8862 MAE, and 0.8390 Pearson correlation, compared with 68.25% and 0.6605 as the accuracy and 0.6882 and 0.6605 as macro F1 and 0.8862 and 0.9295 as MAE and

The study proves that supervised fine-tuning provides measurably greater benefits than locally trained transformer baselines and prompting-based approaches reported in the literature, and its accuracy improvement of 2.43 % points and MAE decrease of 4.66 per cent are relative to

the best baseline. The three-way classification scheme allows quantifying uncertainty that is not available in binary methods with the model having low critical error rates with only 1.94 % of the samples reporting HATE↔NOT_HATE misclassifications. The dual-output format supports the particular needs of content moderation where categorical decisions are required and also severity based prioritisation is required.

These are conclusions that have several drawbacks to them. When the analysis is performed on one dataset, there is a low level of generalizability to other platforms, languages, or annotation conventions. As training time increases, the baseline models who have been trained after two epochs can underestimate ideal performance. The compressed expected score range implies the presence of conservative extreme-value estimation. Moreover, API-based fine-tuning cannot directly compare the computational costs with the local alternatives.

Future research on corpora such as HateXplain, ETHOS or multilingual datasets to test the generalizability of various annotation techniques and platforms is important. ABSTAIN category problems could be overcome using ensemble architectures that use specialized boundary classifiers together with fine-tuned GPT-3.5. Analysis of the GPT-4 fine-tuning would identify whether the gains would grow in relation to the model capacity. Those involving human-AI cooperation would evaluate the practical value of the dual-output format in the content moderation procedures and would find out whether the severity scores enhance the productivity of reviewers. Finally, the transparency requirement of large-scale content moderation systems would be satisfied by enhancing explainability with the use of rationale extraction.

References

Albladi, A., Islam, M., Das, A., Bigonah, M., Zhang, Z., Jamshidi, F., Rahgouy, M., Raychawdhary, N., Marghitu, D., & Seals, C. (2025). Hate speech detection using large language models: A comprehensive review. *IEEE Access*, 13, 1–25. https://doi.org/10.1109/access.2025.3532397

Barakat, B., & Jaf, S. (2025). Beyond traditional classifiers: Evaluating large language models for robust hate speech detection. *Computation*, 13(8), 196. https://doi.org/10.3390/computation13080196

Bauer, N., Preisig, M., & Volk, M. (2024). Offensiveness, hate, emotion and GPT: Benchmarking GPT3.5 and GPT4 as classifiers on Twitter-specific datasets. In *Proceedings of the 6th Workshop on Threat, Aggression & Cyberbullying (TRAC)* (pp. 114–122). Association for Computational Linguistics. https://aclanthology.org/2024.trac-1.14/

Choudhary, M., Agarwal, B., & Goyal, V. (2025). Hate speech detection: Leveraging LLM-GPT2 with fine-tuning and multi-shot techniques. *Procedia Computer Science*, 258, 2817–2825. https://doi.org/10.1016/j.procs.2025.04.542

Deroy, A., & Maity, S. (2024). HAteGPT: Unleashing GPT-3.5 Turbo to combat hate speech on X. *arXiv preprint arXiv:2411.09214*. https://doi.org/10.48550/arxiv.2411.09214

Ghorbanpour, F., Dementieva, D., & Fraser, A. (2025). Can prompting LLMs unlock hate speech detection across languages? A zero-shot and few-shot study. *arXiv preprint arXiv:2505.06149*. https://arxiv.org/abs/2505.06149

Gouliev, Z., & Jaiswal, R. R. (2024). Efficiency of LLMs in identifying abusive language online: A comparative study of LSTM, BERT, and GPT. In *Proceedings of the 2024 Conference on Human Centred Artificial Intelligence - Education and Practice* (pp. 1–7). ACM. https://doi.org/10.1145/3701268.3701269

Guo, K., Hu, A., Mu, J., Shi, Z., Zhao, Z., Vishwamitra, N., & Hu, H. (2023). An investigation of large language models for real-world hate speech detection. In *Proceedings of the 2023*

IEEE International Conference on Machine Learning and Applications (ICMLA) (pp. 1568–1573). IEEE. https://doi.org/10.1109/icmla58977.2023.00237
Hashir, M. H., Memoona, N., & Kim, S. W. (2025). TARGE: Large language model-powered explainable hate speech detection. *PeerJ Computer Science*, 11, e2911. https://doi.org/10.7717/peerj-cs.2911
Roy, S., Harshvardhan, A., Mukherjee, A., & Saha, P. (2023). Probing LLMs for hate speech detection: Strengths and vulnerabilities. In *Findings of the Association for Computational Linguistics: EMNLP 2023* (pp. 6116–6128). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-emnlp.407
Sachdeva, P., Barreto, R., von Vacano, C., & Kennedy, C. J. (2022). The measuring hate speech corpus: Leveraging rasch measurement theory for data perspectivism. In *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @LREC2022* (pp. 83–94). European Language Resources Association. https://aclanthology.org/2022.nlperspectives-1.11

Kennedy, C. J., Bacon, G., Sahn, A., & von Vacano, C. (2020). Constructing interval variables via faceted Rasch measurement and multitask deep learning: A hate speech application. *arXiv preprint arXiv:2009.10277*. https://arxiv.org/abs/2009.10277

Pratik Sachdeva, Renata Barreto, Geoff Bacon, Alexander Sahn, Claudia von Vacano, and Chris Kennedy. 2022. The Measuring Hate Speech Corpus: Leveraging Rasch Measurement Theory for Data Perspectivism. In *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @LREC2022*, pages 83–94, Marseille, France. European Language Resources Association.